



Feynman's Preface

These are the lectures in physics that I gave last year and the year before to the freshman and sophomore classes at Caltech. The lectures are, of course, not verbatim—they have been edited, sometimes extensively and sometimes less so. The lectures form only part of the complete course. The whole group of 180 students gathered in a big lecture room twice a week to hear these lectures and then they broke up into small groups of 15 to 20 students in recitation sections under the guidance of a teaching assistant. In addition, there was a laboratory session once a week.

The special problem we tried to get at with these lectures was to maintain the interest of the very enthusiastic and rather smart students coming out of the high schools and into Caltech. They have heard a lot about how interesting and exciting physics is—the theory of relativity, quantum mechanics, and other modern ideas. By the end of two years of our previous course, many would be very discouraged because there were really very few grand, new, modern ideas presented to them. They were made to study inclined planes, electrostatics, and so forth, and after two years it was quite stultifying. The problem was whether or not we could make a course which would save the more advanced and excited student by maintaining his enthusiasm.

The lectures here are not in any way meant to be a survey course, but are very serious. I thought to address them to the most intelligent in the class and to make sure, if possible, that even the most intelligent student was unable to completely encompass everything that was in the lectures—by putting in suggestions of applications of the ideas and concepts in various directions outside the main line of attack. For this reason, though, I tried very hard to make all the statements as accurate as possible, to point out in every case where the equations and ideas fitted into the body of physics, and how—when they learned more—things would be modified. I also felt that for such students it is important to indicate what it is that they should—if they are sufficiently clever—be able to understand by deduction from what has been said before, and what is being put in as something new. When new ideas came in, I would try either to deduce them if they were deducible, or to explain that it *was* a new idea which hadn't any basis in terms of things they had already learned and which was not supposed to be provable—but was just added in.

At the start of these lectures, I assumed that the students knew something when they came out of high school—such things as geometrical optics, simple chemistry ideas, and so on. I also didn't see that there was any reason to make the lectures

in a definite order, in the sense that I would not be allowed to mention something until I was ready to discuss it in detail. There was a great deal of mention of things **to** come, without complete discussions. These more complete discussions would come later when the preparation became more advanced. Examples are the discussions of inductance, and of energy levels, which are at first brought in in a very qualitative way and are later developed more completely.

At the same time that I was aiming at the more active student, I also wanted to take care of the fellow for whom the extra fireworks and side applications are merely disquieting and who cannot be expected to learn most of the material in the lecture at all. For such students I wanted there to be at least a central core or backbone of material which he *could* get. Even if he didn't understand everything in a lecture, I hoped he wouldn't get nervous. I didn't expect him to understand everything, but only the central and most direct features. It takes, of course, a certain intelligence on his part to see which are the central theorems and central ideas, and which are the more advanced side issues and applications which he may understand only in later years.

In giving these lectures there was one serious difficulty: in the way the course was given, there wasn't any feedback from the students to the lecturer to indicate how well the lectures were going over. This is indeed a very serious difficulty, and I don't know how good the lectures really are. The whole thing was essentially an experiment. And if I did it again I wouldn't do it the same way—I hope I ~~don't~~ have to do it again! I think, though, that things worked out—so far as the physics is concerned—quite satisfactorily in the first year.

In the second year I was not so satisfied. In the first part of the course, dealing with electricity and magnetism, I couldn't think of any really unique or different way of doing it—of any way that would be particularly more exciting than the usual way of presenting it. So I don't think I did very much in the lectures on electricity and magnetism. At the end of the second year I had originally intended to go on, after the electricity and magnetism, by giving some more lectures on the properties of materials, but mainly to take up things like fundamental modes, solutions of the diffusion equation, vibrating systems, orthogonal functions,... developing the first stages of what are usually called "the mathematical methods of physics." In retrospect, I think that if I were doing it again I would go back to that original idea. But since it was not planned that I would be giving these lectures again, it was suggested that it might be a good idea to try to give an introduction to the quantum mechanics—what you will find in Volume III.

It is perfectly clear that students who will major in physics can wait until their third year for quantum mechanics. On the other hand, the argument was made that many of the students in our course study physics as a background for their primary interest in other fields. And the usual way of dealing with quantum mechanics makes that subject almost unavailable for the great majority of students because they have to take so long to learn it. Yet, in its real applications—especially in its more complex applications, such as in electrical engineering and chemistry—the full machinery of the differential equation approach is not actually used. So I tried to describe the principles of quantum mechanics in a way which wouldn't require that one first know the mathematics of partial differential equations. Even for a physicist I think that is an interesting thing to try to do—to present quantum mechanics in this reverse fashion—for several reasons which may be apparent in the lectures themselves. However, I think that the experiment in the quantum mechanics part was not completely successful—in large part because I really did not have enough time at the end (I should, for instance, have had three or four more lectures in order to deal more completely with such matters as energy bands and the spatial dependence of amplitudes). Also, I had never presented the subject this way before, so the lack of feedback was particularly serious. I now believe the quantum mechanics should be given at a later time. Maybe I'll have a chance to do it again someday. Then I'll do it right.

The reason there are no lectures on how to solve problems is because there **were** recitation sections. Although I did put in three lectures in the first year on how to solve problems, they are not included here. Also there was a lecture on inertial

guidance which certainly belongs after the lecture on rotating systems, but which was, unfortunately, omitted. The fifth and sixth lectures are actually due to Matthew Sands, as I was out of town.

The question, of course, is how well this experiment has succeeded. My own point of view—which, however, does not seem to be shared by most of the people who worked with the students—is pessimistic. I don't think I did very well by the students. When I look at the way the majority of the students handled the problems on the examinations, I think that the system is a failure. Of course, my friends point out to me that there were one or two dozen students who—very surprisingly—understood almost everything in all of the lectures, and who were quite active in working with the material and worrying about the many points in an excited and interested way. These people have now, I believe, a first-rate background in physics—and they are, after all, the ones I was trying to get at. But then, "The power of instruction is seldom of much efficacy except in those happy dispositions where it is almost superfluous." (Gibbon)

Still, I didn't want to leave any student completely behind, as perhaps I did. I think one way we could help the students more would be by putting more hard work into developing a set of problems which would elucidate some of the ideas in the lectures. Problems give a good opportunity to fill out the material of the lectures and make more realistic, more complete, and more settled in the mind the ideas that have been exposed.

I think, however, that there isn't any solution to this problem of education other than to realize that the best teaching can be done only when there is a direct individual relationship between a student and a good teacher—a situation in which the student discusses the ideas, thinks about the things, and talks about the things. It's impossible to learn very much by simply sitting in a lecture, or even by simply doing problems that are assigned. But in our modern times we have so many students to teach that we have to try to find some substitute for the ideal. Perhaps my lectures can make some contribution. Perhaps in some small place where there are individual teachers and students, they may get some inspiration or some ideas from the lectures. Perhaps they will have fun thinking them through—or going on to develop some of the ideas further.

RICHARD P. FEYNMAN

June, 1963

Foreword

This book is based upon a course of lectures in introductory physics given by Prof. R. P. Feynman at the California Institute of Technology during the academic year 1961-62; it covers the first year of the two-year introductory course taken by all Caltech freshmen and sophomores, and was followed in 1962-63 by a similar series covering the second year. The lectures constitute a major part of a fundamental revision of the introductory course, carried out over a four-year period.

The need for a basic revision arose both from the rapid development of physics in recent decades and from the fact that entering freshmen have shown a steady increase in mathematical ability as a result of improvements in high school mathematics course content. We hoped to take advantage of this improved mathematical background, and also to introduce enough modern subject matter to make the course challenging, interesting, and more representative of present-day physics.

In order to generate a variety of ideas on what material to include and how to present it, a substantial number of the physics faculty were encouraged to offer their ideas in the form of topical outlines for a revised course. Several of these were presented and were thoroughly and critically discussed. It was agreed almost at once that a basic revision of the course could not be accomplished either by merely adopting a different textbook, or even by writing one *ab initio*, but that the new course should be centered about a set of lectures, to be presented at the rate of two or three per week; the appropriate text material would then be produced as a secondary operation as the course developed, and suitable laboratory experiments would also be arranged to fit the lecture material. Accordingly, a rough outline of the course was established, but this was recognized as being incomplete, tentative, and subject to considerable modification by whoever was to bear the responsibility for actually preparing the lectures.

Concerning the mechanism by which the course would finally be brought to life, several plans were considered. These plans were mostly rather similar, involving a cooperative effort by N staff members who would share the total burden symmetrically and equally: each man would take responsibility for $1/N$ of the material, deliver the lectures, and write text material for his part. However, the unavailability of sufficient staff, and the difficulty of maintaining a uniform point of view because of differences in personality and philosophy of individual participants, made such plans seem unworkable.

The realization that we actually possessed the means to create not just a new and different physics course, but possibly a unique one, came as a happy inspiration to Professor Sands. He suggested that Professor R. P. Feynman prepare and deliver the lectures, and that these be tape-recorded. When transcribed and edited, they would then become the textbook for the new course. This is essentially the plan that was adopted.

It was expected that the necessary editing would be minor, mainly consisting of supplying figures, and checking punctuation and grammar; it was to be done by one or two graduate students on a part-time basis. Unfortunately, this expectation was short-lived. It was, in fact, a major editorial operation to transform the verbatim transcript into readable form, even without the reorganization or revision of the subject matter that was sometimes required. Furthermore, it was not a job for a technical editor or for a graduate student, but one that required the close attention of a professional physicist for from ten to twenty hours per lecture!

The difficulty of the editorial task, together with the need to place the material in the hands of the students as soon as possible, set a strict limit upon the amount of "polishing" of the material that could be accomplished, and thus we were forced to aim toward a preliminary but technically correct product that could be used immediately, rather than one that might be considered final or finished. Because of an urgent need for more copies for our students, and a heartening interest on the part of instructors and students at several other institutions, we decided to publish the material in its preliminary form rather than wait for a further major revision which might never occur. We have no illusions as to the completeness, smoothness, or logical organization of the material; in fact, we plan several minor modifications in the course in the immediate future, and we hope that it will not become static in form or content.

In addition to the lectures, which constitute a centrally important part of the course, it was necessary also to provide suitable exercises to develop the students' experience and ability, and suitable experiments to provide first-hand contact with the lecture material in the laboratory. Neither of these aspects is in as advanced a state as the lecture material, but considerable progress has been made. Some exercises were made up as the lectures progressed, and these were expanded and amplified for use in the following year. However, because we are not yet satisfied that the exercises provide sufficient variety and depth of application of the lecture material to make the student fully aware of the tremendous power being placed at his disposal, the exercises are published separately in a less permanent form in order to encourage frequent revision.

A number of new experiments for the new course have been devised by Professor H. V. Neher. Among these are several which utilize the extremely low friction exhibited by a gas bearing: a novel linear air trough, with which quantitative measurements of one-dimensional motion, impacts, and harmonic motion can be made, and an air-supported, air-driven Maxwell top, with which accelerated rotational motion and gyroscopic precession and nutation can be studied. The development of new laboratory experiments is expected to continue **for a** considerable period of time.

The revision program was under the direction of Professors R. B. Leighton, H. V. Neher, and M. Sands. Officially participating in the program were Professors R. P. Feynman, G. Neugebauer, R. M. Sutton, H. P. Stabler,* F. Strong, and R. Vogt, from the division of Physics, Mathematics and Astronomy, and Professors T. Caughey, M. Plesset, and C. H. Wilts from the division of Engineering Science. The valuable assistance of all those contributing to the revision program is gratefully acknowledged. We are particularly indebted to the Ford Foundation, without whose financial assistance this program could not have been carried out.

ROBERT B. LEIGHTON

July, 1963

* 1961-62, while on leave from Williams College, Williamstown, Mass.

Contents

CHAPTER 1. ATOMS IN MOTION

- 1-1 Introduction 1-1
- 1-2 Matter is made of atoms 1-2
- 1-3 Atomic processes 1-5
- 1-4 Chemical reactions 1-6

CHAPTER 2. BASIC PHYSICS

- 2-1 Introduction 2-1
- 2-2 Physics before 1920 2-3
- 2-3 Quantum physics 2-6
- 2-4 Nuclei and particles 2-8

CHAPTER 3. THE RELATION OF PHYSICS TO OTHER SCIENCES

- 3-1 Introduction 3-1
- 3-2 Chemistry 3-1
- 3-3 Biology 3-2
- 3-4 Astronomy 3-6
- 3-5 Geology 3-7
- 3-6 Psychology 3-8
- 3-7 How did it get that way? 3-9

CHAPTER 4. CONSERVATION OF ENERGY

- 4-1 What is energy? 4-1
- 4-2 Gravitational potential energy 4-2
- 4-3 Kinetic energy 4-5
- 4-4 Other forms of energy 4-6

CHAPTER 5. TIME AND DISTANCE

- 5-1 Motion 5-1
- 5-2 Time 5-1
- 5-3 Short times 5-2
- 5-4 Long times 5-3
- 5-5 Units and standards of time 5-5
- 5-6 Large distances 5-5
- 5-7 Short distances 5-8

CHAPTER 6. PROBABILITY

- 6-1 Chance and likelihood 6-1
- 6-2 Fluctuations 6-3
- 6-3 The random walk 6-5
- 6-4 A probability distribution 6-7
- 6-5 The uncertainty principle 6-10

CHAPTER 7. THE THEORY OF GRAVITATION

- 7-1 Planetary motions 7-1
- 7-2 Kepler's laws 7-1
- 7-3 Development of dynamics 7-2
- 7-4 Newton's law of gravitation 7-3
- 7-5 Universal gravitation 7-5
- 7-6 Cavendish's experiment 7-9
- 7-7 What is gravity? 7-9
- 7-8 Gravity and relativity 7-11

CHAPTER 8. MOTION

- 8-1 Description of motion 8-1
- 8-2 Speed 8-2
- 8-3 Speed as a derivative 8-5
- 8-4 Distance as an integral 8-7
- 8-5 Acceleration 8-8

CHAPTER 9. NEWTON'S LAWS OF DYNAMICS

- 9-1 Momentum and force 9-1
- 9-2 Speed and velocity 9-2
- 9-3 Components of velocity, **acceleration, and force** 9-3
- 9-4 What is the force? 9-3
- 9-5 Meaning of the dynamical equations 9-4
- 9-6 Numerical solution of the equations 9-5
- 9-7 Planetary motions 9-6

CHAPTER 10. CONSERVATION OF MOMENTUM

- 10-1 Newton's Third Law 10-1
- 10-2 Conservation of momentum 10-2
- 10-3 Momentum *is* conserved! 10-5
- 10-4 Momentum and energy 10-7
- 10-5 Relativistic momentum 10-8

CHAPTER 11. VECTORS

- 11-1 Symmetry in physics **11-1**
- 11-2 Translations 11-1
- 11-3 Rotations 11-3
- 11-4 Vectors 11-5
- 11-5** Vector algebra **11-6**
- 11-6 Newton's laws in vector notation **11-7**
- 11-7 Scalar product of vectors 11-8

CHAPTER 12. CHARACTERISTICS OF FORCE

- 12-1 What is a force? 12-1
- 12-2 Friction 12-3
- 12-3 Molecular forces 12-6
- 12-4 Fundamental forces. Fields 12-7
- 12-5 Pseudo forces 12-10
- 12-6 Nuclear forces 12-12

CHAPTER 13. WORK AND POTENTIAL ENERGY (A)

- 13-1 Energy of a falling body 13-1
- 13-2 Work done by gravity 13-3
- 13-3 Summation of energy 13-6
- 13-4 Gravitational field of large objects 13-8

CHAPTER 14. WORK AND POTENTIAL ENERGY (conclusion)

- 14-1 Work 14-1
- 14-2 Constrained motion 14-3
- 14-3 Conservative forces 14-3
- 14-4 Nonconservative forces 14-6
- 14-5 Potentials and fields 14-7

CHAPTER 15. THE SPECIAL THEORY OF RELATIVITY

- 15-1** The principle of relativity 15-1
- 15-2** The Lorentz transformation 15-3
- 15-3** The Michelson-Morley experiment 15-3
- 15-4** Transformation of time 15-5
- 15-5** The Lorentz contraction 15-7
- 15-6** Simultaneity 15-7
- 15-7 Four-vectors 15-8
- 15-8 Relativistic dynamics 15-9
- 15-9** Equivalence of mass and energy 15-10

CHAPTER 16. RELATIVISTIC ENERGY AND MOMENTUM

- 16-1** Relativity and the philosophers **16-1**
- 16-2 The twin paradox 16-3
- 16-3 Transformation of velocities 16-4
- 16-4 Relativistic mass 16-6
- 16-5 Relativistic energy 16-8

CHAPTER 17. SPACE-TIME

- 17-1 The geometry of space-time 17-1
- 17-2 Space-time intervals 17-2
- 17-3 Past, present, and future 17-4
- 17-4 More about four-vectors 17-5
- 17-5 Four-vector algebra 17-7.

CHAPTER 18. ROTATION IN Two DIMENSIONS

- 18-1 The center of mass 18-1
- 18-2 Rotation of a rigid body 18-2
- 18-3 Angular momentum 18-5
- 18-4 Conservation of angular momentum 18-6

CHAPTER 19. CENTER OF MASS; MOMENT OF INERTIA

- 19-1 Properties of the center of mass 19-1
- 19-2 Locating the center of mass 19-4
- 19-3 Finding the moment of inertia 19-5
- 19-4** Rotational kinetic energy 19-7

CHAPTER 20. ROTATION IN SPACE

- 20-1 Torques in three dimensions 20-1
- 20-2 The rotation equations using cross products **20-1**
- 20-3 The gyroscope 20-5
- 20-4 Angular momentum of a solid body 20-8

CHAPTER 21. THE HARMONIC OSCILLATOR

- 21-1 Linear differential equations 21-1
- 21-2 The harmonic oscillator 21-1
- 21-3 Harmonic motion and circular motion 21-4
- 21-4 Initial conditions 21-4
- 21-5 Forced oscillations 21-5

CHAPTER 22. ALGEBRA

- 22-1 Addition and multiplication 22-1
- 22-2 The inverse operations 22-2
- 22-3 Abstraction and generalization 22-3
- 22-4 Approximating irrational numbers 22-4
- 22-5 Complex numbers 22-7
- 22-6 Imaginary exponents 22-9

CHAPTER 23. RESONANCE

- 23-1 Complex numbers and harmonic motion 23-1
- 23-2 The forced oscillator with damping 23-3

23-3 Electrical resonance 23-5**23-4 Resonance in nature 23-7****CHAPTER 24. TRANSIENTS**

- 24-1** The energy of an oscillator 24-1
- 24-2** Damped oscillations 24-2
- 24-3** Electrical transients 24-5

CHAPTER 25. LINEAR SYSTEMS AND REVIEW

- 25-1** Linear differential equations 25-1
- 25-2** Superposition of solutions 25-2
- 25-3** Oscillations in linear systems 25-5
- 25-4** Analogs in physics 25-6
- 25-5** Series and parallel impedances 25-8

CHAPTER 26. OPTICS: THE PRINCIPLE OF LEAST TIME

- 26-1** Light 26-1
- 26-2** Reflection and refraction 26-2
- 26-3** Fermat's principle of least time 26-3
- 26-4** Applications of Fermat's principle 26-5
- 26-5** A more precise statement of Fermat's principle 26-7
- 26-6** How it works 26-8

CHAPTER 27. GEOMETRICAL OPTICS

- 27-1** Introduction 27-1
- 27-2** The focal length of a spherical surface 27-1
- 27-3** The focal length of a lens 27-4
- 27-4** Magnification 27-5
- 27-5** Compound lenses **27-6**
- 27-6** Aberrations 27-7
- 27-7** Resolving power 27-7

CHAPTER 28. ELECTROMAGNETIC RADIATION

- 28-1** Electromagnetism 28-1
- 28-2** Radiation 28-3
- 28-3** The dipole radiator 28-5
- 28-4** Interference 28-6

CHAPTER 29. INTERFERENCE

- 29-1** Electromagnetic waves 29-1
- 29-2** Energy of radiation 29-2
- 29-3** Sinusoidal waves 29-2
- 29-4** Two dipole radiators 29-3
- 29-5** The mathematics of interference 29-5

CHAPTER 30. DIFFRACTION

- 30-1** The resultant amplitude due to n equal oscillators 30-1
- 30-2** The diffraction grating 30-3
- 30-3** Resolving power of a grating **30-5**
- 30-4** The parabolic antenna 30-6
- 30-5** Colored films; crystals 30-7
- 30-6** Diffraction by opaque screens 30-8
- 30-7** The field of a plane of oscillating charges 30-10

CHAPTER 31. THE ORIGIN OF THE REFRACTIVE INDEX

- 31-1** The index of refraction 31-1
- 31-2** The field due to the material 31-4
- 31-3** Dispersion 31-6
- 31-4** Absorption 31-8
- 31-5** The energy carried by an electric wave 31-9
- 31-6** Diffraction of light by a screen 31-10

CHAPTER 32. RADIATION DAMPING. LIGHT SCATTERING

- 32-1 Radiation resistance 32-1
- 32-2 The rate of radiation of energy 32-2
- 32-3 Radiation damping 32-3
- 32-4 Independent sources 32-5
- 32-5 Scattering of light 32-6

CHAPTER 33. POLARIZATION

- 33-1 The electric vector of light 33-1
- 33-2 Polarization of scattered light 33-3
- 33-3 Birefringence 33-3
- 33-4 Polarizers 33-5
- 33-5 Optical activity 33-6
- 33-6 The intensity of reflected light 33-7
- 33-7 Anomalous refraction 33-9

CHAPTER 34. RELATIVISTIC EFFECTS IN RADIATION

- 34-1 Moving sources 34-1
- 34-2 Finding the "apparent" motion 34-2
- 34-3 Synchrotron radiation 34-3
- 34-4 Cosmic synchrotron radiation 34-6
- 34-5 Bremsstrahlung 34-6
- 34-6 The Doppler effect 34-7
- 34-7 The $\pm k$ four-vector 34-9
- 34-8 Aberration 34-10
- 34-9 The momentum of light 34-10

CHAPTER 35. COLOR VISION

- 35-1 The human eye 35-1
- 35-2 Color depends on intensity 35-2
- 35-3 Measuring the color sensation 35-3
- 35-4 The chromaticity diagram 35-6
- 35-5 The mechanism of color vision 35-7
- 35-6 Physiochemistry of color vision 35-9

CHAPTER 36. MECHANISMS OF SEEING

- 36-1 The sensation of color 36-1
- 36-2 The physiology of the eye 36-3
- 36-3 The rod cells 36-6
- 36-4 The compound (insect) eye 36-6
- 36-5 Other eyes 36-9
- 36-6 Neurology of vision 36-9

CHAPTER 37. QUANTUM BEHAVIOR

- 37-1 Atomic mechanics 37-1
- 37-2 An experiment with bullets 37-2
- 37-3 An experiment with waves 37-3
- 37-4 An experiment with electrons 37-4
- 37-5 The interference of electron waves 37-5
- 37-6 Watching the electrons 37-7
- 37-7 First principles of quantum mechanics 37-10
- 37-8 The uncertainty principle 37-11

CHAPTER 38. THE RELATION OF WAVE AND PARTICLE VIEWPOINTS

- 38-1 Probability wave amplitudes 38-1
- 38-2 Measurement of position and momentum 38-2
- 38-3 Crystal diffraction 38-4
- 38-4 The size of an atom 38-5

- 38-5 Energy levels 38-7
- 38-6 Philosophical implications 38-8

CHAPTER 39. THE KINETIC THEORY OF GASES

- 39-1 Properties of matter 39-1
- 39-2 The pressure of a gas 39-2
- 39-3 Compressibility of radiation 39-6
- 39-4 Temperature and kinetic energy 39-6
- 39-5 The ideal gas law 39-10

CHAPTER 40. THE PRINCIPLES OF STATISTICAL MECHANICS

- 40-1 The exponential atmosphere 40-1
- 40-2 The Boltzmann law 40-2
- 40-3 Evaporation of a liquid 40-3
- 40-4 The distribution of molecular speeds 40-4
- 40-5 The specific heats of gases 40-7
- 40-6 The failure of classical physics 40-8

CHAPTER 41. THE BROWNIAN MOVEMENT

- 41-1 Equipartition of energy 41-1
- 41-2 Thermal equilibrium of radiation 41-3
- 41-3 Equipartition and the quantum oscillator 41-6
- 41-4 The random walk 41-8

CHAPTER 42. APPLICATIONS OF KINETIC THEORY

- 42-1 Evaporation 42-1
- 42-2 Thermionic emission 42-4
- 42-3 Thermal ionization 42-5
- 42-4 Chemical kinetics 42-7
- 42-5 Einstein's laws of radiation 42-8

CHAPTER 43. DIFFUSION

- 43-1 Collisions between molecules 43-1
- 43-2 The mean free path 43-3
- 43-3 The drift speed 43-4
- 43-4 Ionic conductivity 43-6
- 43-5 Molecular diffusion 43-7
- 43-6 Thermal conductivity 43-9

CHAPTER 44. THE LAWS OF THERMODYNAMICS

- 44-1 Heat engines; the first law 44-1
- 44-2 The second law 44-3
- 44-3 Reversible engines 44-4
- 44-4 The efficiency of an ideal engine 44-7
- 44-5 The thermodynamic temperature 44-9
- 44-6 Entropy 44-10

CHAPTER 45. ILLUSTRATIONS OF THERMODYNAMICS

- 45-1 Internal energy 45-1
- 45-2 Applications 45-4
- 45-3 The Clausius-Clapeyron equation 45-6

CHAPTER 46. RATCHET AND PAWL

- 46-1 How a ratchet works 46-1
- 46-2 The ratchet as an engine 46-2
- 46-3 Reversibility in mechanics 46-4
- 46-4 Irreversibility 46-5
- 46-5 Order and entropy 46-7

CHAPTER 47. SOUND. THE WAVE EQUATION

- 47-1 Waves 47-1
- 47-2 The propagation of sound 47-3
- 47-3 The wave equation 47-4
- 47-4 Solutions of the wave equation 47-6
- 47-5 **The speed of sound** 47-7

CHAPTER 48. BEATS

- 48-1 Adding two waves 48-1
- 48-2 Beat notes and modulation 48-3
- 48-3 Side bands 48-4
- 48-4 Localized wave trains 48-5
- 48-5 Probability amplitudes for particles 48-7
- 48-6 Waves in three dimensions 48-9
- 48-7 Normal modes 48-10

CHAPTER 49. MODES

- 49-1 The reflection of waves 49-1
- 49-2 Confined waves, with natural frequencies 49-2
- 49-3 Modes in two dimensions 49-3
- 49-4 Coupled pendulums 49-6
- 49-5 Linear systems 49-7

INDEX

CHAPTER 50. HARMONICS

- 50-1 Musical tones 50-1
- 50-2 The Fourier series 50-2
- 50-3 Quality and consonance 50-3
- 50-4 The Fourier coefficients 50-5
- 50-5 The energy theorem 50-7
- 50-6 Nonlinear responses 50-8

CHAPTER 51. WAVES

- 51-1 Bow waves **51-1**
- 51-2 Shock waves 51-2
- 51-3 Waves in solids 51-4
- 51-4 Surface waves 51-7

CHAPTER 52. SYMMETRY IN PHYSICAL LAWS

- 52-1 Symmetry operations 52-1
- 52-2 Symmetry **in space and time** **52-1**
- 52-3 Symmetry and conservation laws 52-3
- 52-4 Mirror reflections 52-4
- 52-5 Polar and axial vectors 52-6
- 52-6 Which hand is right? 52-8
- 52-7 Parity is not conserved! 52-8
- 52-8 Antimatter 52-10
- 52-9 Broken symmetries 52-11

Atoms in Motion

1-1 Introduction

This two-year course in physics is presented from the point of view that you, the reader, are going to be a physicist. This is not necessarily the case of course, but that is what every professor in every subject assumes! If you are going to be a physicist, you will have a lot to study: two hundred years of the most rapidly developing field of knowledge that there is. So much knowledge, in fact, that you might think that you cannot learn all of it in four years, and truly you cannot; you will have to go to graduate school too!

Surprisingly enough, in spite of the tremendous amount of work that has been done for all this time it is possible to condense the enormous mass of results to a large extent—that is, to find *laws* which summarize all our knowledge. Even so, the laws are so hard to grasp that it is unfair to you to start exploring this tremendous subject without some kind of map or outline of the relationship of one part of the subject of science to another. Following these preliminary remarks, the first three chapters will therefore outline the relation of physics to the rest of the sciences, the relations of the sciences to each other, and the meaning of science, to help us develop a "feel" for the subject.

You might ask why we cannot teach physics by just giving the basic laws on page one and then showing how they work in all possible circumstances, as we do in Euclidean geometry, where we state the axioms and then make all sorts of deductions. (So, not satisfied to learn physics in four years, you want to learn it in four minutes?) We cannot do it in this way for two reasons. First, we do not yet know all the basic laws: there is an expanding frontier of ignorance. Second, the correct statement of the laws of physics involves some very unfamiliar ideas which require advanced mathematics for their description. Therefore, one needs a considerable amount of preparatory training even to learn what the *words* mean. No, it is not possible to do it that way. We can only do it piece by piece.

Each piece, or part, of the whole of nature is always merely an *approximation to the complete truth*, or the complete truth so far as we know it. In fact, everything we know is only some kind of approximation, because *we know that we do not know all the laws* as yet. Therefore, things must be learned only to be unlearned again or, more likely, to be corrected.

The principle of science, the definition, almost, is the following: *The test of all knowledge is experiment*. Experiment is the *sole judge* of scientific "truth." But what is the source of knowledge? Where do the laws that are to be tested come from? Experiment, itself, helps to produce these laws, in the sense that it gives us hints. But also needed is *imagination* to create from these hints the great generalizations—to guess at the wonderful, simple, but very strange patterns beneath them all, and then to experiment to check again whether we have made the right guess. This imagining process is so difficult that there is a division of labor in physics: there are *theoretical* physicists who imagine, deduce, and guess at new laws, but do not experiment; and then there are *experimental* physicists who experiment, imagine, deduce, and guess.

We said that the laws of nature are approximate: that we first find the "wrong" ones, and then we find the "right" ones. Now, how can an experiment be "wrong"? First, in a trivial way: if something is wrong with the apparatus that you did not notice. But these things are easily fixed, and checked back and forth. So without snatching at such minor things, how *can* the results of an experiment be wrong? Only by being inaccurate. For example, the mass of an object never seems to

1-1 Introduction

1-2 Matter is made of atoms

1-3 Atomic processes

1-4 Chemical reactions

change; a spinning top has the same weight as a still one. So a "law" was invented: mass is constant, independent of speed. That "law" is now found to be incorrect. Mass is found to increase with velocity, but appreciable increases require velocities near that of light. A *true* law is: if an object moves with a speed of less than one hundred miles a second the mass is constant to within one part in a million. In some such approximate form this is a correct law. So in practice one might think that the new law makes no significant difference. Well, yes and no. For ordinary speeds we can certainly forget it and use the simple constant-mass law as a good approximation. But for high speeds we are wrong, and the higher the speed, the more wrong we are.

Finally, and most interesting, *philosophically we are completely wrong* with the approximate law. Our entire picture of the world has to be altered even though the mass changes only by a little bit. This is a very peculiar thing about the philosophy, or the ideas, behind the laws. Even a very small effect sometimes requires profound changes in our ideas.

Now, what should we teach first? Should we teach the *correct* but unfamiliar law with its strange and difficult conceptual ideas, for example the theory of relativity, four-dimensional space-time, and so on? Or should we first teach the simple "constant-mass" law, which is only approximate, but does not involve such difficult ideas? The first is more exciting, more wonderful, and more fun, but the second is easier to get at first, and is a first step to a real understanding of the second idea. This point arises again and again in teaching physics. At different times we shall have to resolve it in different ways, but at each stage it is worth learning what is now known, how accurate it is, how it fits into everything else, and how it may be changed when we learn more.

Let us now proceed with our outline, or general map, of our understanding of science today (in particular, physics, but also of other sciences on the periphery), so that when we later concentrate on some particular point we will have some idea of the background, why that particular point is interesting, and how it fits into the big structure. So, what *is* our over-all picture of the world?

1-2 Matter is made of atoms

If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generations of creatures, what statement would contain the most information in the fewest words? I believe it is the *atomic hypothesis* (or the *atomic fact*, or whatever you wish to call it) that *all things are made of atoms—little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another*. In that one sentence, you will see, there is an *enormous* amount of information about the world, if just a little imagination and thinking are applied.

To illustrate the power of the atomic idea, suppose that we have a drop of water a quarter of an inch on the side. If we look at it very closely we see nothing but water—smooth, continuous water. Even if we magnify it with the best optical microscope available—roughly two thousand times—then the water drop will be roughly forty feet across, about as big as a large room, and if we looked rather closely, we would *still* see relatively smooth water—but here and there small football-shaped things swimming back and forth. Very interesting. These are paramecia. You may stop at this point and get so curious about the paramecia with their wiggling cilia and twisting bodies that you go no further, except perhaps to magnify the paramecia still more and see inside. This, of course, is a subject for biology, but for the present we pass on and look still more closely at the water material itself, magnifying it two thousand times again. Now the drop of water extends about fifteen miles across, and if we look very closely at it we see a kind of teeming, something which no longer has a smooth appearance—it looks something like a crowd at a football game as seen from a very great distance. In order to see what this teeming is about, we will magnify it another two hundred and fifty times and we will see something similar to what is shown in Fig. 1-1. This is a picture of water magnified a billion times, but idealized in several ways.

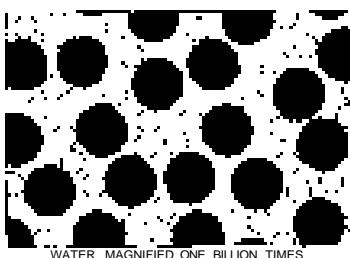


Figure 1-1

In the first place, the particles are drawn in a simple manner with sharp edges, which is inaccurate. Secondly, for simplicity, they are sketched almost schematically in a two-dimensional arrangement, but of course they are moving around in three dimensions. Notice that there are two kinds of "blobs" or circles to represent the atoms of oxygen (black) and hydrogen (white), and that each oxygen has two hydrogens tied to it. (Each little group of an oxygen with its two hydrogens is called a molecule.) The picture is idealized further in that the real particles in nature are continually jiggling and bouncing, turning and twisting around one another. You will have to imagine this as a dynamic rather than a static picture. Another thing that cannot be illustrated in a drawing is the fact that the particles are "stuck together"—that they attract each other, this one pulled by that one, etc. The whole group is "glued together," so to speak. On the other hand, the particles do not squeeze through each other. If you try to squeeze two of them too close together, they repel.

The atoms are 1 or 2×10^{-8} cm in radius. Now 10^{-8} cm is called an *angstrom* (just as another name), so we say they are 1 or 2 angstroms (\AA) in radius. Another way to remember their size is this: if an apple is magnified to the size of the earth, then the atoms in the apple are approximately the size of the original apple.

Now imagine this great drop of water with all of these jiggling particles stuck together and tagging along with each other. The water keeps its volume; it does not fall apart, because of the attraction of the molecules for each other. If the drop is on a slope, where it can move from one place to another, the water will flow, but it does not just disappear—things do not just fly apart—because of the molecular attraction. Now the jiggling motion is what we represent as *heat*: when we increase the temperature, we increase the motion. If we heat the water, the jiggling increases and the volume between the atoms increases, and if the heating continues there comes a time when the pull between the molecules is not enough to hold them together and they *do* fly apart and become separated from one another. Of course, this is how we manufacture steam out of water—by increasing the temperature; the particles fly apart because of the increased motion.

In Fig. 1-2 we have a picture of steam. This picture of steam fails in one respect: at ordinary atmospheric pressure there might be only a few molecules in a whole room, and there certainly would not be as many as three in this figure. Most squares this size would contain none—but we accidentally have two and a half or three in the picture (just so it would not be completely blank). Now in the case of steam we see the characteristic molecules more clearly than in the case of water. For simplicity, the molecules are drawn so that there is a 120° angle between them. In actual fact the angle is $105^\circ 3'$, and the distance between the center of a hydrogen and the center of the oxygen is 0.957 \AA , so we know this molecule very well.

Let us see what some of the properties of steam vapor or any other gas are. The molecules, being separated from one another, will bounce against the walls. Imagine a room with a number of tennis balls (a hundred or so) bouncing around in perpetual motion. When they bombard the wall, this pushes the wall away. (Of course we would have to push the wall back.) This means that the gas exerts a jittery force which our coarse senses (not being ourselves magnified a billion times) feels only as an *average push*. In order to confine a gas we must apply a pressure. Figure 1-3 shows a standard vessel for holding gases (used in all textbooks), a cylinder with a piston in it. Now, it makes no difference what the shapes of water molecules are, so for simplicity we shall draw them as tennis balls or little dots. These things are in perpetual motion in all directions. So many of them are hitting the top piston all the time that to keep it from being patiently knocked out of the tank by this continuous banging, we shall have to hold the piston down by a certain force, which we call the *pressure* (really, the pressure times the area is the force). Clearly, the force is proportional to the area, for if we increase the area but keep the number of molecules per cubic centimeter the same, we increase the number of collisions with the piston in the same proportion as the area was increased.

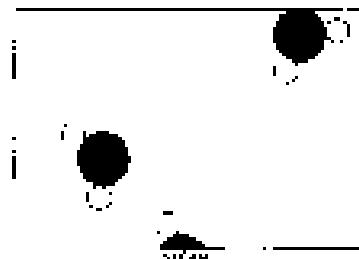


Figure 1-2

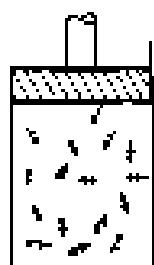


Figure 1-3

Now let us put twice as many molecules in this tank, so as to double the density, and let them have the same speed, i.e., the same temperature. Then, to a close approximation, the number of collisions will be doubled, and since each will be just as "energetic" as before, the pressure is proportional to the density. If we consider the true nature of the forces between the atoms, we would expect a slight decrease in pressure because of the attraction between the atoms, and a slight increase because of the finite volume they occupy. Nevertheless, to an excellent approximation, if the density is low enough that there are not many atoms, *the pressure is proportional to the density*.

We can also see something else: If we increase the temperature without changing the density of the gas, i.e., if we increase the speed of the atoms, what is going to happen to the pressure? Well, the atoms hit harder because they are moving faster, and in addition they hit more often, so the pressure increases. You see how simple the ideas of atomic theory are.

Let us consider another situation. Suppose that the piston moves inward, so that the atoms are slowly compressed into a smaller space. What happens when an atom hits the moving piston? Evidently it picks up speed from the collision. You can try it by bouncing a ping-pong ball from a forward-moving paddle, for example, and you will find that it comes off with more speed than that with which it struck. (Special example: if an atom happens to be standing still and the piston hits it, it will certainly move.) So the atoms are "hotter" when they come away from the piston than they were before they struck it. Therefore all the atoms which are in the vessel will have picked up speed. This means that *when we compress a gas slowly, the temperature of the gas increases*. So, under slow compression, a gas will increase in temperature, and under slow expansion it will decrease in temperature.

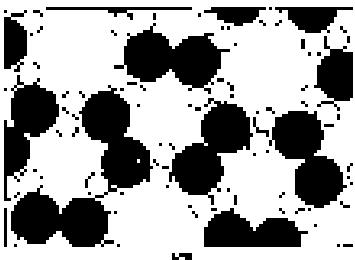


Figure 1-4

We now return to our drop of water and look in another direction. Suppose that we decrease the temperature of our drop of water. Suppose that the jiggling of the molecules of the atoms in the water is steadily decreasing. We know that there are forces of attraction between the atoms, so that after a while they will not be able to jiggle so well. What will happen at very low temperatures is indicated in Fig. 1-4: the molecules lock into a new pattern which is *ice*. This particular schematic diagram of ice is wrong because it is in two dimensions, but it is right qualitatively. The interesting point is that the material has a *definite place for every atom*, and you can easily appreciate that if somehow or other we were to hold all the atoms at one end of the drop in a certain arrangement, each atom in a certain place, then because of the structure of interconnections, which is rigid, the other end miles away (at our magnified scale) will have a definite location. So if we hold a needle of ice at one end, the other end resists our pushing it aside, unlike the case of water, in which the structure is broken down because of the increased jiggling so that the atoms all move around in different ways. The difference between solids and liquids is, then, that in a solid the atoms are arranged in some kind of an array, called a *crystalline array*, and they do not have a random position at long distances; the position of the atoms on one side of the crystal is determined by that of other atoms millions of atoms away on the other side of the crystal. Figure 1-4 is an invented arrangement for ice, and although it contains many of the correct features of ice, it is not the true arrangement. One of the correct features is that there is a part of the symmetry that is hexagonal. You can see that if we turn the picture around an axis by 120° , the picture returns to itself. So there is a *symmetry* in the ice which accounts for the six-sided appearance of snowflakes. Another thing we can see from Fig. 1-4 is why ice shrinks when it melts. The particular crystal pattern of ice shown here has many "holes" in it, as does the true ice structure. When the organization breaks down, these holes can be occupied by molecules. Most simple substances, with the exception of water and type metal, *expand* upon melting, because the atoms are closely packed in the solid crystal and upon melting need more room to jiggle around, but an open structure collapses, as in the case of water.

Now although ice has a "rigid" crystalline form, its temperature can change—ice has heat. If we wish, we can change the amount of heat. What is the heat in

the case of ice? The atoms are not standing still. They are jiggling and vibrating. So even though there is a definite order to the crystal—a definite structure—all of the atoms are vibrating "in place." As we increase the temperature, they vibrate with greater and greater amplitude, until they shake themselves out of place. We call this *melting*. As we decrease the temperature, the vibration decreases and decreases until, at absolute zero, there is a minimum amount of vibration that the atoms can have, but *not zero*. This minimum amount of motion that atoms can have is not enough to melt a substance, with one exception: helium. Helium merely decreases the atomic motions as much as it can, but even at absolute zero there is still enough motion to keep it from freezing. Helium, even at absolute zero, does not freeze, unless the pressure is made so great as to make the atoms squash together. If we increase the pressure, we *can* make it solidify.

1-3 Atomic processes

So much for the description of solids, liquids, and gases from the atomic point of view. However, the atomic hypothesis also describes *processes*, and so we shall now look at a number of processes from an atomic standpoint. The first process that we shall look at is associated with the surface of the water. What happens at the surface of the water? We shall now make the picture more complicated—and more realistic—by imagining that the surface is in air. Figure 1-5 shows the surface of water in air. We see the water molecules as before, forming a body of liquid water, but now we also see the surface of the water. Above the surface we find a number of things: First of all there are water molecules, as in steam. This is *water vapor*, which is always found above liquid water. (There is an equilibrium between the steam vapor and the water which will be described later.) In addition we find some other molecules—here two oxygen atoms stuck together by themselves, forming an *oxygen molecule*; there two nitrogen atoms also stuck together to make a nitrogen molecule. Air consists almost entirely of nitrogen, oxygen, some water vapor, and lesser amounts of carbon dioxide, argon, and other things. So above the water surface is the air, a gas, containing some water vapor. Now what is happening in this picture? The molecules in the water are always jiggling around. From time to time, one on the surface happens to be hit a little harder than usual, and gets knocked away. It is hard to see that happening in the picture because it is a *still* picture. But we can imagine that one molecule near the surface has just been hit and is flying out, or perhaps another one has been hit and is flying out. Thus, molecule by molecule, the water disappears—it evaporates. But if we *close* the vessel above, after a while we shall find a large number of molecules of water amongst the air molecules. From time to time, one of these vapor molecules comes flying down to the water and gets stuck again. So we see that what looks like a dead, uninteresting thing—a glass of water with a cover, that has been sitting there for perhaps twenty years—really contains a dynamic and interesting phenomenon which is going on all the time. To our eyes, our crude eyes, nothing is changing, but if we could see it a billion times magnified, we would see that from its own point of view it is always changing: molecules are leaving the surface, molecules are coming back.

Why do we see *no change*? Because just as many molecules are leaving as are coming back! In the long run "nothing happens." If we then take the top of the vessel off and blow the moist air away, replacing it with dry air, then the number of molecules leaving is just the same as it was before, because this depends on the jiggling of the water, but the number coming back is greatly reduced because there are so many fewer water molecules above the water. Therefore there are more going out than coming in, and the water evaporates. Hence, if you wish to evaporate water turn on the fan!

Here is something else: Which molecules leave? When a molecule leaves it is due to an accidental, extra accumulation of a little bit more than ordinary energy, which it needs if it is to break away from the attractions of its neighbors. Therefore, since those that leave have more energy than the average, the ones that are left have *less* average motion than they had before. So the liquid gradually

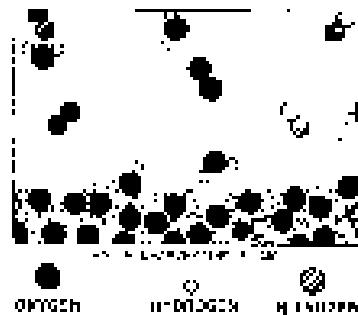


Figure 1-5

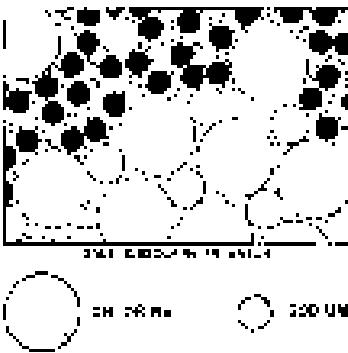


Figure 1-6

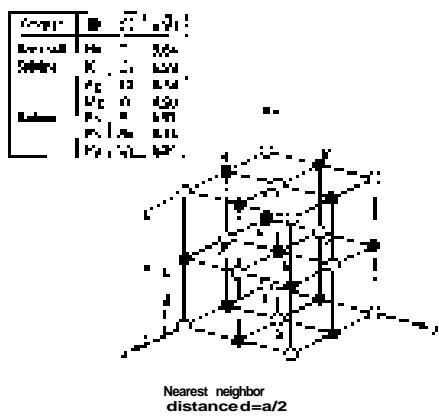


Figure 1-7

cools if it evaporates. Of course, when a molecule of vapor comes from the air to the water below there is a sudden great attraction as the molecule approaches the surface. This speeds up the incoming molecule and results in generation of heat. So when they leave they take away heat; when they come back they generate heat. Of course when there is no net evaporation the result is nothing—the water is not changing temperature. If we blow on the water so as to maintain a continuous preponderance in the number evaporating, then the water is cooled. Hence, blow on soup to cool it!

Of course you should realize that the processes just described are more complicated than we have indicated. Not only does the water go into the air, but also, from time to time, one of the oxygen or nitrogen molecules will come in and "get lost" in the mass of water molecules, and work its way into the water. Thus the air dissolves in the water; oxygen and nitrogen molecules will work their way into the water and the water will contain air. If we suddenly take the air away from the vessel, then the air molecules will leave more rapidly than they come in, and in doing so will make bubbles. This is very bad for divers, as you may know.

Now we go on to another process. In Fig. 1-6 we see, from an atomic point of view, a solid dissolving in water. If we put a crystal of salt in the water, what will happen? Salt is a solid, a crystal, an organized arrangement of "salt atoms." Figure 1-7 is an illustration of the three-dimensional structure of common salt, sodium chloride. Strictly speaking, the crystal is not made of atoms, but of what we call *ions*. An ion is an atom which either has a few extra electrons or has lost a few electrons. In a salt crystal we find chlorine ions (chlorine atoms with an extra electron) and sodium ions (sodium atoms with one electron missing). The ions all stick together by electrical attraction in the solid salt, but when we put them in the water we find, because of the attractions of the negative oxygen and positive hydrogen for the ions, that some of the ions jiggle loose. In Fig. 1-6 we see a chlorine ion getting loose, and other atoms floating in the water in the form of ions. This picture was made with some care. Notice, for example, that the hydrogen ends of the water molecules are more likely to be near the chlorine ion, while near the sodium ion we are more likely to find the oxygen end, because the sodium is positive and the oxygen end of the water is negative, and they attract electrically. Can we tell from this picture whether the salt is *dissolving in water* or *crystallizing out of water*? Of course we *cannot* tell, because while some of the atoms are leaving the crystal other atoms are rejoining it. The process is a *dynamic* one, just as in the case of evaporation, and it depends on whether there is more or less salt in the water than the amount needed for equilibrium. By equilibrium we mean that situation in which the rate at which atoms are leaving just matches the rate at which they are coming back. If there is almost no salt in the water, more atoms leave than return, and the salt dissolves. If, on the other hand, there are too many "salt atoms," more return than leave, and the salt is crystallizing.

In passing, we mention that the concept of a *molecule* of a substance is only approximate and exists only for a certain class of substances. It is clear in the case of water that the three atoms are actually stuck together. It is not so clear in the case of sodium chloride in the solid. There is just an arrangement of sodium and chlorine ions in a cubic pattern. There is no natural way to group them as "molecules of salt."

Returning to our discussion of solution and precipitation, if we increase the temperature of the salt solution, then the rate at which atoms are taken away is increased, and so is the rate at which atoms are brought back. It turns out to be very difficult, in general, to predict which way it is going to go, whether more or less of the solid will dissolve. Most substances dissolve more, but some substances dissolve less, as the temperature increases.

1-4 Chemical reactions

In all of the processes which have been described so far, the atoms and the ions have not changed partners, but of course there are circumstances in which the atoms do change combinations, forming new molecules. This is illustrated in

Fig. 1-8. A process in which the rearrangement of the atomic partners occurs is what we call a *chemical reaction*. The other processes so far described are called physical processes, but there is no sharp distinction between the two. (Nature does not care what we call it, she just keeps on doing it.) This figure is supposed to represent carbon burning in oxygen. In the case of oxygen, *two* oxygen atoms stick together very strongly. (Why do not *three* or *even four* stick together? That is one of the very peculiar characteristics of such atomic processes. Atoms are very special: they like certain particular partners, certain particular directions, and so on. It is the job of physics to analyze why each one wants what it wants. At any rate, two oxygen atoms form, saturated and happy, a molecule.)

The carbon atoms are supposed to be in a solid crystal (which could be graphite or diamond*). Now, for example, one of the oxygen molecules can come over to the carbon, and each atom can pick up a carbon atom and go flying off in a new combination—"carbon-oxygen"—which is a molecule of the gas called carbon monoxide. It is given the chemical name CO. It is very simple: the letters "CO" are practically a picture of that molecule. But carbon attracts oxygen much more than oxygen attracts oxygen or carbon attracts carbon. Therefore in this process the oxygen may arrive with only a little energy, but the oxygen and carbon will snap together with a tremendous vengeance and commotion, and everything near them will pick up the energy. A large amount of motion energy, kinetic energy, is thus generated. This of course is *burning*; we are getting *heat* from the combination of oxygen and carbon. The heat is ordinarily in the form of the molecular motion of the hot gas, but in certain circumstances it can be so enormous that it generates *light*. That is how one *gets flames*.

In addition, the carbon monoxide is not quite satisfied. It is possible for it to attach another oxygen, so that we might have a much more complicated reaction in which the oxygen is combining with the carbon, while at the same time there happens to be a collision with a carbon monoxide molecule. One oxygen atom could attach itself to the CO and ultimately form a molecule, composed of one carbon and two oxygens, which is designated CO₂ and called carbon dioxide. If we burn the carbon with very little oxygen in a very rapid reaction (for example, in an automobile engine, where the explosion is so fast that there is not time for it to make carbon dioxide) a considerable amount of carbon monoxide is formed. In many such rearrangements, a very large amount of energy is released, forming explosions, flames, etc., depending on the reactions. Chemists have studied these arrangements of the atoms, and found that every substance is some type of *arrangement of atoms*.

To illustrate this idea, let us consider another example. If we go into a field of small violets, we know what "that smell" is. It is some kind of *molecule*, or arrangement of atoms, that has worked its way into our noses. First of all, *how* did it work its way in? That is rather easy. If the smell is some kind of molecule in the air, jiggling around and being knocked every which way, it might have *accidentally* worked its way into the nose. Certainly it has no particular desire to get into our nose. It is merely one helpless part of a jostling crowd of molecules, and in its aimless wanderings this particular chunk of matter happens to find itself in the nose.

Now chemists can take special molecules like the odor of violets, and analyze them and tell us the *exact arrangement* of the atoms in space. We know that the carbon dioxide molecule is straight and symmetrical: O—C—O. (That can be determined easily, too, by physical methods.) However, even for the vastly more complicated arrangements of atoms that there are in chemistry, one can, by a long, remarkable process of detective work, find the arrangements of the atoms. Figure 1-9 is a picture of the air in the neighborhood of a violet; again we find nitrogen and oxygen in the air, and water vapor. (Why is there water vapor? Because the violet is *wet*. All plants transpire.) However, we also see a "monster" composed of carbon atoms, hydrogen atoms, and oxygen atoms, which have picked a certain particular pattern in which to be arranged. It is a much more complicated arrange-

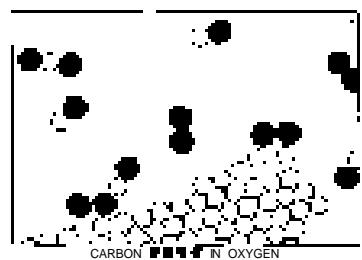


Figure 1-8

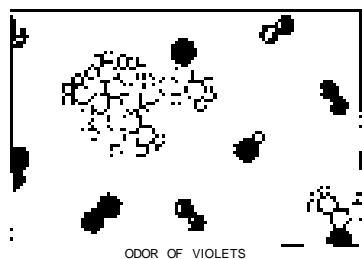


Figure 1-9

*One can burn a diamond in air.

ment than that of carbon dioxide; in fact, it is an enormously complicated arrangement. Unfortunately, we cannot picture all that is really known about it chemically, because the precise arrangement of all the atoms is actually known in three dimensions, while our picture is in only two dimensions. The six carbons which form a ring do not form a flat ring, but a kind of "puckered" ring. All of the angles and distances are known. So a chemical *formula* is merely a picture of such a molecule. When the chemist writes such a thing on the blackboard, he is trying to "draw," roughly speaking, in two dimensions. For example, we see a "ring" of six carbons, and a "chain" of carbons hanging on the end, with an oxygen second from the end, three hydrogens tied to that carbon, two carbons and three hydrogens sticking up here, etc.

How does the chemist find what the arrangement is? He mixes bottles full of stuff together, and if it turns red, it tells him that it consists of one hydrogen and two carbons tied on here; if it turns blue, on the other hand, that is not the way it is at all. This is one of the most fantastic pieces of detective work that has ever been done—organic chemistry. To discover the arrangement of the atoms in these enormously complicated arrays the chemist looks at what happens when he mixes two ~~different~~ substances together. The physicist could never quite believe that the chemist knew what he was talking about when he described the arrangement of the atoms. For about twenty years it has been possible, in some cases, to look at such molecules (not quite as complicated as this one, but some which contain parts of it) by a physical method, and it has been possible to locate every atom, not by looking at colors, but by *measuring where they are*. And lo and behold!, the chemists are almost always correct.

It turns out, in fact, that in the odor of violets there are three slightly different molecules, which differ only in the arrangement of the hydrogen atoms.

One problem of chemistry is to name a substance, so that we will know what it is. Find a name for this shape! Not only must the name tell the shape, but it must also tell that here is an oxygen atom, there a hydrogen—exactly what and where each atom is. So we can appreciate that the chemical names must be complex in order to be complete. You see that the name of this thing in the more complete form that will tell you the structure of it is 4-(2, 2, 3, 6 tetramethyl-5-cyclohexanyl)-3-buten-2-one, and that tells you that this is the arrangement. We can appreciate the difficulties that the chemists have, and also appreciate the reason for such long names. It is not that they wish to be obscure, but they have an extremely difficult problem in trying to describe the molecules in words!

How do we *know* that there are atoms? By one of the tricks mentioned earlier: we make the *hypothesis* that there are atoms, and one after the other results come out the way we predict, as they ought to if things *are* made of atoms. There is also somewhat more direct evidence, a good example of which is the following: The atoms are so small that you cannot see them with a light microscope—in fact, not even with an *electron* microscope. (With a light microscope you can only see things which are much bigger.) Now if the atoms are always in motion, say in water, and we put a big ball of something in the water, a ball much bigger than the atoms, the ball will jiggle around—much as in a push ball game, where a great big ball is pushed around by a lot of people. The people are pushing in various directions, and the ball moves around the field in an irregular fashion. So, in the same way, the "large ball" will move because of the inequalities of the collisions on one side to the other, from one moment to the next. Therefore, if we look at very tiny particles (colloids) in water through an excellent microscope, we see a perpetual jiggling of the particles, which is the result of the bombardment of the atoms. This is called the *Brownian motion*.

We can see further evidence for atoms in the structure of crystals. In many cases the structures deduced by x-ray analysis agree in their spatial "shapes" with the forms actually exhibited by crystals as they occur in nature. The angles between the various "faces" of a crystal agree, within seconds of arc, with angles deduced on the assumption that a crystal is made of many "layers" of atoms.

Everything is made of atoms. That is the key hypothesis. The most important hypothesis in all of biology, for example, is that *everything that animals do, atoms*

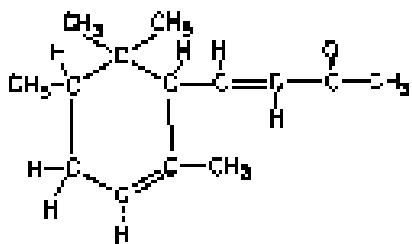


Fig. 1-10. The substance pictured is α-irone.

do. In other words, *there is nothing that living things do that cannot be understood from the point of view that they are made of atoms acting according to the laws of physics.* This was not known from the beginning: it took some experimenting and theorizing to suggest this hypothesis, but now it is accepted, and it is the most useful theory for producing new ideas in the field of biology.

If a piece of steel or a piece of salt, consisting of atoms one next to the other, can have such interesting properties; if water—which is nothing but these little blobs, mile upon mile of the same thing over the earth—can form waves and foam, and make rushing noises and strange patterns as it runs over cement; if all of this, all the life of a stream of water, can be nothing but a pile of atoms, *how much more is possible?* If instead of arranging the atoms in some definite pattern, again and again repeated, on and on, or even forming little lumps of complexity like the odor of violets, we make an arrangement which is *always different* from place to place, with different kinds of atoms arranged in many ways, continually changing, not repeating, how much more marvelously is it possible that this thing might behave? Is it possible that that "thing" walking back and forth in front of you, talking to you, is a great glob of these atoms in a very complex arrangement, such that the sheer complexity of it staggers the imagination as to what it can do? When we say we are a pile of atoms, we do not mean we are *merely* a pile of atoms, because a pile of atoms which is not repeated from one to the other might well have the possibilities which you see before you in the mirror.

Basic Physics

2-1 Introduction

In this chapter, we shall examine the most fundamental ideas that we have about physics—the nature of things as we see them at the present time. We shall not discuss the history of how we know that all these ideas are true; you will learn these details in due time.

The things with which we concern ourselves in science appear in myriad forms, and with a multitude of attributes. For example, if we stand on the shore and look at the sea, we see the water, the waves breaking, the foam, the sloshing motion of the water, the sound, the air, the winds and the clouds, the sun and the blue sky, and light; there is sand and there are rocks of various hardness and permanence, color and texture. There are animals and seaweed, hunger and disease, and the observer on the beach; there may be even happiness and thought. Any other spot in nature has a similar variety of things and influences. It is always as complicated as that, no matter where it is. Curiosity demands that we ask questions, that we try to put things together and try to understand this multitude of aspects as perhaps resulting from the action of a relatively small number of elemental things and forces acting in an infinite variety of combinations.

For example: Is the sand other than the rocks? That is, is the sand perhaps nothing but a great number of very tiny stones? Is the moon a great rock? If we understood rocks, would we also understand the sand and the moon? Is the wind a sloshing of the air analogous to the sloshing motion of the water in the sea? What common features do different movements have? What is common to different kinds of sound? How many different colors are there? And so on. In this way we try gradually to analyze all things, to put together things which at first sight look different, with the hope that we may be able to *reduce* the number of *different* things and thereby understand them better.

A few hundred years ago, a method was devised to find partial answers to such questions. *Observation, reason, and experiment* make up what we call the *scientific method*. We shall have to limit ourselves to a bare description of **our** basic view of what is sometimes called *fundamental physics*, or fundamental ideas which have arisen from the application of the scientific method.

What do we mean by "understanding" something? We can imagine that this complicated array of moving things which constitutes "the world" is something like a great chess game being played by the gods, and we are observers of the game. We do not know what the rules of the game are; all we are allowed to do is to *watch* the playing. Of course, if we watch long enough, we may eventually catch on to a few of the rules. *The rules of the game* are what we mean by *fundamental physics*. Even if we knew every rule, however, we might not be able to understand why a particular move is made in the game, merely because it is too complicated and our minds are limited. If you play chess you must know that it is easy to learn all the rules, and yet it is often very hard to select the best move or to understand why a player moves as he does. So it is in nature, only much more so; but we may be able at least to find all the rules. Actually, we do not have all the rules now. (Every once in a while something like castling is going on that we still do not understand.) Aside from not knowing all of the rules, what we really can explain in terms of those rules is very limited, because almost all situations are so enormously complicated that we cannot follow the plays of the game using the rules, much less tell what is going to happen next. We must, therefore, limit ourselves to the more basic question of the rules of the game. If we know the rules, we consider that we "understand" the world.

2-1 Introduction

2-2 Physics before 1920

2-3 Quantum physics

2-4 Nuclei and particles

How can we tell whether the rules which we "guess" at are really right if we cannot analyze the game very well? There are, roughly speaking, three ways. First, there may be situations where nature has arranged, or we arrange nature, to be simple and to have so few parts that we can predict exactly what will happen, and thus we can check how our rules work. (In one corner of the board there may be only a few chess pieces at work, and that we can figure out exactly.)

A second good way to check rules is in terms of less specific rules derived from them. For example, the rule on the move of a bishop on a chessboard is that it moves only on the diagonal. One can deduce, no matter how many moves may be made, that a certain bishop will always be on a red square. So, without being able to follow the details, we can always check our idea about the bishop's motion by finding out whether it is always on a red square. Of course it will be, for a long time, until all of a sudden we find that it is on a *black* square (what happened of course, is that in the meantime it was captured, another pawn crossed for queening, and it turned into a bishop on a black square). That is the way it is in physics. For a long time we will have a rule that works excellently in an over-all way, even when we cannot follow the details, and then some time we may discover a *new rule*. From the point of view of basic physics, the most interesting phenomena are of course in the *new* places, the places where the rules do not work—not the places where they *do* work! That is the way in which we discover new rules.

The third way to tell whether our ideas are right is relatively crude but probably the most powerful of them all. That is, by rough *approximation*. While we may not be able to tell why Alekhine moves *this particular piece*, perhaps we can *roughly* understand that he is gathering his pieces around the king to protect it, more or less, since that is the sensible thing to do in the circumstances. In the same way, we can often understand nature, more or less, without being able to see what *every little piece* is doing, in terms of our understanding of the game.

At first the phenomena of nature were roughly divided into classes, like heat, electricity, mechanics, magnetism, properties of substances, chemical phenomena, light or optics, x-rays, nuclear physics, gravitation, meson phenomena, etc. However, the aim is to see *complete nature* as different aspects of *one set* of phenomena. That is the problem in basic theoretical physics, today—to *find the laws behind experiment*; to *amalgamate these classes*. Historically, we have always been able to amalgamate them, but as time goes on new things are found. We were amalgamating very well, when all of a sudden x-rays were found. Then we amalgamated some more, and mesons were found. Therefore, at any stage of the game, it always looks rather messy. A great deal is amalgamated, but there are always many wires or threads hanging out in all directions. That is the situation today, which we shall try to describe.

Some historic examples of amalgamation are the following. First, take *heat* and *mechanics*. When atoms are in motion, the more motion, the more heat the system contains, and so *heat and all temperature effects can be represented by the laws of mechanics*. Another tremendous amalgamation was the discovery of the relation between electricity, magnetism, and light, which were found to be different aspects of the same thing, which we call today the *electromagnetic field*. Another amalgamation is the unification of chemical phenomena, the various properties of various substances, and the behavior of atomic particles, which is in the *quantum mechanics of chemistry*.

The question is, of course, is it going to be possible to amalgamate *everything*, and merely discover that this world represents different aspects of *one* thing? Nobody knows. All we know is that as we go along, we find that we can amalgamate pieces, and then we find some pieces that do not fit, and we keep trying to put the jigsaw puzzle together. Whether there are a finite number of pieces, and whether there is even a border to the puzzle, is of course unknown. It will never be known until we finish the picture, if ever. What we wish to do here is to see to what extent this amalgamation process has gone on, and what the situation is at present, in understanding basic phenomena in terms of the smallest set of principles. To express it in a simple manner, *what are things made of and how few elements are there?*

2-2 Physics before 1920

It is a little difficult to begin at once with the present view, so we shall first see how things looked in about 1920 and then take a few things out of that picture. Before 1920, our world picture was something like this: The "stage" on which the universe goes is the three-dimensional *space* of geometry, as described by Euclid, and things change in a medium called *time*. The elements on the stage are *particles*, for example the atoms, which have some *properties*. First, the property of inertia: if a particle is moving it keeps on going in the same direction unless *forces* act upon it. The second element, then, is *forces*, which were then thought to be of two varieties: First, an enormously complicated, detailed kind of interaction force which held the various atoms in different combinations in a complicated way, which determined whether salt would dissolve faster or slower when we raise the temperature. The other force that was known was a long-range interaction—a smooth and quiet attraction—which varied inversely as the square of the distance, and was called *gravitation*. This law was known and was very simple. *Why* things remain in motion when they are moving, or *why* there is a law of gravitation was, of course, not known.

A description of nature is what we are concerned with here. From this point of view, then, a gas, and indeed *all* matter, is a myriad of moving particles. Thus many of the things we saw while standing at the seashore can immediately be connected. First the pressure: this comes from the collisions of the atoms with the walls or whatever; the drift of the atoms, if they are all moving in one direction on the average, is wind; the *random* internal motions are the *heat*. There are waves of excess density, where too many particles have collected, and so as they push off they push up piles of particles farther out, and so on. This wave of excess density is *sound*. It is a tremendous achievement to be able to understand so much. Some of these things were described in the previous chapter.

What *kinds* of particles are there? There were considered to be 92 at that time: 92 different kinds of atoms were ultimately discovered. They had different names associated with their chemical properties.

The next part of the problem was, *what are the short-range forces?* Why does carbon attract one oxygen or perhaps two oxygens, but not three oxygens? What is the machinery of interaction between atoms? Is it gravitation? The answer is no. Gravity is entirely too weak. But imagine a force analogous to gravity, varying inversely with the square of the distance, but enormously more powerful and having one difference. In gravity everything attracts everything else, but now imagine that there are *two kinds* of "things," and that this new force (which is the electrical force, of course) has the property that likes *repel* but unlikes *attract*. The "thing" that carries this strong interaction is called *charge*.

Then what do we have? Suppose that we have two unlikes that attract each other, a plus and a minus, and that they stick very close together. Suppose we have another charge some distance away. Would it feel any attraction? It would feel *practically none*, because if the first two are equal in size, the attraction for the one and the repulsion for the other balance out. Therefore there is very little force at any appreciable distance. On the other hand, if we get *very close* with the extra charge, *attraction* arises, because the repulsion of likes and attraction of unlikes will tend to bring unlikes closer together and push likes farther apart. Then the repulsion will be *less* than the attraction. This is the reason why the atoms, which are constituted out of plus and minus electric charges, feel very little force when they are separated by appreciable distance (aside from gravity). When they come close together, they can "see inside" each other and rearrange their charges, with the result that they have a very strong interaction. The ultimate basis of an interaction between the atoms is *electrical*. Since this force is so enormous, all the plusses and all minuses will normally come together in as intimate a combination as they can. All things, even ourselves, are made of fine-grained, enormously strongly interacting plus and minus parts, all neatly balanced out. Once in a while, by accident, we may rub off a few minuses or a few plusses (usually it is easier to rub off minuses), and in those circumstances we find the force of electricity *unbalanced*, and we can then see the effects of these electrical attractions.

To give an idea of how much stronger electricity is than gravitation, consider two grains of sand, a millimeter across, thirty meters apart. If the force between them were not balanced, if everything attracted everything else instead of likes repelling, so that there were no cancellation, how much force would there be? There would be a force of *three million tons* between the two! You see, there is very, *very* little excess or deficit of the number of negative or positive charges necessary to produce appreciable electrical effects. This is, of course, the reason why you cannot see the difference between an electrically charged or uncharged thing—so few particles are involved that they hardly make a difference in the weight or size of an object.

With this picture the atoms were easier to understand. They were thought to have a "nucleus" at the center, which is positively electrically charged and very massive, and the nucleus is surrounded by a certain number of "electrons" which are very light and negatively charged. Now we go a little ahead in our story to remark that in the nucleus itself there were found two kinds of particles, protons and neutrons, almost of the same weight and very heavy. The protons are electrically charged and the neutrons are neutral. If we have an atom with six protons inside its nucleus, and this is surrounded by six electrons (the negative particles in the ordinary world of matter are all electrons, and these are very light compared with the protons and neutrons which make nuclei), this would be atom number six in the chemical table, and it is called carbon. Atom number eight is called oxygen, etc., because the chemical properties depend upon the electrons on the *outside*, and in fact only upon *how many* electrons there are. So the *chemical* properties of a substance depend only on a number, the number of electrons. (The whole list of elements of the chemists really could have been called 1, 2, 3, 4, 5, etc. Instead of saying "carbon," we could say "element six," meaning six electrons, but of course, when the elements were first discovered, it was not known that they could be numbered that way, and secondly, it would make everything look rather complicated. It is better to have names and symbols for these things, rather than to call everything by number.)

More was discovered about the electrical force. The natural interpretation of electrical interaction is that two objects simply attract each other: plus against minus. However, this was discovered to be an inadequate idea to represent it. A more adequate representation of the situation is to say that the existence of the positive charge, in some sense, distorts, or creates a "condition" in space, so that when we put the negative charge in, it feels a force. This potentiality for producing a force is called an *electric field*. When we put an electron in an electric field, we say it is "pulled." We then have two rules: (a) charges make a field, and (b) charges in fields have forces on them and move. The reason for this will become clear when we discuss the following phenomena: If we were to charge a body, say a comb, electrically, and then place a charged piece of paper at a distance and move the comb back and forth, the paper will respond by always pointing to the comb. If we shake it faster, it will be discovered that the paper is a little behind, *there is a delay* in the action. (At the first stage, when we move the comb rather slowly, we find a complication which is *magnetism*. Magnetic influences have to do with *charges in relative motion*, so magnetic forces and electric forces can really be attributed to one field, as two different aspects of exactly the same thing. A changing electric field cannot exist without magnetism.) If we move the charged paper farther out, the delay is greater. Then an interesting thing is observed. Although the forces between two charged objects should go inversely as the *square* of the distance, it is found, when we shake a charge, that the influence extends *very much farther out* than we would guess at first sight. That is, the effect falls off more slowly than the inverse square.

Here is an analogy: If we are in a pool of water and there is a floating cork very close by, we can move it "directly" by pushing the water with another cork. If you looked only at the two *corks*, all you would see would be that one moved immediately in response to the motion of the other—there is some kind of "*interaction*" between them. Of course, what we really do is to disturb the *water*; the *water* then disturbs the other cork. We could make up a "law" that if you pushed

the water a little bit, an object close by in the water would move. If it were farther away, of course, the second cork would scarcely move, for we move the water *locally*. On the other hand, if we jiggle the cork a new phenomenon is involved, in which the motion of the water moves the water there, etc., and *waves* travel away, so that by jiggling, there is an influence *wry much farther out*, an oscillatory influence, that cannot be understood from the direct interaction. Therefore the idea of direct interaction must be replaced with the existence of the water, or—in the electrical case, with what we call the *electromagnetic field*.

The electromagnetic field can carry waves; some of these waves are *light*, others are used in *radio broadcasts*, but the general name is *electromagnetic waves*. These oscillatory waves can have various *frequencies*. The only thing that is really different from one wave to another is the *frequency of oscillation*. If we shake a charge back and forth more and more rapidly, and look at the effects, we get a whole series of different kinds of effects, which are all unified by specifying but one number, the number of oscillations per second. The usual "pickup" that we get from electric currents in the circuits in the walls of a building have a frequency of about one hundred cycles per second. If we increase the frequency to 500 or 1000 kilocycles (1 kilocycle = 1000 cycles) per second, we are "on the air," for this is the frequency range which is used for radio broadcasts. (Of course it has nothing to do with the *air!* We can have radio broadcasts without any air.) If we again increase the frequency, we come into the range that is used for FM and TV. Going still further, we use certain short waves, for example for *radar*. Still higher, and we do not need an instrument to "see" the stuff, we can see it with the human eye. In the range of frequency from 5×10^{14} to 5×10^{15} cycles per second our eyes would see the oscillation of the charged comb, if we could shake it that fast, as red, blue, or violet light, depending on the frequency. Frequencies below this range are called infrared, and above it, ultraviolet. The fact that we can see in a particular frequency range makes that part of the electromagnetic spectrum no more impressive than the other parts from a physicist's standpoint, but from a human standpoint, of course, it *is* more interesting. If we go up even higher in frequency, we get x-rays. X-rays are nothing but very high-frequency light. If we go still higher, we get gamma rays. These two terms, x-rays and gamma rays, are used almost synonymously. Usually electromagnetic rays coming from nuclei are called gamma rays, while those of high energy from atoms are called x-rays, but at the same frequency they are indistinguishable physically, no matter what their source. If we go to still higher frequencies, say to 10^{24} cycles per second, we find that we can make those waves artificially, for example with the synchrotron here at Caltech. We can find electromagnetic waves with stupendously high frequencies—with even a thousand times more rapid oscillation—in the waves found in *cosmic rays*. These waves cannot be controlled by us.

Table 2-1
The Electromagnetic Spectrum

Freq. per sec. in oscillations/sec.	Name	Rough Example
10^2	Electrical disturbance	Wind
$5 \times 10^4 - 10^5$	Radio broadcast	
10^5	FM—TV	
10^{10}	Radar	Waves
$5 \times 10^{14} - 10^{15}$	Light	
10^{17}	X rays	
10^{20}	γ rays, nuclear	
10^{22}	γ rays, "artificial"	Particle
10^{24}	γ rays, in cosmic rays	

2-3 Quantum physics

Having described the idea of the electromagnetic field, and that this field can carry waves, we soon learn that these waves actually behave in a strange way which seems very unwavelike. At higher frequencies they behave much more like *particles!* It is *quantum mechanics*, discovered just after 1920, which explains this strange behavior. In the years before 1920, the picture of space as a three-dimensional space, and of time as a separate thing, was changed by Einstein, first into a combination which we call space-time, and then still further into a *curved* space-time to represent gravitation. So the "stage" is changed into space-time, and gravitation is presumably a modification of space-time. Then it was also found that the rules for the motions of particles were incorrect. The mechanical rules of "inertia" and "forces" are *wrong*—Newton's laws are *wrong*—in the world of atoms. Instead, it was discovered that things on a small scale behave *nothing like* things on a large scale. That is what makes physics difficult—and very interesting. It is hard because the way things behave on a small scale is so "unnatural"; we have no direct experience with it. Here things behave like nothing we know of, so that it is impossible to describe this behavior in any other than analytic ways. It is difficult, and takes a lot of imagination.

Quantum mechanics has many aspects. In the first place, the idea that a particle has a definite location and a definite speed is no longer allowed; that is wrong. To give an example of how wrong classical physics is, there is a rule in quantum mechanics that says that one cannot know both where something is and how fast it is moving. The uncertainty of the momentum and the uncertainty of the position are complementary, and the product of the two is constant. We can write the law like this: $Dx Dp \geq h/2p$, but we shall explain it in more detail later. This rule is the explanation of a very mysterious paradox: if the atoms are made out of plus and minus charges, why don't the minus charges simply sit on top of the plus charges (they attract each other) and get so close as to completely cancel them out? *Why are atoms so big?* Why is the nucleus at the center with the electrons around it? It was first thought that this was because the nucleus was so big; but no, the nucleus is *very small*. An atom has a diameter of about 10^{-8} cm. The nucleus has a diameter of about 10^{-13} cm. If we had an atom and wished to see the nucleus, we would have to magnify it until the whole atom was the size of a large room, and then the nucleus would be a bare speck which you could just about make out with the eye, but very nearly *all the weight* of the atom is in that infinitesimal *nucleus*. What keeps the electrons from simply falling in? This principle: If they were in the nucleus, we would know their position precisely, and the uncertainty principle would then require that they have a very *large* (but uncertain) momentum, i.e., a very large *kinetic energy*. With this energy they would break away from the nucleus. They make a compromise: they leave themselves a little room for this uncertainty and then jiggle with a certain amount of minimum motion in accordance with this rule. (Remember that when a crystal is cooled to absolute zero, we said that the atoms do not stop moving, they still jiggle. Why? If they stopped moving, we would know where they were and that they had zero motion, and that is against the uncertainty principle. We cannot know where they are and how fast they are moving, so they must be continually wiggling in there!)

Another most interesting change in the ideas and philosophy of science brought about by quantum mechanics is this: it is not possible to predict *exactly* what will happen in any circumstance. For example, it is possible to arrange an atom which is ready to emit light, and we can measure when it has emitted light by picking up a photon particle, which we shall describe shortly. We cannot, however, predict *when* it is going to emit the light or, with several atoms, *which one* is going to. You may say that this is because there are some internal "wheels" which we have not looked at closely enough. No, there *are* no internal wheels; nature, as we understand it today, behaves in such a way that *it is fundamentally impossible* to make a precise prediction of *exactly what will happen* in a given experiment. This is a horrible thing; in fact, philosophers have said before that one of the fundamental requisites of science is that whenever you set **up** the same

conditions, the same thing must happen. This is simply *not true*, it is *not* a fundamental condition of science. The fact is that the same thing does not happen, that we can find only an average, statistically, **as to** what happens. Nevertheless, science has not completely collapsed. Philosophers, incidentally, say a great deal about what is *absolutely necessary* for science, and it is always, so far as one can see, rather naive, and probably wrong. For example, some philosopher or other said it is fundamental to the scientific effort that if an experiment is performed in, say, Stockholm, and then the same experiment is done in, say, Quito, the *same results* must occur. That is quite false. It is not necessary that *science* do that; it may be a *fact of experience*, but it is not necessary. For example, if one of the experiments is to look out at the sky and see the aurora borealis in Stockholm, you do not see it in Quito; that is a different phenomenon. "But," you say, "that is something that has to do with the outside; can you close yourself up in a box in Stockholm and pull down the shade and get any difference?" Surely. If we take a pendulum on a universal joint, and pull it out and let go, then the pendulum will swing almost in a plane, but not quite. Slowly the plane keeps changing in Stockholm, but not in Quito. The blinds are down, too. The fact that this happened does not bring on the destruction of science. What *is* the fundamental hypothesis of science, the fundamental philosophy? We stated it in the first chapter: *the sole test of the validity of any idea is experiment*. If it turns out that most experiments work out the same in Quito as they do in Stockholm, then those "most experiments" will be used to formulate some general law, and those experiments which do not come out the same we will say were a result of the environment near Stockholm. We will invent some way to summarize the results of the experiment, and we do not have to be told ahead of time what this way will look like. If we are told that the same experiment will always produce the same result, that is all very well, but if when we try it, it does *not*, then it does *not*. We just have to take what we see, and then formulate all the rest of our ideas in terms of our actual experience.

Returning again to quantum mechanics and fundamental physics, we cannot go into details of the quantum-mechanical principles at this time, of course, because these are rather difficult to understand. We shall assume that they are there, and go on to describe what some of the consequences are. One of the consequences is that things which we used to consider as waves also behave like particles, and particles behave like waves; in fact everything behaves the same way. There is no distinction between a wave and a particle. So quantum mechanics *unifies* the idea of the field and its waves, and the particles, all into one. Now it is true that when the frequency is low, the field aspect of the phenomenon is more evident, or more useful as an approximate description in terms of everyday experiences. But as the frequency increases, the particle aspects of the phenomenon become more evident with the equipment with which we usually make the measurements. In fact, although we mentioned many frequencies, no phenomenon directly involving a frequency has yet been detected above approximately 10^{12} cycles per ~~second~~. We only *deduce* the higher frequencies from the energy of the particles, by a rule which assumes that the particle-wave idea of quantum mechanics is valid.

Thus we have a new view of electromagnetic interaction. We have a new kind of *particle* to add to the electron, the proton, and the neutron. That new particle is called a *photon*. The new view of the interaction of electrons and protons that is electromagnetic theory, but with everything quantum-mechanically correct, is called *quantum electrodynamics*. This fundamental theory of the interaction of light and matter, or electric field and charges, is our greatest success so far in physics. In this one theory we have the basic rules for all ordinary phenomena except for gravitation and nuclear processes. For example, out of quantum electrodynamics come all known electrical, mechanical, and chemical laws: the laws for the collision of billiard balls, the motions of wires in magnetic fields, the specific heat of carbon monoxide, the color of neon signs, the density of salt, and the reactions of hydrogen and oxygen to make water are all consequences of this one law. All these details can be worked out if the situation is simple enough for us to make an approximation, which is almost never, but often we can understand more

or less what is happening. At the present time no exceptions are found to the quantum-electrodynamic laws outside the nucleus, and there we do not know whether there is an exception because we simply do not know what is going on in the nucleus.

In principle, then, quantum electrodynamics is the theory of all chemistry, and of life, if life is ultimately reduced to chemistry and therefore just to physics because chemistry is already reduced (the part of physics which is involved in chemistry being already known). Furthermore, the same quantum electrodynamics, this great thing, predicts a lot of new things. In the first place, it tells the properties of very high-energy photons, gamma rays, etc. It predicted another very remarkable thing: besides the electron, there should be another particle of the same mass, but of opposite charge, called a *positron*, and these two, coming together, could annihilate each other with the emission of light or gamma rays. (After all, light and gamma rays are all the same, they are just different points on a frequency scale.) The generalization of this, that for each particle there is an antiparticle, turns out to be true. In the case of electrons, the antiparticle has another name—it is called a positron, but for most other particles, it is called anti-so-and-so, like antiproton or antineutron. In quantum electrodynamics, *two numbers* are put in and most of the other numbers in the world are supposed to come out. The two numbers that are put in are called the mass of the electron and the charge of the electron. Actually, that is not quite true, for we have a whole set of numbers for chemistry which tells how heavy the nuclei are. That leads us to the next part.

2-4 Nuclei and particles

What are the nuclei made of, and how are they held together? It is found that the nuclei are held together by enormous forces. When these are released, the energy released is tremendous compared with chemical energy, in the same ratio as the atomic bomb explosion is to a TNT explosion, because, of course, the atomic bomb has to do with changes inside the nucleus, while the explosion of TNT has to do with the changes of the electrons on the outside of the atoms. The question is, what are the forces which hold the protons and neutrons together in the nucleus? Just as the electrical interaction can be connected to a particle, a photon, Yukawa suggested that the forces between neutrons and protons also have a field of some kind, and that when this field jiggles it behaves like a particle. Thus there could be some other particles in the world besides protons and neutrons, and he was able to deduce the properties of these particles from the already known characteristics of nuclear forces. For example, he predicted they should have a mass of two or three hundred times that of an electron; and lo and behold, in cosmic rays there was discovered a particle of the right mass! But it later turned out to be the wrong particle. It was called a m-meson, or muon.

However, a little while later, in 1947 or 1948, another particle was found, the p-meson, or pion, which satisfied Yukawa's criterion. Besides the proton and the neutron, then, in order to get nuclear forces we must add the pion. Now, you say, "Oh great!, with this theory we make quantum nucleodynamics using the pions just like Yukawa wanted to do, and see if it works, and everything will be explained." Bad luck. It turns out that the calculations that are involved in this theory are so difficult that no one has ever been able to figure out what the consequences of the theory are, or to check it against experiment, and this has been going on now for almost twenty years!

So we are stuck with a theory, and we do not know whether it is right or wrong, but we do know that it is a *little* wrong, or at least incomplete. While we have been dawdling around theoretically, trying to calculate the consequences of this theory, the experimentalists have been discovering some things. For example, they had already discovered this m-meson or muon, and we do not yet know where it fits. Also, in cosmic rays, a large number of other "extra" particles were found. It turns out that today we have approximately thirty particles, and it is very difficult to understand the relationships of all these particles, and what nature

, wants them for, or what the connections are from one to another. We do not today understand these various particles as different aspects of the same thing, and the fact that we have so many unconnected particles is a representation of the fact that we have so much unconnected information without a good theory. After the great successes of quantum electrodynamics, there is a certain amount of knowledge of nuclear physics which is rough knowledge, sort of half experience and half theory, assuming a type of force between protons and neutrons and seeing what will happen, but not really understanding where the force comes from. Aside from that, we have made very little progress. We have collected an enormous number of chemical elements. In the chemical case, there suddenly appeared a relationship among these elements which was unexpected, and which is embodied in the periodic table of Mendeleev. For example, sodium and potassium are about the same in their chemical properties and are found in the same column in the Mendeleev chart. We have been seeking a Mendeleev-type chart for the new particles. One such chart of the new particles was made independently by Gell-Mann in the U.S.A. and Nishijima in Japan. The basis of their classification is a new number, like the electric charge, which can be assigned to each particle, called its "strangeness," S . This number is conserved, like the electric charge, in reactions which take place by nuclear forces.

In Table 2-2 are listed all the particles. We cannot discuss them much at this stage, but the table will at least show you how much we do not know. Underneath each particle its mass is given in a certain unit, called the Mev. One Mev is equal to 1.782×10^{-27} gram. The reason this unit was chosen is historical, and we shall not go into it now. More massive particles are put higher up on the chart; we see that a neutron and a proton have almost the same mass. In vertical columns we have put the particles with the same electrical charge, all neutral objects in one column, all positively charged ones to the right of this one, and all negatively charged objects to the left.

Particles are shown with a solid line and "resonances" with a dashed one. Several particles have been omitted from the table. These include the important zero-mass, zero-charge particles, the photon and the graviton, which do not fall into the baryon-meson-lepton classification scheme, and also some of the newer resonances (K^* , η , r). The antiparticles of the mesons are listed in the table, but the antiparticles of the leptons and baryons would have to be listed in another table which would look exactly like this one reflected on the zero-charge column. Although all of the particles except the electron, neutrino, photon, graviton, and proton are unstable, decay products have been shown only for the resonances. Strangeness assignments are not applicable for leptons, since they do not interact strongly with nuclei.

All particles which are together with the neutrons and protons are called *baryons*, and the following ones exist: There is a "lambda," with a mass of 1154 Mev, and three others, called sigmas, minus, neutral, and plus, with several masses almost the same. There are groups or multiplets with almost the same mass, within one or two percent. Each particle in a *multiplet* has the same strangeness. The first *multiplet* is the proton-neutron doublet, and then there is a singlet (the lambda) then the sigma triplet, and finally the xi doublet. Very recently, in 1961, even a few more particles were found. Or *are* they particles? They live so short a time, they disintegrate almost instantaneously, as soon as they are formed, that we do not know whether they should be considered as new particles, or some kind of "resonance" interaction of a certain definite energy between the A and T products into which they disintegrate.

In addition to the baryons the other particles which are involved in the nuclear interaction are called *mesons*. There are first the pions, which come in three varieties, positive, negative, and neutral; they form another multiplet. We have also found some new things called A' -mesons, and they occur as a doublet, π^+ and K^0 . Also, every particle has its antiparticle, unless a particle is *its own* antiparticle. For example, the π^- and the π^+ are antiparticles, but the π^0 is its own antiparticle. The K^- and K^+ are antiparticles, and the K^0 and K^0 . In addition, in 1961 we also found some more mesons or *maybe* mesons which disintegrate almost immediately.

Table 2-2. Elementary Particles

MASs	-E	CHARGE	STRANGENESS
π^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 0$
π^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
π^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 0$
η	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 1$
η'	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 1$
Λ	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Σ	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Ξ	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Λ'	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Σ'	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Ξ'	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 0$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 0$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 0$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 0$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 0$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 0$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 0$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 1$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 1$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 1$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 1$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 1$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 1$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 1$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 1$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 1$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 2$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 2$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 2$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 2$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 2$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 2$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 2$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 2$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 2$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 3$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 3$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 3$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 3$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 3$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 3$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 3$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 3$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 3$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 4$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 4$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 4$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 4$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 4$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 4$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 4$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 4$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 4$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 5$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 5$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 5$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 5$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 5$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 5$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 5$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 5$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 5$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 6$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 6$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 6$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 6$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 6$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 6$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 6$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 6$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 6$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 7$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 7$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 7$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 7$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 7$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 7$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 7$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 7$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 7$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 8$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 8$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 8$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 8$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 8$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 8$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 8$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 8$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 8$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 9$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 9$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 9$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 9$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 9$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 9$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 9$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 9$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 9$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 10$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 10$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 10$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 10$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 10$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 10$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 10$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 10$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 10$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 11$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 11$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 11$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 11$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 11$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 11$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 11$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 11$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 11$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 12$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 12$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 12$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 12$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 12$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 12$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 12$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 12$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 12$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 13$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 13$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 13$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 13$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 13$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 13$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 13$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 13$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 13$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 14$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 14$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 14$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 14$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 14$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 14$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 14$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 14$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 14$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 15$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 15$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 15$
Λ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 15$
Σ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 15$
Ξ^+	$\frac{1}{2} +$	$\frac{1}{2} +$	$S = 15$
Λ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 15$
Σ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 15$
Ξ^-	$\frac{1}{2} -$	$\frac{1}{2} -$	$S = 15$
Λ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 16$
Σ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 16$
Ξ^0	$\frac{1}{2} 0$	$\frac{1}{2} 0$	$S = 16$
$\Lambda^+</math$			

A thing called w which goes into three pions has a mass 780 on this scale, and somewhat less certain is an object which disintegrates into two pions. These particles, called mesons and baryons, and the antiparticles of the mesons are on the same chart, but the antiparticles of the baryons must be put on another chart, "reflected" through the charge-zero column.

Just as Mendeleev's chart was very good, except for the fact that there were a number of rare earth elements which were hanging out loose from it, so we have a number of things hanging out loose from this chart—particles which do not interact strongly in nuclei, have nothing to do with a nuclear interaction, and do not have a strong interaction (I mean the powerful kind of interaction of nuclear energy). These are called leptons, and they are the following: there is the electron, which has a very small mass on this scale, only 0.510 Mev. Then there is that other, the μ meson, the muon, which has a mass much higher, 206 times as heavy as an electron. So far as we can tell, by all experiments so far, the difference between the electron and the muon is nothing but the mass. Everything works exactly the same for the muon as for the electron, except that one is heavier than the other. Why is there another one heavier; what is the use for it? We do not know. In addition, there is a lepton which is neutral, called a neutrino, and this particle has zero mass. In fact, it is now known that there are *two* different kinds of neutrinos, one related to electrons and the other related to muons.

Finally, we have two other particles which do not interact strongly with the nuclear ones: one is a photon, and perhaps, if the field of gravity also has a quantum-mechanical analog (a quantum theory of gravitation has not yet been worked out), then there will be a particle, a graviton, which will have zero mass.

What is this "zero mass"? The masses given here are the masses of the particles *at rest*. The fact that a particle has zero mass means, in a way, that it cannot *be at rest*. A photon is never at rest, it is always moving at 186,000 miles a second. We will understand more what mass means when we understand the theory of relativity, which will come in due time.

Thus we are confronted with a large number of particles, which together seem to be the fundamental constituents of matter. Fortunately, these particles are not *all* different in their *interactions* with one another. In fact, there seem to be just *four kinds* of interaction between particles which, in the order of decreasing strength, are the nuclear force, electrical interactions, the beta-decay interaction, and gravity. The photon is coupled to all charged particles and the strength of the interaction is measured by some number, which is 1/137. The detailed law of this coupling is known, that is quantum electrodynamics. Gravity is coupled to all *energy*, but its coupling is extremely weak, much weaker than that of electricity. This law is also known. Then there are the so-called weak decays—beta decay, which causes the neutron to disintegrate into proton, electron, and neutrino, relatively slowly. This law is only partly known. The so-called strong interaction, the meson-baryon interaction, has a strength of 1 in this scale, and the law is completely unknown, although there are a number of known rules, such as that the number of baryons does not change in any reaction.

Table 2-3. Elementary Interactions

Coupling	Strength*	Law
Photon to charged particles	$\sim 10^{-4}$	Law known
Gravity to all energy	$\sim 10^{-5}$	known
Weak decays	Law $\sim 10^{-5}$	known
Mesons to baryons	~ 1	Law unknown (some rules known)

* The "strength" is a dimensionless measure of the coupling constant involved in each interaction (\sim means "approximately").

This then, is the horrible condition of our physics today. To summarize it, I would say this: outside the nucleus, we seem to know all; inside it, quantum mechanics is valid—the principles of quantum mechanics have not been found to fail. The stage on which we put all of our knowledge, we would say, is relativistic space-time; perhaps gravity is involved in space-time. We do not know how the universe got started, and we have never made experiments which check our ideas of space and time accurately, below some tiny distance, so we only *know* that our ideas work above that distance. We should also add that the rules of the game are the quantum-mechanical principles, and those principles apply, so far as we can tell, to the new particles as well as to the old. The origin of the forces in nuclei leads us to new particles, but unfortunately they appear in great profusion and we lack a complete understanding of their interrelationship, although we already know that there are some very surprising relationships among them. We seem gradually to be groping toward an understanding of the world of subatomic particles, but we really do not know how far we have yet to go in this task.

The Relation of Physics to Other Sciences

3-1 Introduction

Physics is the most fundamental and all-inclusive of the sciences, and has had a profound effect on all scientific development. In fact, physics is the present-day equivalent of what used to be called *natural philosophy*, from which most of our modern sciences arose. Students of many fields find themselves studying physics because of the basic role it plays in all phenomena. In this chapter we shall try to explain what the fundamental problems in the other sciences are, but of course it is impossible in so small a space really to deal with the complex, subtle, beautiful matters in these other fields. Lack of space also prevents our discussing the relation of physics to engineering, industry, society, and war, or even the most remarkable relationship between mathematics and physics. (Mathematics is not a science from our point of view, in the sense that it is not a *natural* science. The test of its validity is not experiment.) We must, incidentally, make it clear from the beginning that if a thing is not a science, it is not necessarily bad. For example, love is not a science. So, if something is said not to be a science, it does not mean that there is something wrong with it; it just means that it is not a science.

3-2 Chemistry

The science which is perhaps the most deeply affected by physics is chemistry. Historically, the early days of chemistry dealt almost entirely with what we now call inorganic chemistry, the chemistry of substances which are not associated with living things. Considerable analysis was required to discover the existence of the many elements and their relationships—how they make the various relatively simple compounds found in rocks, earth, etc. This early chemistry was very important for physics. The interaction between the two sciences was very great because the theory of atoms was substantiated to a large extent by experiments in chemistry. The theory of chemistry, i.e., of the reactions themselves, was summarized to a large extent in the periodic chart of Mendeleev, which brings out many strange relationships among the various elements, and it was the collection of rules as to which substance is combined with which, and how, that constituted inorganic chemistry. All these rules were ultimately explained in principle by quantum mechanics, so that theoretical chemistry is in fact physics. On the other hand, it must be emphasized that this explanation is *in principle*. We have already discussed the difference between knowing the rules of the game of chess, and being able to play. So it is that we may know the rules, but we cannot play very well. It turns out to be very difficult to predict precisely what will happen in a given chemical reaction; nevertheless, the deepest part of theoretical chemistry must end up in quantum mechanics.

There is also a branch of physics and chemistry which was developed by both sciences together, and which is extremely important. This is the method of statistics applied in a situation in which there are mechanical laws, which is aptly called *statistical mechanics*. In any chemical situation a large number of atoms are involved, and we have seen that the atoms are all jiggling around in a very random and complicated way. If we could analyze each collision, and be able to follow in detail the motion of each molecule, we might hope to figure out what would happen, but the many numbers needed to keep track of all these molecules exceeds so enormously the capacity of any computer, and certainly the capacity of

3-1 Introduction

3-2 Chemistry

3-3 Biology

3-4 Astronomy

3-5 Geology

3-6 Psychology

3-7 How did it get that way?

the mind, that it was important to develop a method for dealing with such complicated situations. Statistical mechanics, then, is the science of the phenomena of heat, or thermodynamics. Inorganic chemistry is, as a science, now reduced essentially to what are called physical chemistry and quantum chemistry; physical chemistry to study the rates at which reactions occur and what is happening in detail (How do the molecules hit? Which pieces fly off first?, etc.), and quantum chemistry to help us understand what happens in terms of the physical laws.

The other branch of chemistry is *organic chemistry*, the chemistry of the substances which are associated with living things. For a time it was believed that the substances which are associated with living things were so marvelous that they could not be made by hand, from inorganic materials. This is not at all true—they are just the same as the substances made in inorganic chemistry, but more complicated arrangements of atoms are involved. Organic chemistry obviously has a very close relationship to the biology which supplies its substances, and to industry, and furthermore, much physical chemistry and quantum mechanics can be applied to organic as well as to inorganic compounds. However, the main problems of organic chemistry are not in these aspects, but rather in the analysis and synthesis of the substances which are formed in biological systems, in living things. This leads imperceptibly, in steps, toward biochemistry, and then into biology itself, or molecular biology.

3-3 Biology

Thus we come to the science of *biology*, which is the study of living things. In the early days of biology, the biologists had to deal with the purely descriptive problem of finding out *what* living things there were, and so they just had to count such things as the hairs of the limbs of fleas. After these matters were worked out with a great deal of interest, the biologists went into the *machinery* inside the living bodies, first from a gross standpoint, naturally, because it takes some effort to get into the finer details.

There was an interesting early relationship between physics and biology in which biology helped physics in the discovery of the *conservation of energy*, which was first demonstrated by Mayer in connection with the amount of heat taken in and given out by a living creature.

If we look at the processes of biology of living animals more closely, we see many physical phenomena: the circulation of blood, pumps, pressure, etc. There are nerves: we know what is happening when we step on a sharp stone, and that somehow or other the information goes from the leg up. It is interesting how that happens. In their study of nerves, the biologists have come to the conclusion that nerves are very fine tubes with a complex wall which is very thin; through this wall the cell pumps ions, so that there are positive ions on the outside and negative ions on the inside, like a capacitor. Now this membrane has an interesting property; if it "discharges" in one place, i.e., if some of the ions were able to move through one place, so that the electric voltage is reduced there, that electrical influence makes itself felt on the ions in the neighborhood, and it affects the membrane in such a way that it lets the ions through at neighboring points also. This in turn affects it farther along, etc., and so there is a wave of "penetrability" of the membrane which runs down the fiber when it is "excited" at one end by stepping on the sharp stone. This wave is somewhat analogous to a long sequence of vertical dominoes; if the end one is pushed over, that one pushes the next, etc. Of course this will transmit only one message unless the dominoes are set up again; and similarly in the nerve cell, there are processes which pump the ions slowly out again, to get the nerve ready for the next impulse. So it is that we know what we are doing (or at least where we are). Of course the electrical effects associated with this nerve impulse can be picked up with electrical instruments, and because there *are* electrical effects, obviously the physics of electrical effects has had a great deal of influence on understanding the phenomenon.

The opposite effect is that, from somewhere in the brain, a message is sent out along a nerve. What happens at the end of the nerve? There the nerve branches

out into fine little things, connected to a structure near a muscle, called an end-plate. For reasons which are not exactly understood, when the impulse reaches the end of the nerve, little packets of a chemical called acetylcholine are shot off (five or ten molecules at a time) and they affect the muscle fiber and make it contract—how simple! What makes a muscle contract? A muscle is a very large number of fibers close together, containing two different substances, myosin and actomyosin, but the machinery by which the chemical reaction induced by acetylcholine can modify the dimensions of the molecule is not yet known. Thus the fundamental processes in the muscle that make mechanical motions are not known.

Biology is such an enormously wide field that there are hosts of other problems that we cannot mention at all—problems on how vision works (what the light does in the eye), how hearing works, etc. (The way in which *thinking* works we shall discuss later under psychology.) Now, these things concerning biology which we have just discussed are, from a biological standpoint, really not fundamental, at the bottom of life, in the sense that even if we understood them we still would not understand life itself. To illustrate: the men who study nerves feel their work is very important, because after all you cannot have animals without nerves. But you *can* have *life* without nerves. Plants have neither nerves nor muscles, but they are working, they are alive, just the same. So for the fundamental problems of biology we must look deeper; when we do, we discover that all living things have a great many characteristics in common. The most common feature is that they are made of *cells*, within each of which is complex machinery for doing things chemically. In plant cells, for example, there is machinery for picking up light and generating sucrose, which is consumed in the dark to keep the plant alive. When the plant is eaten the sucrose itself generates in the animal a series of chemical reactions very closely related to photosynthesis (and its opposite effect in the dark) in plants.

In the cells of living systems there are many elaborate chemical reactions, in which one compound is changed into another and another. To give some impression of the enormous efforts that have gone into the study of biochemistry, the chart in Fig. 3-1 summarizes our knowledge to date on just one small part of the many series of reactions which occur in cells, perhaps a percent or so of it.

Here we see a whole series of molecules which change from one to another in a sequence or cycle of rather small steps. It is called the Krebs cycle, the respiratory cycle. Each of the chemicals and each of the steps is fairly simple, in terms of what change is made in the molecule, but—and this is a centrally important discovery in biochemistry—these changes are *relatively difficult to accomplish in a laboratory*. If we have one substance and another very similar substance, the one does not just turn into the other, because the two forms are usually separated by

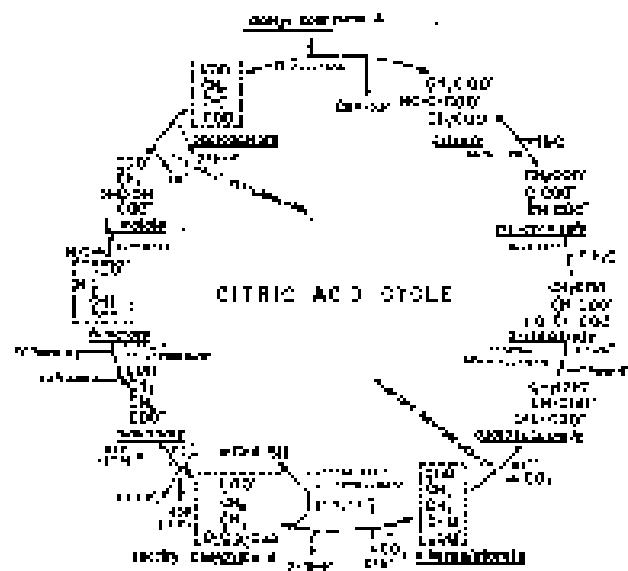


Fig. 3-1. The Krebs cycle.

an energy barrier or "hill." Consider this analogy: If we wanted to take an object from one place to another, at the same level but on the other side of a hill, we could push it over the top, but to do so requires the addition of some energy. Thus most chemical reactions do not occur, because there is what is called an *activation energy* in the way. In order to add an extra atom to our chemical requires that we get it *close* enough that some rearrangement can occur; then it will stick. But if we cannot give it enough energy to get it close enough, it will not go to completion, it will just go part way up the "hill" and back down again. However, if we could literally take the molecules in our hands and push and pull the atoms around in such a way as to open a hole to let the new atom in, and then let it snap back, we would have found another way, *around* the hill, which would not require extra energy, and the reaction would go easily. Now there actually *are*, in the cells, *very* large molecules, much larger than the ones whose changes we have been describing, which in some complicated way hold the smaller molecules just right, so that the reaction can occur easily. These very large and complicated things are called *enzymes*. (They were first called fermenters, because they were originally discovered in the fermentation of sugar. In fact, some of the first reactions in the cycle were discovered there.) In the presence of an enzyme the reaction will go.

An enzyme is made of another substance called *protein*. Enzymes are very big and complicated, and each one is different, each being built to control a certain special reaction. The names of the enzymes are written in Fig. 3-1 at each reaction. (Sometimes the same enzyme may control two reactions.) We emphasize that the enzymes themselves are not involved in the reaction directly. They do not change; they merely let an atom go from one place to another. Having done so, the enzyme is ready to do it to the next molecule, like a machine in a factory. Of course, there must be a supply of certain atoms and a way of disposing of other atoms. Take hydrogen, for example: there are enzymes which have special units on them which carry the hydrogen for all chemical reactions. For example, there are three or four hydrogen-reducing enzymes which are used all over our cycle in different places. It is interesting that the machinery which liberates some hydrogen at one place will take that hydrogen and use it somewhere else.

The most important feature of the cycle of Fig. 3-1 is the transformation from GDP to GTP (guanadine-di-phosphate to guanadine-tri-phosphate) because the one substance has much more energy in it than the other. Just as there is a "box" in certain enzymes for carrying hydrogen atoms around, there are special *energy-carrying* "boxes" which involve the triphosphate group. So, GTP has more energy than GDP and if the cycle is going one way, we are producing molecules which have extra energy and which can go drive some other cycle which *requires* energy, for example the contraction of muscle. The muscle will not contract unless there is GTP. We can take muscle fiber, put it in water, and add GTP, and the fibers contract, changing GTP to GDP if the right enzymes are present. So the real system is in the GDP-GTP transformation; in the dark the GTP which has been stored up during the day is used to run the whole cycle around the other way. An enzyme you see, does not care in which direction the reaction goes, for if it did it would violate one of the laws of physics.

Physics is of great importance in biology and other sciences for still another reason, that has to do with *experimental techniques*. In fact, if it were not for the great development of experimental physics, these biochemistry charts would not be known today. The reason is that the most useful tool of all for analyzing this fantastically complex system is to *label* the atoms which are used in the reactions. Thus, if we could introduce into the cycle some carbon dioxide which has a "green mark" on it, and then measure after three seconds where the green mark is, and again measure after ten seconds, etc., we could trace out the course of the reactions. What are the "green marks"? They are different *isotopes*. We recall that the chemical properties of atoms are determined by the number of ~~electrons~~, not by the mass of the nucleus. But there can be, for example in carbon, six neutrons or seven neutrons, together with the six protons which all carbon nuclei have. Chemically, the two atoms C₁₂ and C₁₃ are the same, but they differ in weight and they have different nuclear properties, and so they are distinguishable.

By using these isotopes of different weights, or even radioactive isotopes like C₁₄, which provide a more sensitive means for tracing very small quantities, it is possible to trace the reactions.

Now, we return to the description of enzymes and proteins. All proteins are not enzymes, but all enzymes are proteins. There are many proteins, such as the proteins in muscle, the structural proteins which are, for example, in cartilage and hair, skin, etc., that are not themselves enzymes. However, proteins are a very characteristic substance of life: first of all they make up all the enzymes, and second, they make up much of the rest of living material. Proteins have a very interesting and simple structure. They are a series, or chain, of different *amino acids*. There are twenty different amino acids, and they all can combine with each other to form chains in which the backbone is CO-NH, etc. Proteins are nothing but chains of various ones of these twenty amino acids. Each of the amino acids probably serves some special purpose. Some, for example, have a sulphur atom at a certain place; when two sulphur atoms are in the same protein, they form a bond, that is, they tie the chain together at two points and form a loop. Another has extra oxygen atoms which make it an acidic substance, another has a basic characteristic. Some of them have big groups hanging out to one side, so that they take up a lot of space. One of the amino acids, called proline, is not really an amino acid, but imino acid. There is a slight difference, with the result that when proline is in the chain, there is a kink in the chain. If we wished to manufacture a particular protein, we would give these instructions: put one of those sulphur hooks here; next, add something to take up space; then attach something to put a kink in the chain. In this way, we will get a complicated-looking chain, hooked together and having some complex structure; this is presumably just the manner in which all the various enzymes are made. One of the great triumphs in recent times (since 1960), was at last to discover the exact spatial atomic arrangement of certain proteins, which involve some fifty-six or sixty amino acids in a row. Over a thousand atoms (more nearly two thousand, if we count the hydrogen atoms) have been located in a complex pattern in two proteins. The first was hemoglobin. One of the sad aspects of this discovery is that we cannot see anything from the pattern; we do not understand why it works the way it does. Of course, that is the next problem to be attacked.

Another problem is how do the enzymes know what to be? A red-eyed fly makes a red-eyed fly baby, and so the information for the whole pattern of enzymes to make red pigment must be passed from one fly to the next. This is done by a substance in the nucleus of the cell, not a protein, called DNA (short for deoxyribose nucleic acid). This is the key substance which is passed from one cell to another (for instance, sperm cells consist mostly of DNA) and carries the information as to how to make the enzymes. DNA is the "blueprint." What does the blueprint look like and how does it work? First, the blueprint must be able to reproduce itself. Secondly, it must be able to instruct the protein. Concerning the reproduction, we might think that this proceeds like cell reproduction. Cells simply grow bigger and then divide in half. Must it be thus with DNA molecules, then, that they too grow bigger and divide in half? Every *atom* certainly does not grow bigger and divide in half! No, it is impossible to reproduce a molecule except by some more clever way.

The structure of the substance DNA was studied for a long time, first chemically to find the composition, and then with x-rays to find the pattern in space. The result was the following remarkable discovery: The DNA molecule is a pair of chains, twisted upon each other. The backbone of each of these chains, which are analogous to the chains of proteins but chemically quite different, is a series of sugar and phosphate groups, as shown in Fig. 3-2. Now we see how the chain can contain instructions, for if we could split this chain down the middle, we would have a series BAADC . . . and every living thing could have a different series. Thus perhaps, in some way, the specific *instructions* for the manufacture of proteins are contained in the specific *series* of the DNA.

Attached to each sugar along the line, and linking the two chains together, are certain pairs of cross-links. However, they are not all of the same kind; there are

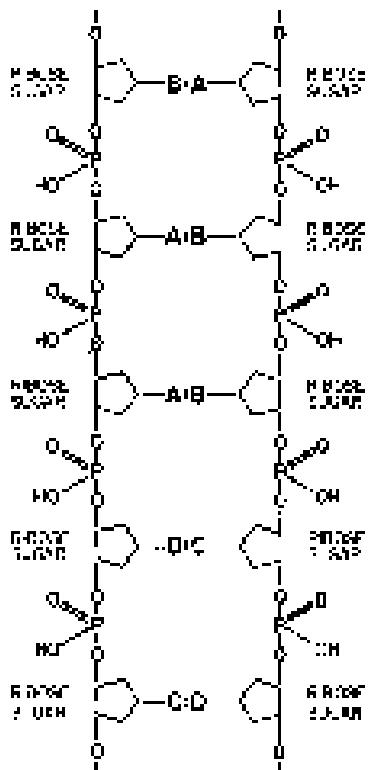


Fig. 3-2. Schematic diagram of DNA.

four kinds, called adenine, thymine, cytosine, and guanine, but let us call them *A*, *B*, *C*, and *D*. The interesting thing is that only certain pairs can sit opposite each other, for example *A* with *B* and *C* with *D*. These pairs are put on the two chains in such a way that they "fit together," and have a strong energy of interaction. However, *C* will not fit with *A*, and *B* will not fit with *C*; they will only fit in pairs, *A* against *B* and *C* against *D*. Therefore if one is *C*, the other must be *D*, etc. Whatever the letters may be in one chain, each one must have its specific complementary letter on the other chain.

What then about reproduction? Suppose we split this chain in two. How can we make another one just like it? If, in the substances of the cells, there is a manufacturing department which brings up phosphate, sugar, and *A*, *B*, *C*, *D* units not connected in a chain, the only ones which will attach to our split chain will be the correct ones, the complements of *BAADC* . . . , namely, *ABBCD* . . . Thus what happens is that the chain splits down the middle during cell division, one half ultimately to go with one cell, the other half to end up in the other cell; when separated, a new complementary chain is made by each half-chain.

Next comes the question, precisely how does the order of the *A*, *B*, *C*, *D* units determine the arrangement of the amino acids in the protein? This is the central unsolved problem in biology today. The first clues, or pieces of information, however, are these: There are in the cell tiny particles called microsomes, and it is now known that that is the place where proteins are made. But the microsomes are not in the nucleus, where the DNA and its instructions are. Something seems to be the matter. However, it is also known that little molecule pieces come off the DNA—not as long as the big DNA molecule that carries all the information itself, but like a small section of it. This is called RNA, but that is not essential. It is a kind of copy of the DNA, a short copy. The RNA, which somehow carries a message as to what kind of protein to make goes over to the microsome; that is known. When it gets there, protein is synthesized at the microsome. That is also known. However, the details of how the amino acids come in and are arranged in accordance with a code that is on the RNA are, as yet, still unknown. We do not know how to read it. If we knew, for example, the "lineup" *A*, *B*, *C*, *C*, *A*, we could not tell you what protein is to be made.

Certainly no subject or field is making more progress on so many fronts at the present moment, than biology, and if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that *all things are made of atoms*, and that everything that living things do can be understood in terms of the jigglings and wiggles of atoms.

3-4 Astronomy

In this rapid-fire explanation of the whole world, we must now turn to astronomy. Astronomy is older than physics. In fact, it got physics started by showing the beautiful simplicity of the motion of the stars and planets, the understanding of which was the *beginning* of physics. But the most remarkable discovery in all of astronomy is that *the stars are made of atoms of the same kind as those on the earth**. How was this done? Atoms liberate light which has definite fre-

* How I'm rushing through this! How much each sentence in this brief story contains. "The stars are made of the same atoms as the earth." I usually pick one small topic like this to give a lecture on. Poets say science takes away from the beauty of the stars—mere globs of gas atoms. Nothing is "mere." I too can see the stars on a desert night, and feel them. But do I see less or more? The vastness of the heavens stretches my imagination—stuck on this carousel my little eye can catch one-million-year-old light. A vast pattern—of which I am a part—perhaps my stuff was belched from some forgotten star, as one is belching there. Or see them with the greater eye of Palomar, rushing all apart from some common starting point when they were perhaps all together. What is the pattern, or the meaning, or the *why*? It does not do harm to the mystery to know a little about it. For far more marvelous is the truth than any artists of the past imagined! Why do the poets of the present not speak of it? What men are poets who can speak of Jupiter if he were like a man, but if he is an immense spinning sphere of methane and ammonia must be silent?

quencies, something like the timbre of a musical instrument, which has definite pitches or frequencies of sound. When we are listening to several different tones we can tell them apart, but when we look with our eyes at a mixture of colors we cannot tell the parts from which it was made, because the eye is nowhere near as discerning as the ear in this connection. However, with a spectroscope we *can* analyze the frequencies of the light waves and in this way we can see the very tunes of the atoms that are in the different stars. As a matter of fact, two of the chemical elements were discovered on a star before they were discovered on the earth. Helium was discovered on the sun, whence its name, and technetium was discovered in certain cool stars. This, of course, permits us to make headway in understanding the stars, because they are made of the same kinds of atoms which are on the earth. Now we know a great deal about the atoms, especially concerning their behavior under conditions of high temperature but not very great density, so that we can analyze by statistical mechanics the behavior of the stellar substance. Even though we cannot reproduce the conditions on the earth, using the basic physical laws we often can tell precisely, or very closely, what will happen. So it is that physics aids astronomy. Strange as it may seem, we understand the distribution of matter in the interior of the sun far better than we understand the interior of the earth. What goes on *inside* a star is better understood than one might guess from the difficulty of having to look at a little dot of light through a telescope, because we can *calculate* what the atoms in the stars should do in most circumstances.

One of the most impressive discoveries was the origin of the energy of the stars, that makes them continue to burn. One of the men who discovered this was out with his girl friend the night after he realized that *nuclear reactions* must be going on in the stars in order to make them shine. She said "Look at how pretty the stars shine!" He said "Yes, and right now I am the only man in the world who knows *why* they shine." She merely laughed at him. She was not impressed with being out with the only man who, at that moment, knew why stars shine. Well, it is sad to be alone, but that is the way it is in this world.

It is the nuclear "burning" of hydrogen which supplies the energy of the sun; the hydrogen is converted into helium. Furthermore, ultimately, the manufacture of various chemical elements proceeds in the centers of the stars, from hydrogen. The stuff of which *we* are made, was "cooked" once, in a star, and spit out. How do we know? Because there is a clue. The proportion of the different isotopes—how much C¹², how much C¹³, etc., is something which is never changed by *chemical* reactions, because the chemical reactions are so much the same for the two. The proportions are purely the result of *nuclear* reactions. By looking at the proportions of the isotopes in the cold, dead ember which we are, we can discover what the *furnace* was like in which the stuff of which we are made was formed. That furnace was like the stars, and so it is very likely that our elements were "made" in the stars and spit out in the explosions which we call novae and supernovae. Astronomy is so close to physics that we shall study many astronomical things as we go along.

3-5 Geology

We turn now to what are called *earth sciences*, or *geology*. First, meteorology and the weather. Of course the *instruments* of meteorology are physical instruments, and the development of experimental physics made these instruments possible, as was explained before. However, the theory of meteorology has never been satisfactorily worked out by the physicist. "Well," you say, "there is nothing but air, and we know the equations of the motions of air." Yes we do. "So if we know the condition of air today, why can't we figure out the condition of the air tomorrow?" First, we do not *really* know what the condition is today, because the air is swirling and twisting everywhere. It turns out to be very sensitive, and even unstable. If you have ever seen water run smoothly over a dam, and then turn into a large number of blobs and drops as it falls, you will understand what I mean by unstable. You know the condition of the water before it goes over the

spillway; it is perfectly smooth; but the moment it begins to fall, where do the drops begin? What determines how big the lumps are going to be and where they will be? That is not known, because the water is unstable. Even a smooth moving mass of air, in going over a mountain turns into complex whirlpools and eddies. In many fields we find this situation of *turbulent flow* that we cannot analyze today. Quickly we leave the subject of weather, and discuss geology!

The question basic to geology is, what makes the earth the way it is? The most obvious processes are in front of your very eyes, the erosion processes of the rivers, the winds, etc. It is easy enough to understand these, but for every bit of erosion there is an equal amount of something else going on. Mountains are no lower today, on the average, than they were in the past. There must be *mountain-forming* processes. You will find, if you study geology, that there *are* mountain-forming processes and vulcanism, which nobody understands but which is half of geology. The phenomenon of volcanoes is really not understood. What makes an earthquake is, ultimately, not understood. It is understood that if something is pushing something else, it snaps and will slide—that is all right. But what pushes, and why? The theory is that there are currents inside the earth—circulating currents, due to the difference in temperature inside and outside—which, in their motion, push the surface slightly. Thus if there are two opposite circulations next to each other, the matter will collect in the region where they meet and make belts of mountains which are in unhappy stressed conditions, and so produce volcanoes and earthquakes.

What about the inside of the earth? A great deal is known about the speed of earthquake waves through the earth and the density of distribution of the earth. However, physicists have been unable to get a good theory as to how dense a substance should be at the pressures that would be expected at the center of the earth. In other words, we cannot figure out the properties of matter very well in these circumstances. We do much less well with the earth than we do with the conditions of matter in the stars. The mathematics involved seems a little too difficult, so far, but perhaps it will not be too long before someone realizes that it is an important problem, and really work it out. The other aspect, of course, is that even if we did know the density, we cannot figure out the circulating currents. Nor can we really work out the properties of rocks at high pressure. We cannot tell how fast the rocks should "give"; that must all be worked out by experiment.

3-6 Psychology

Next, we consider the science of *psychology*. Incidentally, psychoanalysis is not a science: it is at best a medical process, and perhaps even more like witch-doctoring. It has a theory as to what causes disease—lots of different "spirits," etc. The witch doctor has a theory that a disease like malaria is caused by a spirit which comes into the air; it is not cured by shaking a snake over it, but quinine does help malaria. So, if you are sick, I would advise that you go to the witch doctor because he is the man in the tribe who knows the most about the disease; on the other hand, his knowledge is not science. Psychoanalysis has not been checked carefully by experiment, and there is no way to find a list of the number of cases in which it works, the number of cases in which it does not work, etc.

The other branches of psychology, which involve things like the physiology of sensation—what happens in the eye, and what happens in the brain—are, if you wish, less interesting. But some small but real progress has been made in studying them. One of the most interesting technical problems may or may not be called psychology. The central problem of the mind, if you will, or the nervous system, is this: when an animal learns something, it can do something different than it could before, and its brain cell must have changed too, if it is made out of atoms. *In what way is it different?* We do not know where to look, or what to look for, when something is memorized. We do not know what it means, or what change there is in the nervous system, when a fact is learned. This is a very important problem which has not been solved at all. Assuming, however, that there is some kind of memory thing, the brain is such an enormous mass of interconnect-

ing wires and nerves that it probably cannot be analyzed in a straightforward manner. There is an analog of this to computing machines and computing elements, in that they also have a lot of lines, and they have some kind of element, analogous, perhaps, to the synapse, or connection of one nerve to another. This is a very interesting subject which we have not the time to discuss further—the relationship between thinking and computing machines. It must be appreciated, of course, that this subject will tell us very little about the real complexities of ordinary human behavior. All human beings are so different. It will be a long time before we get there. We must start much further back. If we could even figure out how a *dog* works, we would have gone pretty far. Dogs are easier to understand, but nobody yet knows how dogs work.

3-7 How did it get that way?

In order for physics to be useful to other sciences in a *theoretical* way, other than in the invention of instruments, the science in question must supply to the physicist a description of the object in a physicist's language. They can say "why does a frog jump?", and the physicist cannot answer. If they tell him what a frog is, that there are so many molecules, there is a nerve here, etc., that is different. If they will tell us, more or less, what the earth or the stars are like, then we can figure it out. In order for physical theory to be of any use, we must know where the atoms are located. In order to understand the chemistry, we must know exactly what atoms are present, for otherwise we cannot analyze it. That is but one limitation, of course.

There is another *kind* of problem in the sister sciences which does not exist in physics; we might call it, for lack of a better term, the historical question. How did it get that way? If we understand all about biology, we will want **to** know how all the things which are on the earth got there. There is the theory of evolution, an important part of biology. In geology, we not only want to know how the mountains are forming, but how the entire earth was formed in the beginning, the origin of the solar system, etc. That, of course, leads us to want to know what kind of matter there was in the world. How did the stars evolve? What were the initial conditions? That is the problem of astronomical history. A great deal has been found out about the formation of stars, the formation of elements from which we were made, and even a little about the origin of the universe.

There is no historical question being studied in physics at the present time. We do not have a question, "Here are the laws of physics, how did they get that way?" We do not imagine, at the moment, that the laws of physics are somehow changing with time, that they were different in the past than they are at present. Of course they *may* be, and the moment we find they *are*, the historical question of physics will be wrapped up with the rest of the history of the universe, and then the physicist will be talking about the same problems as astronomers, geologists, and biologists.

Finally, there is a physical problem that is common to many fields, that is very old, and that has not been solved. It is not the problem of finding new fundamental particles, but something left over from a long time ago—over a hundred years. Nobody in physics has really been able to analyze it mathematically satisfactorily in spite of its importance to the sister sciences. It is the analysis of *circulating or turbulent fluids*. If we watch the evolution of a star, there comes a point where we can deduce that it is going to start convection, and thereafter we can no longer deduce what should happen. A few million years later the star explodes, but we cannot figure out the reason. We cannot analyze the weather. We do not know the patterns of motions that there should be inside the earth. The simplest form of the problem is to take a pipe that is very long and push water through it at high speed. We ask: to push a given amount of water through that pipe, how much pressure is needed? No one can analyze it from first principles and the properties of water. If the water flows very slowly, or if we use a thick goo like honey, then we can do it nicely. You will find that in your textbook.

What we really cannot do is deal with actual, wet water running through a pipe. That is the central problem which we ought to solve some day, and we have not.

A poet once said, "The whole universe is in a glass of wine." We will probably never know in what sense he meant that, for poets do not write to be understood. But it is true that if we look at a glass of wine closely enough we see the entire universe. There are the things of physics: the twisting liquid which evaporates depending on the wind and weather, the reflections in the glass, and our imagination adds the atoms. The glass is a distillation of the earth's rocks, and in its composition we see the secrets of the universe's age, and the evolution of stars. What strange array of chemicals are in the wine? How did they come to be? There are the ferments, the enzymes, the substrates, and the products. There in wine is found the great generalization: all life is fermentation. Nobody can discover the chemistry of wine without discovering, as did Louis Pasteur, the cause of much disease. How vivid is the claret, pressing its existence into the consciousness that watches it! If our small minds, for some convenience, divide this glass of wine, this universe, into parts—physics, biology, geology, astronomy, psychology, and so on—remember that nature does not know it! So let us put it all back together, not forgetting ultimately what it is for. Let it give us one more final pleasure: drink it and forget it all!

Conservation of Energy

4-1 What is energy?

In this chapter, we begin our more detailed study of the different aspects of physics, having finished our description of things in general. To illustrate the ideas and the kind of reasoning that might be used in theoretical physics, we shall now examine one of the most basic laws of physics, the conservation of energy.

There is a fact, or if you wish, a *law*, governing all natural phenomena that are known to date. There is no known exception to this law—it is exact so far as we know. The law is called the *conservation of energy*. It states that there is a certain quantity, which we call energy, that does not change in the manifold changes which nature undergoes. That is a most abstract idea, because it is a mathematical principle; it says that there is a numerical quantity which does not change when something happens. It is not a description of a mechanism, or anything concrete; it is just a strange fact that we can calculate some number and when we finish watching nature go through her tricks and calculate the number again, it is the same. (Something like the bishop on a red square, and after a number of moves—details unknown—it is still on some red square. It is a law of this nature.) Since it is an abstract idea, we shall illustrate the meaning of it by an analogy.

Imagine a child, perhaps "Dennis the Menace," who has blocks which are absolutely indestructible, and cannot be divided into pieces. Each is the same as the other. Let us suppose that he has 28 blocks. His mother puts him with his 28 blocks into a room at the beginning of the day. At the end of the day, being curious, she counts the blocks very carefully, and discovers a phenomenal law—no matter what he does with the blocks, there are always 28 remaining! This continues for a number of days, until one day there are only 27 blocks, but a little investigating shows that there is one under the rug—she must look everywhere to be sure that the number of blocks has not changed. One day, however, the number appears to change—there are only 26 blocks. Careful investigation indicates that the window was open, and upon looking outside, the other two blocks are found. Another day, careful count indicates that there are 30 blocks! This causes considerable consternation, until it is realized that Bruce came to visit, bringing his blocks with him, and he left a few at Dennis' house. After she has disposed of the extra blocks, she closes the window, does not let Bruce in, and then everything is going along all right, until one time she counts and finds only 25 blocks. However, there is a box in the room, a toy box, and the mother goes to open the toy box, but the boy says "No, do not open my toy box," and screams. Mother is not allowed to open the toy box. Being extremely curious, and somewhat ingenious, she invents a scheme! She knows that a block weighs three ounces, so she weighs the box at a time when she sees 28 blocks, and it weighs 16 ounces. The next time she wishes to check, she weighs the box again, subtracts sixteen ounces and divides by three. She discovers the following:

$$\left(\begin{array}{l} \text{number of} \\ \text{blocks seen} \end{array} \right) + \frac{\text{weight of box}}{\text{A ounce}} = \text{constant.} \quad (4.1)$$

There then appear to be some new deviations, but careful study indicates that the dirty water in the bathtub is changing its level. The child is throwing blocks into the water, and she cannot see them because it is so dirty, but she can find out how many blocks are in the water by adding another term to her formula. Since the original height of the water was 6 inches and each block raises the water a quarter

4-1 What is energy?

4-2 Gravitational potential energy

4-3 Kinetic energy

4-4 Other forms of energy

of an inch, this new formula would be:

$$\begin{aligned} \left(\frac{\text{number of}}{\text{blocks seen}} \right) & - \frac{\text{(weight of box)}}{3 \text{ ounces}} = \frac{16 \text{ ounces}}{3 \text{ ounces}} \\ + \frac{\text{(height of water)}}{1/2 \text{ inch}} & = \text{constant.} \quad (4.2) \end{aligned}$$

In the gradual increase in the complexity of her world, she finds a whole series of terms representing ways of calculating how many blocks are in places where she is not allowed to look. As a result, she finds a complex formula, a quantity which *has to be computed*, which always stays the same in her situation.

What is the analogy of this to the conservation of energy? The most remarkable aspect that must be abstracted from this picture is that *there are no blocks*. Take away the first terms in (4.1) and (4.2) and we find ourselves calculating more or less abstract things. The analogy has the following points. First, when we are calculating the energy, sometimes some of it leaves the system and goes away, or sometimes some comes in. In order to verify the conservation of energy, we must be careful that we have not put any in or taken any out. Second, the energy has a large number of *different forms*, and there is a formula for each one. These are: gravitational energy, kinetic energy, heat energy, elastic energy, electrical energy, chemical energy, radiant energy, nuclear energy, mass energy. If we total up the formulas for each of these contributions, it will not change except for energy going in and out.

It is important to realize that in physics today, we have no knowledge of what energy *is*. We do not have a picture that energy comes in little blobs of a definite amount. It is not that way. However, there are formulas for calculating some numerical quantity, and when we add it all together it gives ~~"the"~~—*always*, the same number. It is an abstract thing in that it does not tell us the mechanism or the *reasons* for the various formulas.

4-2 Gravitational potential energy

Conservation of energy can be understood only if we have the formula for all of its forms. I wish to discuss the formula for gravitational energy near the surface of the Earth, and I wish to derive this formula in a way which has nothing to do with history but is simply a line of reasoning invented for this particular lecture to give you an illustration of the remarkable fact that a great deal about nature can be extracted from a few facts and close reasoning. It is an illustration of the kind of work theoretical physicists become involved in. It is patterned after a most excellent argument by Mr. Carnot on the efficiency of steam engines.*

Consider weight-lifting machines—machines which have the property that they lift one weight by lowering another. Let us also make a hypothesis: that *there is no such thing as perpetual motion* with these weight-lifting machines. (In fact, that there is no perpetual motion at all is a general statement of the law of conservation of energy.) We must be careful to define perpetual motion. First, let us do it for weight-lifting machines. If, when we have lifted and lowered a lot of weights and restored the machine to the original condition, we find that the net result is to have *lifted a weight*, then we have a perpetual motion machine because we can use that lifted weight to run something else. That is, *provided* the machine which lifted the weight is brought back to its exact *original condition*, and furthermore that it is completely *self-contained*—that it has not received the energy to lift that weight from some external source—like Bruce's blocks.

A very simple weight-lifting machine is shown in Fig. 4-1. This machine lifts weights three units "strong." We place three units on one balance pan, and one unit on the other. However, in order to get it actually to work, we must lift a little weight off the left pan. On the other hand, we could lift a one-unit weight

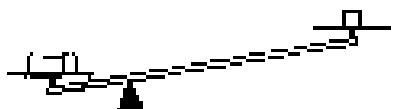


Fig. 4-1. Simple weight-lifting machine.

* Our point here is not so much the result, (4.3), which in fact you may already know, as the possibility of arriving at it by theoretical reasoning.

by lowering the three-unit weight, if we cheat a little by lifting a little weight off the other pan. Of course, we realize that with any *actual* lifting machine, we must add a little extra to get it to run. This we disregard, *temporarily*. Ideal machines, although they do not exist, do not require anything extra. A machine that we actually use can be, in a sense, *almost* reversible: that is, if it will lift the weight of three by lowering a weight of one, then it will also lift nearly the weight of one the same amount by lowering the weight of three.

We imagine that there are two classes of machines, those that are *not* reversible, which includes all real machines, and those that *are* reversible, which of course are actually not attainable no matter how careful we may be in our design of bearings, levers, etc. We suppose, however, that there is such a thing—a reversible machine—which lowers one unit of weight (a pound or any other unit) by one unit of distance, and at the same time lifts a three-unit weight. Call this reversible machine, Machine A. Suppose this particular reversible machine lifts the three-unit weight a distance X . Then suppose we have another machine, Machine B, which is not necessarily reversible, which also lowers a unit weight a unit distance, but which lifts three units a distance Y . We can now prove that Y is not higher than X ; that is, it is impossible to build a machine that will lift a weight *any higher* than it will be lifted by a reversible machine. Let us see why. Let us suppose that Y were higher than X . We take a one-unit weight and lower it one unit height with Machine B, and that lifts the three-unit weight up a distance V . Then we could lower the weight from Y to X , *obtaining free power*, and use the reversible Machine A, running backwards, to lower the three-unit weight a distance X and lift the one-unit weight by one unit height. This will put the one-unit weight back where it was before, and leave both machines ready to be used again! We would therefore have perpetual motion if Y were higher than X , which we assumed was impossible. With those assumptions, we thus deduce that **F is not higher than X**, so that of all machines that can be designed, the reversible machine is the best.

We can also see that all reversible machines must lift to *exactly the same height*. Suppose that B were really reversible also. The argument that Y is not higher than X is, of course, just as good as it was before, but we can also make our argument the other way around, using the machines in the opposite order, and prove that X is not higher than Y . This, then, is a very remarkable observation because it permits us to analyze the height to which different machines are going to lift something *without looking at the interior mechanism*. We know at once that if somebody makes an enormously elaborate series of levers that lift three units a certain distance by lowering one unit by one unit distance, and we compare it with a simple lever which does the same thing and is fundamentally reversible, his machine will lift it no higher, but perhaps less high. If his machine is reversible, we also know exactly *how* high it will lift. To summarize: every reversible machine, no matter how it operates, which drops one pound one foot and lifts a three-pound weight always lifts it the same distance, X . This is clearly a universal law of great utility. The next question is, of course, what is **T**?

Suppose we have a reversible machine which is going to lift this distance X , three for one. We set up three balls in a rack which does not move, as shown in Fig. 4-2. One ball is held on a stage at a distance one foot above the ground. The machine can lift three balls, lowering one by a distance 1. Now, we have arranged that the platform which holds three balls has a floor and two shelves, exactly spaced at distance X , and further, that the rack which holds the balls is spaced at distance X , (a). First we roll the balls horizontally from the rack to the shelves, (b), and we suppose that this takes no energy because we do not change the height. The reversible machine then operates: it lowers the single ball to the floor, and it lifts the rack a distance X , (c). Now we have ingeniously arranged the rack so that these balls are again even with the platforms. Thus we unload the balls onto the rack, (d); having unloaded the balls, we can restore the machine to its original condition. Now we have three balls on the upper three shelves and one at the bottom. But the strange thing is that, in a certain way of speaking, we have not lifted *two* of them at all because, after all, there were balls on shelves 2 ~~and 1~~.

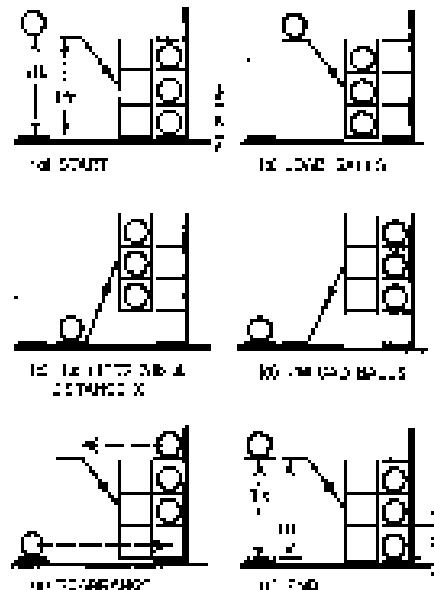


Fig. 4-2. A reversible machine.

before. The resulting effect has been to lift *one ball* a distance $3X$. Now, if $3X$ exceeds one foot, then we can *lower* the ball to return the machine to the initial condition, (f), and we can run the apparatus again. Therefore $3X$ cannot exceed one foot, for if $3X$ exceeds one foot we can make perpetual motion. Likewise, we can prove that *one foot cannot exceed $3X$* , by making the whole machine run the opposite way, since it is a reversible machine. Therefore $3X$ is neither greater nor less than a foot, and we discover then, by argument alone, the law that $X = \frac{1}{3}$ foot. The generalization is clear: one pound falls a certain distance in operating a reversible machine; then the machine can lift p pounds this distance divided by p . Another way of putting the result is that three pounds times the height lifted, which in our problem was X , is equal to one pound times the distance lowered, which is one foot in this case. If we take all the weights and multiply them by the heights at which they are now, above the floor, let the machine operate, and then multiply all the weights by all the heights again, *there will be no change*. (We have to generalize the example where we moved only one weight to the case where when we lower one we lift several different ones—but that is easy.)

We call the sum of the weights times the heights *gravitational potential energy*—the energy which an object has because of its relationship in space, relative to the earth. The formula for gravitational energy, then, so long as we are not too far from the earth (the force weakens as we go higher) is

$$\left(\begin{array}{l} \text{(gravitational} \\ \text{potential energy)} \\ \text{for one object!} \end{array} \right) = (\text{weight!}) \times (\text{height}). \quad (4.3)$$

It is a very beautiful line of reasoning. The only problem is that perhaps it is not true. (After all, nature does not *have* to go along with our reasoning.) For example, perhaps perpetual motion is, in fact, possible. Some of the assumptions may be wrong, or we may have made a mistake in reasoning, so it is always necessary to check. It turns out experimentally, in fact, to be true.

The general name of energy which has to do with location relative to something else is called *potential energy*. In this particular case, of course, we call it *gravitational potential energy*. If it is a question of electrical forces against which we are working, instead of gravitational forces, if we are "lifting" charges away from other charges with a lot of levers, then the energy content is called *electrical potential energy*. The general principle is that the change in the energy is the force times the distance that the force is pushed, and that this is a change in energy in general:

$$\left(\begin{array}{l} \text{(change in)} \\ \text{energy} \end{array} \right) = (\text{force}) \times \left(\begin{array}{l} \text{(distance force)} \\ \text{acts through} \end{array} \right). \quad (4.4)$$

We will return to many of these other kinds of energy as we continue the course.

The principle of the conservation of energy is very useful for deducing what will happen in a number of circumstances. In high school we learned a lot of laws about pulleys and levers used in different ways. We can now see that these "laws" are *all the same thing*, and that we did not have to memorize 75 rules to figure it out. A simple example is a smooth inclined plane which is, happily, a three-four-five triangle (Fig. 4-3). We hang a one-pound weight on the inclined plane with a pulley, and on the other side of the pulley, a weight W . We want to know how heavy W must be to balance the one pound on the plane. How can we figure that out? If we say it is just balanced, it is reversible and so can move up and down, and we can consider the following situation. In the initial circumstance, (a), the one pound weight is at the bottom and weight W is at the top. When W has slipped down in a reversible way, we have a one-pound weight at the top and the weight W the slant distance, (b), or five feet, from the plane in which it was before. We *lifted* the one-pound weight only *three* feet and we lowered W pounds by *five* feet. Therefore $W = \frac{5}{3}$ of a pound. Note that we deduced this from the *conservation of energy*, and not from force components. Cleverness, however, is relative. It can be deduced in a way which is even more brilliant, discovered by

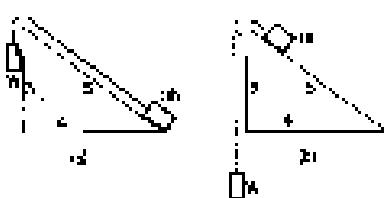


Fig. 4-3. Reversible slope.

Stevinus and inscribed on his tombstone. Figure 4-4 explains that it has to be $\frac{1}{5}$ of a pound, because the chain does not go around. It is evident that the lower part of the chain is balanced by itself, so that the pull of the five weights on one side must balance the pull of three weights on the other, or whatever the ratio of the legs. You see, by looking at this diagram, that W must be $\frac{1}{5}$ of a pound. (If you get an epitaph like that on your gravestone, you are doing fine.)

Let us now illustrate the energy principle with a more complicated problem, the screw jack shown in Fig. 4-5. A handle 20 inches long is used to turn the screw, which has 10 threads to the inch. We would like to know how much force would be needed at the handle to lift one ton (2000 pounds). If we want to lift the ton one inch, say, then we must turn the handle around ten times. When it goes around once it goes approximately 126 inches. The handle must thus travel 1260 inches, and if we used various pulleys, etc., we would be lifting our one ton with an unknown smaller weight W applied to the end of the handle. So we find out that W is about 1.6 pounds. This is a result of the conservation of energy.

Take now the somewhat more complicated example shown in Fig. 4-6. A rod or bar, 8 feet long, is supported at one end. In the middle of the bar is a weight of 60 pounds, and at a distance of two feet from the support there is a weight of 100 pounds. How hard do we have to lift the end of the bar in order to keep it balanced, disregarding the weight of the bar? Suppose we put a pulley at one end and hang a weight on the pulley. How big would the weight W have to be in order for it to balance? We imagine that the weight falls any arbitrary distance—to make it easy for ourselves suppose it goes down 4 inches—how high would the two load weights rise? The center rises 2 inches, and the point a quarter of the way from the fixed end lifts 1 inch. Therefore, the principle that the sum of the heights times the weights does not change tells us that the weight W times 4 inches down, plus 60 pounds times 2 inches up, plus 100 pounds times 1 inch has to add up to nothing:

$$-4W + (2)(60) = 11(100) = 0, \quad W = 55 \text{ lb.} \quad (4.5)$$

Thus we must have a 55-pound weight to balance the bar. In this way we can work out the laws of "balance"—the statics of complicated bridge arrangements, and so on. This approach is called *the principle of virtual work*, because in order to apply this argument we had to *imagine* that the structure moves a little—even though it is not *really* moving or even *movable*. We use the very small imagined motion to apply the principle of conservation of energy.

4-3 Kinetic energy

To illustrate another type of energy we consider a pendulum (Fig. 4-7). If we pull the mass aside and release it, it swings back and forth. In its motion, it loses height in going from either end to the center. Where does the potential energy go? ~~Gravitational~~ energy disappears when it is down at the bottom; nevertheless, it will climb up again. The gravitational energy must have gone into another form. Evidently it is by virtue of its *motion* that it is able to climb up again, so we have the conversion of gravitational energy into some other form when it reaches the bottom.

We must get a formula for the energy of motion. Now, recalling our arguments about reversible machines, we can easily see that in the motion at the bottom must be a quantity of energy which permits it to rise a certain height, and which has nothing to do with the *machinery* by which it comes up or the *path* by which it comes up. So we have an equivalence formula something like the one we wrote for the child's blocks. We have another form to represent the energy. It is easy to say what it is. The kinetic energy at the bottom equals the weight times the height that it could go, corresponding to its velocity: $K.E. = WH$. What we need is the formula which tells us the height by some rule that has to do with the motion of objects. If we start something out with a certain velocity, say straight up, it will reach a certain height; we do not know what it is yet, but it depends on the velocity—there is a formula for that. Then to find the formula for kinetic energy

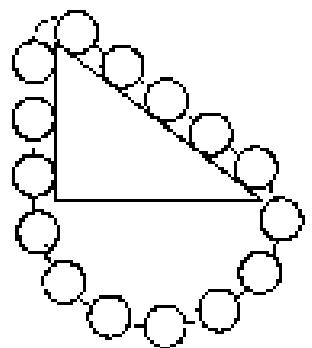


Fig. 4-4. The principle of Stevinus.

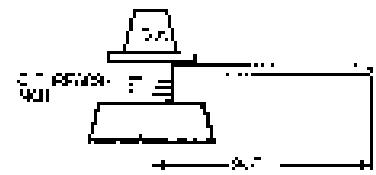


Fig. 4-5. A screw jack.

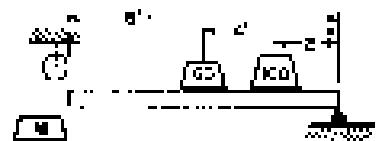


Fig. 4-6. Weighted rod suspended on one end.

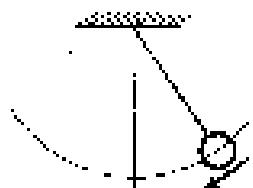


Fig. 4-7. Pendulum.

for an object moving with velocity V , we must calculate the height that it could reach, and multiply by the weight. We shall soon find that we can write it this way:

$$K.E. = \frac{1}{2} m V^2 \quad (4.6)$$

Of course, the fact that motion has energy has nothing to do with the fact that we are in a gravitational field. It makes no difference *where* the motion came from. This is a general formula for various velocities. Both (4.3) and (4.6) are approximate formulas, the first because it is incorrect when the heights are great, i.e., when the heights are so high that gravity is weakening; the second, because of the relativistic correction at high speeds. However, when we do finally get the exact formula for the energy, then the law of conservation of energy is correct.

44 Other forms of energy

We can continue in this way to illustrate the existence of energy in other forms. First, consider elastic energy. If we pull down on a spring, we must do some work, for when we have it down, we can lift weights with it. Therefore in its stretched condition it has a possibility of doing some work. If we were to evaluate the sums of weights times heights, it would not check out—we must add something else to account for the fact that the spring is under tension. Elastic energy is the formula for a spring when it is stretched. How much energy is it? If we let go, the elastic energy, as the spring passes through the equilibrium point, is converted to kinetic energy and it goes back and forth between compressing or stretching the spring and kinetic energy of motion. (There is also some gravitational energy going in and out, but we can do this experiment "sideways" if we like.) It keeps going until the losses—Aha! We have cheated all the way through by putting on little weights to move things or saying that the machines are reversible, or that they go on forever, but we can see that things do stop, eventually. Where is the energy when the spring has finished moving up and down? This brings in *another* form of energy: *heat energy*.

Inside a spring or a lever there are crystals which are made up of lots of atoms, and with great care and delicacy in the arrangement of the parts one can try to adjust things so that as something rolls on something else, none of the atoms do any jiggling at all. But one must be very careful. Ordinarily when things roll, there is bumping and jiggling because of the irregularities of the material, and the atoms start to wiggle inside. So we lose track of that energy; we find the atoms are wiggling inside in a random and confused manner after the motion slows down. There is still kinetic energy, all right, but it is not associated with visible motion. What a dream! How do we *know* there is still kinetic energy? It turns out that with thermometers you can find out that, in fact, the spring or the lever is *warmer*, and that there is really an increase of kinetic energy by a definite amount. We call this form of energy *heat energy*, but we know that it is not really a new form, it is just kinetic energy—internal motion. (One of the difficulties with all these experiments with matter that we do on a large scale is that we cannot really demonstrate the conservation of energy and we cannot really make our reversible machines, because every time we move a large clump of stuff, the atoms do not remain absolutely undisturbed, and so a certain amount of random motion goes into the atomic system. We cannot see it, but we can measure it with thermometers, etc.)

There are many other forms of energy, and of course we cannot describe them in any more detail just now. There is electrical energy, which has to do with pushing and pulling by electric charges. There is radiant energy, the energy of light, which we know is a form of electrical energy because light can be represented as wigglings in the electromagnetic field. There is chemical energy, the energy which is released in chemical reactions. Actually, elastic energy is, to a certain extent, like chemical energy, because chemical energy is the energy of the attraction of the atoms, one for the other, and so is elastic energy. Our modern understanding is the following: chemical energy has two parts, kinetic energy of the electrons inside the atoms, so part of it is kinetic, and electrical energy of interaction of the

electrons and the protons—the rest of it, therefore, is electrical. Next we come to nuclear energy, the energy which is involved with the arrangement of particles inside the nucleus, and we have formulas for that, but we do not have the fundamental laws. We know that it is not electrical, not gravitational, and not purely chemical, but we do not know what it is. It seems to be an additional form of energy. Finally, associated with the relativity theory, there is a modification of the laws of kinetic energy, or whatever you wish to call it, so that kinetic energy is combined with another thing called *mass energy*. An object has energy from its sheer *existence*. If I have a positron and an electron, standing still doing nothing—never mind gravity, never mind anything—and they come together and disappear, radiant energy will be liberated, in a definite amount, and the amount can be calculated. All we need know is the mass of the object. It does not depend on what it is—we make two things disappear, and we get a certain amount of energy. The formula was first found by Einstein; it is $E = mc^2$.

It is obvious from our discussion that the law of conservation of energy is enormously useful in making analyses, as we have illustrated in a few examples without knowing all the formulas. If we had all the formulas for all kinds of energy, we could analyze how many processes should work without having to go into the details. Therefore conservation laws are very interesting. The question naturally arises as to what other conservation laws there are in physics. There are two other conservation laws which are analogous to the conservation of energy. One is called the conservation of linear momentum. The other is called the conservation of angular momentum. We will find out more about these later. In the last analysis, we do not understand the conservation laws deeply. We do not understand the conservation of energy. We do not understand energy as a certain number of little blobs. You may have heard that photons come out in blobs and that the energy of a photon is Planck's constant times the frequency. That is true, but since the frequency of light can be anything, there is no law that says that energy has to be a certain definite amount. Unlike Dennis' blocks, there can be any amount of energy, at least as presently understood. So we do not understand this energy as counting something at the moment, but just as a mathematical quantity, which is an abstract and rather peculiar circumstance. In quantum mechanics it turns out that the conservation of energy is very closely related to another important property of the world, *things do not depend on the absolute time*. We can set up an experiment at a given moment and try it out, and then do the same experiment at a later moment, and it will behave in exactly the same way. Whether this is strictly true or not, we do not know. If we assume that it is true, and add the principles of ~~classical~~ mechanics, then we can deduce the principle of the conservation of energy. It is a rather subtle and interesting thing, and it is not easy to explain. The other conservation laws are also linked together. The conservation of momentum is associated in quantum mechanics with the proposition that it makes no difference ~~when~~ you do the experiment, the results will always be the same. As independence in space has to do with the conservation of momentum, independence of time has to do with the conservation of energy, and finally, if we turn our apparatus, this too makes no difference, and so the invariance of the world to angular orientation is related to the conservation of *angular momentum*. Besides these, there are three other conservation laws, that are exact so far as we can tell today, which are much simpler to understand because they are in the nature of counting blocks.

The first of the three is the ~~conservation~~ of charge, and that merely means that you count how many positive, minus how many negative electrical charges you have, and the number is never changed. You may get rid of a positive with a negative, but you do not create any net excess of positives over negatives. Two other laws are analogous to this one—one is called the *conservation of baryons*. There are a number of strange particles, a neutron and a proton are examples, which are called baryons. In any reaction whatever in nature, if we count how many baryons are coming into a process, the number of baryons* which come out

*Counting antibaryons as —1 baryon.

will be exactly the same. There is another law, the *conservation of leptons*. We can say that the group of particles called leptons are: electron, mu meson, and neutrino. There is an antielectron which is a positron, that is, a -1 lepton. Counting the total number of leptons in a reaction reveals that the number in and out never changes, at least so far as we know at present.

These are the six conservation laws, three of them subtle, involving space and time, and three of them simple, in the sense of counting something.

With regard to the conservation of energy, we should note that *available* energy is another matter—there is a lot of jiggling around in the atoms of the water of the sea, because the sea has a certain temperature, but it is impossible to get them herded into a definite motion without taking energy from somewhere else. That is, although we know for a fact that energy is conserved, the energy available for human utility is not conserved so easily. The laws which govern how much energy is available are called the *laws of thermodynamics* and involve a concept called entropy for irreversible thermodynamic processes.

Finally, we remark on the question of where we can get our supplies of energy today. Our supplies of energy are from the sun, rain, coal, uranium, and hydrogen. The sun makes the rain, and the coal also, so that all these are from the sun. Although energy is conserved, nature does not seem to be interested in it; she liberates a lot of energy from the sun, but only one part in two billion falls on the earth. Nature has conservation of energy, but does not really care; she spends a lot of it in all directions. We have already obtained energy from uranium; we can also get energy from hydrogen, but at present only in an explosive and dangerous condition. If it can be controlled in thermonuclear reactions, it turns out that the energy that can be obtained from 10 quarts of water per second is equal to all of the electrical power generated in the United States. With 150 gallons of running water a minute, you have enough fuel to supply all the energy which is used in the United States today! Therefore it is up to the physicist to figure out how to liberate us from the need for having energy. It can be done.

Time and Distance

5-1 Motion

In this chapter we shall consider some aspects of the concepts of *time* and *distance*. It has been emphasized earlier that physics, as do all the sciences, depends on *observation*. One might also say that the development of the physical sciences to their present form has depended to a large extent on the emphasis which has been placed on the making of *quantitative* observations. Only with quantitative observations can one arrive at quantitative relationships, which are the heart of physics.

Many people would like to place the beginnings of physics with the work done 350 years ago by Galileo, and to call him the first physicist. Until that time, the study of motion had been a philosophical one based on arguments that could be thought up in one's head. Most of the arguments had been presented by Aristotle and other Greek philosophers, and were taken as "proven." Galileo was skeptical, and did an experiment on motion which was essentially this: He allowed a ball to roll down an inclined trough and observed the motion. He did not, however, just look; he measured *how far* the ball went in *how long a time*.

The way to measure a distance was well known long before Galileo, but there were no accurate ways of measuring time, particularly short times. Although he later devised more satisfactory clocks (though not like the ones we know), Galileo's first experiments on motion were done by using his pulse to count off equal intervals of time. Let us do the same.

We may count off beats of a pulse as the ball rolls down the track: "one . . . two . . . three . . . four . . . five . . . six . . . seven . . . eight . . ." We ask a friend to make a small mark at the location of the ball at each count; we can then measure the *distance* the ball travelled from the point of release in one, or two, or three, etc., equal intervals of time. Galileo expressed the result of *his* observations in this way: if the location of the ball is marked at 1, 2, 3, 4, . . . units of time from the instant of its release, those marks are distant from the starting point in proportion to the numbers 1, 4, 9, 16, . . . Today we would say the distance is proportional to the square of the time:

$$D \propto t^2$$

The study of *motion*, which is basic to all of physics, treats with the questions: where? and when?

5-2 Time

Let us consider first what we mean by *time*. What *is* time? It would be nice if we could find a good definition of time. Webster defines "a time" as "a period," and the latter as "a time," which doesn't seem to be very useful. Perhaps we should say: "Time is what happens when nothing else happens." Which also doesn't get us very far. Maybe it is just as well if we face the fact that time is one of the things we probably cannot define (in the dictionary sense), and just say that it is what we already know it to be: it is how long we wait!

What really matters anyway is not how we *define* time, but how we measure it. One way of measuring time is to utilize something which happens over and over again in a regular fashion—something which is *periodic*. For example, a day. A day seems to happen over and over again. But when you begin to think

5-1 Motion

5-2 Time

5-3 Short times

5-4 Long times

5-5 Units and standards of time

5-6 Large distances

5-7 Short distances

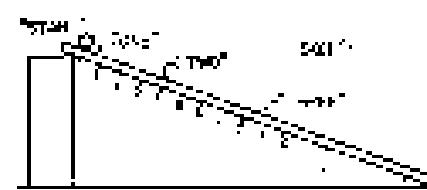


Fig. 5-1. A ball rolls down an inclined track.

about it, you might well ask: "Are days periodic; are they regular? Are all days the same length?" One certainly has the impression that days in summer are longer than days in winter. Of course, some of the days in winter seem to get awfully long if one is very bored. You have certainly heard someone say, "My, but this has been a long day!"

It does seem, however, that days are about the same length *on the average*. Is there any way we can test whether the days are the same length—either from one day to the next, or at least on the average? One way is to make a comparison with some other periodic phenomenon. Let us see how such a comparison might be made with an hour glass. With an hour glass, we can "create" a periodic occurrence if we have someone standing by it day and night to turn it over whenever the last grain of sand runs out.

We could then count the turnings of the glass from each morning to the next. We would find, this time, that the number of "hours" (i.e., turnings of the glass) was not the same each "day." We should distrust the sun, or the glass, or both. After some thought, it might occur to us to count the "hours" from noon to noon. (Noon is here defined *not* as 12:00 o'clock, but that instant when the sun is at its highest point.) We would find, this time, that the number of "hours" each day is the same.

We now have some confidence that both the "hour" and the "day" have a regular periodicity, i.e., mark off successive equal intervals of time, although we have not *proved* that either one is "really" periodic. Someone might question whether there might not be some omnipotent being who would slow down the flow of sand every night and speed it up during the day. Our experiment does not, of course, give us an answer to this sort of question. All we can say is that we find that a regularity of one kind fits together with a regularity of another kind. We can just say that we base our *definition* of time on the repetition of some apparently periodic event.

5-3 Short times

We should now notice that in the process of checking on the reproducibility of the day, we have received an important by-product. We have found a way of measuring, more accurately, *fractions* of a day. We have found a way of counting time in smaller pieces. Can we carry the process further, and learn to measure even smaller intervals of time?

Galileo decided that a given pendulum always swings back and forth in equal intervals of time so long as the size of the swing is kept small. A test comparing the number of swings of a pendulum in one "hour" shows that such is indeed the case. We can in this way mark fractions of an hour. If we use a mechanical device to count the swings—and to keep them going—we have the pendulum clock of our grandfathers.

Let us agree that if our pendulum oscillates 3600 times in one hour (and if there are 24 such hours in a day), we shall call each period of the pendulum one "second." We have then divided our original unit of time into approximately 10s parts. We can apply the same principles to divide the second into smaller and smaller intervals. It is, you will realize, not practical to make mechanical pendulums which go arbitrarily fast, but we can now make *electrical* pendulums, called oscillators, which can provide a periodic ~~succession~~ with a very short period of swing. In these electronic oscillators it is an electrical current which swings to and fro, in a manner analogous to the swinging of the bob of the pendulum.

We can make a series of such electronic oscillators, each with a period 10 times shorter than the previous one. We may "calibrate" each oscillator against the next slower one by counting the number of swings it makes for one swing of the slower oscillator. When the period of oscillation of our clock is shorter than a fraction of a second, we cannot count the oscillations without the help of some device which extends our powers of observation. One such device is the electron-beam oscilloscope, which acts as a sort of microscope for short times. This device plots on a fluorescent screen a graph of electrical current (or voltage) versus time.

By connecting the oscilloscope to two of our oscillators in sequence, so that it plots a graph first of the current in one of our oscillators and then of the current in the other, we get two graphs like those shown in Fig. 5-2. We can readily determine the number of periods of the faster oscillator in one period of the slower oscillator.

With modern electronic techniques, oscillators have been built with periods as short as about 10^{-11} second, and they have been calibrated (by comparison methods such as we have described) in terms of our standard unit of time, the second. With the invention and perfection of the "laser," or light amplifier, in the past few years, it has become possible to make oscillators with even shorter periods than 10^{-11} second, but it has not yet been possible to calibrate them by the methods which have been described, although it will no doubt soon be possible.

Times shorter than 10^{-11} second have been measured, but by a different technique. In effect, a different *definition* of "time" has been used. One way has been to observe the *distance* between two happenings on a moving object. If, for example, the headlights of a moving automobile are turned on and then off, we can figure out *how long* the lights were on if we know *where* they were turned on and off and how fast the car was moving. The time is the distance over which the lights were on divided by the speed.

Within the past few years, just such a technique was used to measure the lifetime of the π^0 -meson. By observing in a microscope the minute tracks left in a photographic emulsion in which π^0 -mesons had been created one saw that a π^0 -meson (known to be travelling at a certain speed nearly that of light) went a distance of about 10^{-7} meter, on the average, before disintegrating. It lived for only about 10^{-16} sec. It should be emphasized that we have here used a somewhat different definition of "time" than before. So long as there are no inconsistencies in our understanding, however, we feel fairly confident that our definitions are sufficiently equivalent.

By extending our techniques—and if necessary our definitions—still further we can infer the time duration of still faster physical events. We can speak of the period of a nuclear vibration. We can speak of the lifetime of the newly discovered strange resonances (particles) mentioned in Chapter 2. Their complete life occupies a time span of only 10^{-24} second, approximately the time it would take light (which moves at the fastest known speed) to cross the nucleus of hydrogen (the smallest known object).

What about still smaller times? Does "time" exist on a still smaller scale? Does it make any sense to speak of smaller times if we cannot measure—or perhaps even think sensibly about—something which happens in a shorter time? Perhaps not. These are some of the open questions which you will be asking and perhaps answering in the next twenty or thirty years.

5-4 Long times

Let us now consider times longer than one day. Measurement of longer times is easy; we just count the days—so long as there is someone around to do the counting. First we find that there is another natural periodicity: the year, about 365 days. We have also discovered that nature has sometimes provided a counter for the years, in the form of tree rings or river-bottom sediments. In some cases we can use these natural time markers to determine the time which has passed since some early event.

When we cannot count the years for the measurement of long times, we must look for other ways to measure. One of the most successful is the use of radioactive material as a "clock." In this case we do not have a periodic occurrence, as for the day or the pendulum, but a new kind of "regularity." We find that the radioactivity of a particular sample of material decreases by the same *fraction* for successive equal increases in its age. If we plot a graph of the radioactivity observed as a function of time (say in days), we obtain a curve like that shown in Fig. 5-3. We observe that if the radioactivity decreases to one-half in T days (called the "half-life"), then it decreases to one-quarter in another T days, and so

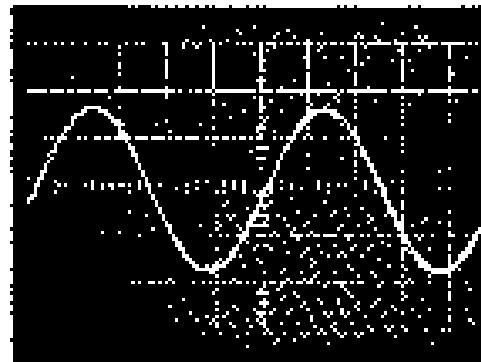


Fig. 5-2. Two views of an oscilloscope screen. In (a) the oscilloscope is connected to one oscillator, in (b) it is connected to an oscillator with a period one-tenth as long.

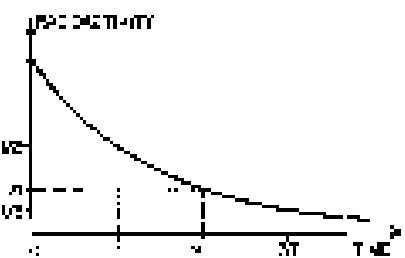


Fig. 5-3. The decrease with time of radioactivity. The activity decreases by one-half in each "half-life," T .

TIME	MEAN LIFE OF	SECONDS	
		SECONDS	SECONDS
10 ⁻¹	Age of universe	3.16 x 10 ¹⁷	
10 ⁻²	Age of earth	4.5 x 10 ⁹	4.5 x 10 ⁹
10 ⁻³	Earthman	10 ⁶	
10 ⁻⁴	Age of pyramid	10 ¹²	
10 ⁻⁵	Age of U.S.	10 ¹⁰	10 ¹⁰
10 ⁻⁶	Life of a man	10 ⁷	10 ⁷
10 ⁻⁷	One day	10 ⁴	
10 ⁻⁸	Light goes from sun to earth	10 ³	10 ³
10 ⁻⁹	One hour sec	10 ²	
10 ⁻¹⁰	1/10 of one minute	10 ¹	
10 ⁻¹¹	Period of red waves	10 ⁻²	10 ⁻²
10 ⁻¹²	Light travels one atom	10 ⁻³	10 ⁻³
10 ⁻¹³	Period of nuclear vibration	10 ⁻⁴	10 ⁻⁴
10 ⁻¹⁴	Light crosses nucleus	10 ⁻⁵	Strong particle
		1 2 3 4 5 6 7	

on. In an arbitrary time interval t there are t/T "half-lives," and the fraction left after this time t is $(\frac{1}{2})^{t/T}$.

If we knew that a piece of material, say a piece of wood, had contained an amount A of radioactive material when it was formed, and we found out by a direct measurement that it now contains the amount B , we could compute the age of the object, t , by solving the equation

$$(1/2)^{t/T} = B/A.$$

There are, fortunately, cases in which we can know the amount of radioactivity that was in an object when it was formed. We know, for example, that the carbon dioxide in the air contains a certain small fraction of the radioactive carbon isotope C¹⁴ (replenished continuously by the action of cosmic rays). If we measure the total carbon content of an object, we know that a certain fraction of that amount was originally the radioactive C¹⁴; we know, therefore, the starting amount A to use in the formula above. Carbon-14 has a half-life of 5000 years. By careful measurements we can measure the amount left after 20 half-lives or so and can therefore "date" organic objects which grew as long as 100,000 years ago.

We would like to know, and we think we do know, the life of still older things. Much of our knowledge is based on the measurements of other radioactive isotopes which have different half-lives. If we make measurements with an isotope with a longer half-life, then we are able to measure longer times. Uranium, for example, has an isotope whose half-life is about 10⁹ years, so that if some material was formed with uranium in it 10⁹ years ago, only half the uranium would remain today. When the uranium disintegrates, it changes into lead. Consider a piece of rock which was formed a long time ago in some chemical process. Lead, being of a chemical nature different from uranium, would appear in one part of the rock and uranium would appear in another part of the rock. The uranium and lead

would be separate. If we look at that piece of rock today, where there should only be uranium we will now find a certain fraction of uranium and a certain fraction of lead. By comparing these fractions, we can tell what percent of the uranium disappeared and changed into lead. By this method, the age of certain rocks has been determined to be several billion years. An extension of this method, not using particular rocks but looking at the uranium and lead in the oceans and using averages over the earth, has been used to determine (within the past few years) that the age of the earth itself is approximately 5.5 billion years.

It is encouraging that the age of the earth is found to be the same as the age of the meteorites which land on the earth, as determined by the uranium method. It appears that the earth was formed out of rocks floating in space, and that the meteorites are, quite likely, some of that material left over. At some time more than five billion years ago, the universe started. It is now believed that at least our part of the universe had its beginning about ten or twelve billion years ago. We do not know what happened before then. In fact, we may well ask again: Does the question make any sense? Does an earlier time have any meaning?

5-5 Units and standards of time

We have implied that it is convenient if we start with some standard unit of time, say a day or a second, and refer all other times to some multiple or fraction of this unit. What shall we take as our basic standard of time? Shall we take the human pulse? If we compare pulses, we find that they seem to vary a lot. On comparing two clocks, one finds they do not vary so much. You might then say, well, let us take a clock. But whose clock? There is a story of a Swiss boy who wanted all of the clocks in his town to ring noon at the same time. So he went around trying to convince everyone of the value of this. Everyone thought it was a marvelous idea so long as all of the other clocks rang noon when his did! It is rather difficult to decide whose clock we should take as a standard. Fortunately, we all share one clock—the earth. For a long time the rotational period of the earth has been taken as the basic standard of time. As measurements have been made more and more precise, however, it has been found that the rotation of the earth is not exactly periodic, when measured in terms of the best clocks. These "best" clocks are those which we have reason to believe are accurate because they agree with each other. We now believe that, for various reasons, some days are longer than others, some days are shorter, and on the average the period of the earth becomes a little longer as the centuries pass.

Until very recently we had found nothing much better than the earth's period, so all clocks have been related to the length of the day, and the second has been defined as $1/86400$ of an average day. Recently we have been gaining experience with some natural oscillators which we now believe would provide a more constant time reference than the earth, and which are also based on a natural phenomenon available to everyone. These are the so-called "atomic clocks." Their basic internal period is that of an atomic vibration which is very insensitive to the temperature or any other external effects. These clocks keep time to an accuracy of one part in 10^9 or better. Within the past two years an improved atomic clock which operates on the vibration of the hydrogen atom has been designed and built by Professor Norman Ramsey at Harvard University. He believes that this clock might be 100 times more accurate still. Measurements now in progress will show whether this is true or not.

We may expect that since it has been possible to build clocks much more accurate than astronomical time, there will soon be an agreement among scientists to define the unit of time in terms of one of the atomic clock standards.

5-6 Large distances

Let us now turn to the question of *distance*. How far, or how big, are things? Everybody knows that the way you measure distance is to start with a stick and count. Or start with a thumb and count. You begin with a unit and count. How

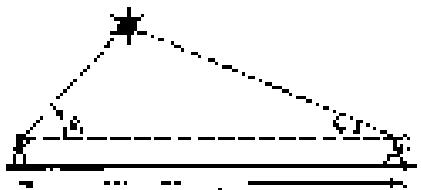


Fig. 5-4. The height of a Sputnik is determined by triangulation

does one measure smaller things? How does one subdivide distance? In the same way that we subdivided time: we take a smaller unit and count the number of such units it takes to make up the longer unit. So we can measure smaller and smaller lengths.

But we do not always mean by distance what one gets by counting off with a meter stick. It would be difficult to measure the horizontal distance between two mountain tops using only a meter stick. We have found by experience that distance can be measured in another fashion: by triangulation. Although this means that we are really using a different definition of distance, when they can both be used they agree with each other. Space is more or less what Euclid thought it was, so the two types of definitions of distance agree. Since they do agree on the earth it gives us some confidence in using triangulation for still larger distances. For example, we were able to use triangulation to measure the height of the first Sputnik. We found that it was roughly 5×10^5 meters high. By more careful measurements the distance to the moon can be measured in the same way. Two telescopes at different places on the earth can give us the two angles we need. It has been found in this way that the moon is 4×10^8 meters away.

We cannot do the same with the sun, or at least no one has been able to yet. The accuracy with which one can focus on a given point on the sun and with which one can measure angles is not good enough to permit us to measure the distance to the sun. Then how can we measure the distance to the sun? We must invent an extension of the idea of triangulation. We measure the relative distances of all the planets by astronomical observations of where the planets appear to be, and we get a picture of the solar system with the proper relative distances of everything, but with no *absolute* distance. One absolute measurement is then required, which has been obtained in a number of ways. One of the ways, which was believed until recently to be the most accurate, was to measure the distance from the earth to Eros, one of the small planetoids which passes near the earth every now and then. By triangulation on this little object, one could get the one required scale measurement. Knowing the relative distances of the rest, we can then tell the distance, for example, from the earth to the sun, or from the earth to Pluto.

Within the past year there has been a big improvement in our knowledge of the scale of the solar system. At the Jet Propulsion Laboratory the distance from the earth to Venus was measured quite accurately by a direct radar observation. This, of course, is a still different type of inferred distance. We say we know the speed at which light travels (and therefore, at which radar waves travel), and we assume that it is the same speed everywhere between the earth and Venus. We send the radio wave out, and count the time until the reflected wave comes back. From the *time* we infer a *distance*, assuming we know the speed. We have really another definition of a measurement of distance.

How do we measure the distance to a star, which is much farther away? Fortunately, we can go back to our triangulation method, because the earth moving around the sun gives us a large baseline for measurements of objects outside the solar system. If we focus a telescope on a star in summer and in winter, we might hope to determine these two angles accurately enough to be able to measure the distance to a star.

What if the stars are too far away for us to use triangulation? Astronomers are always inventing new ways of measuring distance. They find, for example, that they can estimate the size and brightness of a star by its color. The color and brightness of many nearby stars—whose distances are known by triangulation—have been measured, and it is found that there is a smooth relationship between the color and the intrinsic brightness of stars (in most cases). If one now measures the color of a distant star, one may use the color-brightness relationship to determine the intrinsic brightness of the star. By measuring how bright the star *appears* to us at the earth (or perhaps we should say how *dim* it appears), we can compute how far away it is. (For a given intrinsic brightness, the apparent brightness decreases with the square of the distance.) A nice confirmation of the correctness of this method of measuring stellar distances is given by the results obtained for groups of stars known as globular clusters. A photograph of such a group is

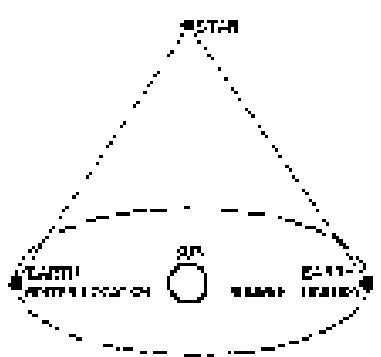


Fig. 5-5. The distance of nearby stars can be measured by triangulation, using the diameter of the earth's orbit as a baseline.

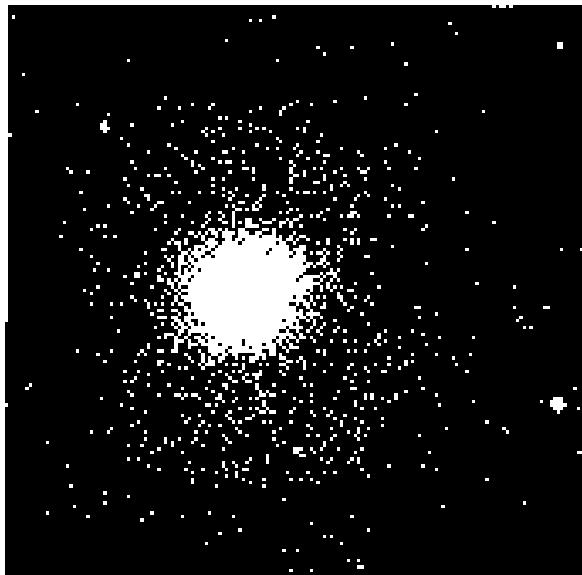


Fig. 5-6. A cluster of stars near the center of our galaxy. Their distance from the earth is 30,000 light-years, or about 3×10^{20} meters.

shown in Fig. 5-6. Just from looking at the photograph one is convinced that these stars are all together. The same result is obtained from distance measurements by the color-brightness method.

A study of many globular clusters gives another important bit of information. It is found that there is a high concentration of such clusters in a certain part of the sky and that most of them are about the same distance from us. Coupling this information with other evidence, we conclude that this concentration of clusters marks the center of our galaxy. We then know the distance to the center of the galaxy—about 10^{20} meters.

Knowing the size of our own galaxy, we have a key to the measurement of still larger distances—the distances to other galaxies. Figure 5-7 is a photograph of a galaxy, which has much the same shape as our own. Probably it is the same size, too. (Other evidence supports the idea that galaxies are all about the same size.) If it is the same size as ours, we can tell its distance. We measure the angle it subtends in the sky; we know its diameter, and we compute its distance—triangulation again!



Fig. 5-7. A spiral galaxy like our own. Presuming that its diameter is similar to that of our own galaxy we may compute its distance from its apparent size. It is 30 million light-years (3×10^{23} meters) from the earth.

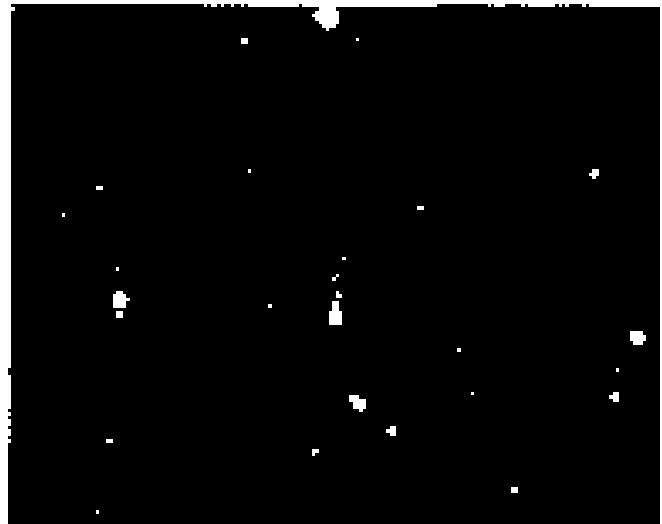


Fig. 5-8. The most distant object, 3C295 in BOOTES (indicated by the cross), measured by the 200-inch telescope to date (1960).

Photographs of exceedingly distant galaxies have recently been obtained with the giant Palomar telescope. One is shown in Fig. 5-8. It is now believed that some of these galaxies are about halfway to the limit of the universe— 10^{26} meters away—the largest distance we can contemplate!

5-7 Short distances

Now let's think about smaller distances. Subdividing the meter is easy. Without much difficulty we can mark off one thousand equal spaces which add up to one meter. With somewhat more difficulty, but in a similar way (using a good microscope), we can mark off a thousand equal subdivisions of the millimeter to make a scale of microns (millionths of a meter). It is difficult to continue to smaller scales, because we cannot "see" objects smaller than the wavelength of visible light (about 5×10^{-7} meter).

We need not stop, however, at what we can see. With an electron microscope, we can continue the process by making photographs on a still smaller scale, say down to 10^{-8} meter (Fig. 5-9). By indirect measurements—by a kind of triangulation on a microscopic scale—we can continue to measure to smaller and smaller scales. First, from an observation of the way light of short wavelength (x-radiation) is reflected from a pattern of marks of known separation, we determine the wave-

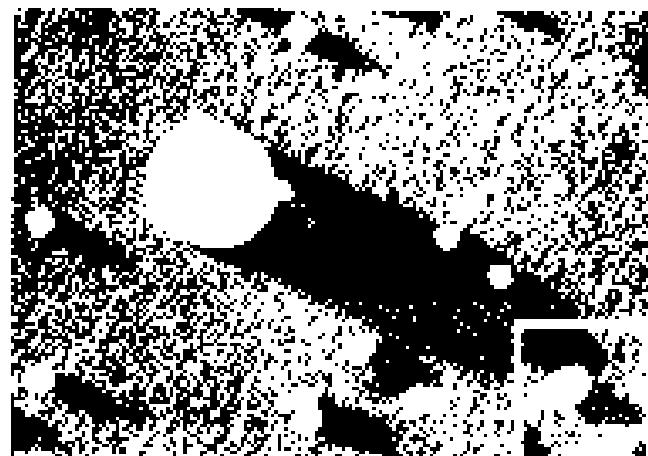


Fig. 5-9. Electron micrograph of some virus molecules. The "large" sphere is for calibration and is known to have a diameter of 2×10^{-7} meter (2000 Å).

DISTANCES	
LIGHT-YEARS	METERS
10^{27}	10^{27} meters
10^1	Edge of universe
10^6	Nearest neighbor galaxy
10^{21}	To center of our galaxy
10^8	To nearest star
10^{15}	Radius of orbit of Pluto
10^{11}	To the sun
10^9	In the moon
10^3	Height of a Spurri
10^2	Height of a TV antenna tower Height of a child
10^{-3}	A grain of salt
10^{-4}	A virus
10^{-5}	Radius of an atom
10^{-10}	Radius of a nucleus
10^{-13}	10^{-13} meters

length of the light vibrations. Then, from the pattern of the scattering of the same light from a crystal, we can determine the relative location of the atoms in the crystal, obtaining results which agree with the atomic spacings also determined by chemical means. We find in this way that atoms have a diameter of about 10^{-10} meter.

There is a large "gap" in physical sizes between the typical atomic dimension of about 10^{-10} meter and the nuclear dimensions 10^{-13} meter, 10^3 times smaller. For nuclear sizes, a different way of measuring size becomes convenient. We measure the *apparent area*, σ , called the effective *cross section*. If we wish the radius, we can obtain it from $\sigma = \pi r^2$ since nuclei are nearly spherical.

Measurement of a nuclear cross section can be made by passing a beam of high-energy particles through a thin slab of material and observing the number of particles which do not get through. These high-energy particles will plow right through the thin cloud of electrons and will be stopped or deflected only if they hit the concentrated weight of a nucleus. Suppose we have a piece of material 1 centimeter thick. There will be about 10^8 atomic layers. But the nuclei are so small that there is little chance that any nucleus will lie behind another. We might *imagine* that a highly magnified view of the situation—looking along the particle beam—would look like Fig. 5-10.

The chance that a very small particle will hit a nucleus on the trip through is just the total area covered by the profiles of the nuclei divided by the total area in the picture. Suppose that we know that in an area A of our slab of material there are N atoms (each with one nucleus, of course). Then the total area "covered" by the nuclei is $N\sigma/A$. Now let the number of particles of our beam which arrive at the slab be n_1 and the number which come out the other side be n_2 . The fraction which do *not* get through is $(n_1 - n_2)/n_1$, which should just equal the

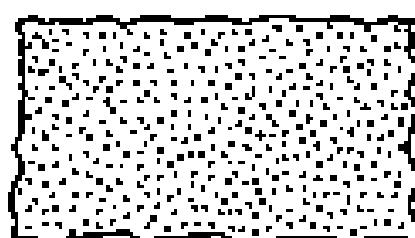


Fig. 5-10. Imagined view through a block of carbon 1 cm thick if only the nuclei were observed.

fraction of the area covered. We can obtain the radius of the nucleus from the equation*

$$\pi^2 = \sigma = \frac{A}{N} \frac{\sigma_0 - \sigma_1}{\sigma_L}$$

From such an experiment we find that the radii of the nuclei are from about 1 to 6 times 10^{-13} meter. The length unit 10^{-13} meter is called *the fermi*, in honor of Enrico Fermi (1901-1958).

What do we find if we go to smaller distances? Can we measure smaller distances? Such questions are not yet answerable. It has been suggested that the still unsolved mystery of nuclear forces may be unravelled only by some modification of our idea of space, or measurement, at such small distances.

It might be thought that it would be a good idea to use some natural length as our unit of length—say the radius of the earth or some fraction of it. The meter was originally intended to be such a unit and was defined to be $(\pi/2) \times 10^{-13}$ times the earth's radius. It is neither convenient nor very accurate to determine the unit of length in this way. For a long time it has been agreed internationally that the meter would be defined as the distance between two scratches on a bar kept in a special laboratory in France. More recently, it has been realized that this definition is neither as precise as would be useful, nor as permanent or universal as one would like. It is currently being considered that a new definition be adopted, an agreed-upon (arbitrary) number of wavelengths of a chosen spectral line.

Measurements of distance and of time give results which depend on the observer. Two observers moving with respect to each other will not measure the same distances and times when measuring what appear to be the same things. Distances and time intervals have different magnitudes, depending on the coordinate system (or "frame of reference") used for making the measurements. We shall study this subject in more detail in a later chapter.

Perfectly precise measurements of distances or times are not permitted by the laws of nature. We have mentioned earlier that the errors in a measurement of the position of an object must be at least as large as

$$\Delta x = h/\Delta p,$$

where h is a small quantity called "Planck's constant" and Δp is the error in our knowledge of the momentum (mass times velocity) of the object whose position we are measuring. It was also mentioned that the uncertainty in position measurements is related to the wave nature of particles.

The relativity of space and time implies that time measurements have also a minimum error, given in fact by

$$\Delta t = h/2E,$$

where $h/2E$ is the error in our knowledge of the energy of the process whose time period we are measuring. If we wish to know *more* precisely *when* something happened we must know less about *what* happened, because our knowledge of the energy involved will be less. The time uncertainty is also related to the wave nature of matter.

* This equation is right only if the area covered by the nuclei is a small fraction of the total, i.e., if $A/N \ll \pi R^2$ is much less than 1. Otherwise we must make a correction for the fact that some nuclei will be partly obscured by the nuclei in front of them.

Probability

"The true logic of this world is in the calculus of probabilities."

—Isaac Todhunter

6-1 Chance and likelihood

"Chance" is a word often used without real everyday living. The radio reports speaking of tomorrow's weather may say, "There is a 50% chance of rain." You might ask, "There is a 50% chance that it won't rain for the next ten years?" Because after all, there were chance. A seismologist may be interested in the question, "What is the chance that there will be an earthquake somewhere in Los Angeles in California next year?" A physicist might ask the question, "What is the chance that a particular geiger counter will register radioactive counts in the next ten seconds?" A politician or a citizen might be interested in the question, "What is the chance that there will be a nuclear war within the next ten years?" You may be interested in the chance that you will learn something here. In chapter 1.

By chance, we mean something like a guess. Why do we make guesses? We make guesses when we wish to make a decision but have incomplete information or uncertain knowledge. We want to make a guess as to what things are, or what things can happen. Often we wish to make a guess because we have to make a decision. For example, shall I take my chances with the incorrect? For what in the moment should I design a new building? Shall I build myself a fallout shelter? Shall I change my goals in entrepreneurial negotiations? Shall I go to class today?

Sometimes we make guesses because we wish, with our limit of knowledge, to say as much as we can about some situation. Really, any generalization is of the nature of a guess. Any theorized theory is a kind of guesswork. There are several types of more or less educated guesses. The theory of probability is a way for formulating better guesses. The language of probability allows us to speak specifically about some situation which may be highly variable, but which does have some consistent, average behavior.

Let us consider the flipping of a coin. If the two sides head and tail are "heads" we base our way of knowing what to expect on the number of experiments done. You would flip a coin a large number of times; there should be about equal numbers of heads and tails. We say, "The probability that a toss will land heads is 0.5."

You speak of probability only for observations carried out because being made in the future. By the "potential" of a particular outcome of an observation we mean our estimate for the most likely frequency of a number of repeated observations that will yield that particular outcome. If we imagine repeating an observation such as tossing a die 1000 times, and if we call N_j the outcome of the most likely number of outcomes that will give some specified result j , say the result "heads," then by $N_j/1000$, the probability of observing j , we mean

$$P(j) = N_j/1000 \quad (6.1)$$

Our definition requires several comments. First of all, we may speak of a probability of something happening only if the outcome is a possible outcome of some reasonable observation. It is not clear exactly what makes any sense at all. "What is the probability that there is a ghost in that house?"

6-1 Chance and likelihood

6-2 Likelihood

6-3 The random walk

6-4 A probability distribution

6-5 The uncertainty principle

You may object that no situation is exactly repeatable. That is right. Every different observation must have at least one different aspect. All we can say is that the "unrepeated" observations coincide for our intended purpose, appear to be equivalent. We should assume, at least, that each observation was made from an approximately parallel situation, and especially with the same day or of approximately the same. (If we could do experiments here in a controlled way, our estimates of the chance of winning are different than if we do not.)

We should emphasize the word "best". In Fig. 16, the curve is entitled to represent numbers based on actual observations. It is our best estimate of what would occur in a large number of trials. Probability depends therefore on our knowledge and our state of mind or estimate. In this, you cannot argue with it! But this only means there is a certain amount of agreement in the estimated sense on many things, so that different people will make the same estimate. Probabilities need not however, be "theoretical" numbers. Since they depend on our ignorance, they may become different if our knowledge changes.

You may have noticed another ratio between we expect of our daily coin of probability. We have referred to it as our estimate of the most likely number... I mean that we expect to obtain exactly N_0 , but that we expect a number near N_0 , and then N_1 numbers. N_1 is more likely than any other number in the scoring. If we toss a coin, say, 10 times, we would expect that the number of heads would not be very likely to be exactly 10. But a close value would be more likely, say, 9, 10, 11, 8, 9, 10, 11. However, if we knew closely, we would decide that 10 heads is about twice that any other number. We would write $P(N_0) = 0.5$.

Why did we choose to be more ready than my sister is? Well, we might have argued that because of the "theoretical" number of heads, the likely number of heads is N_0 , i.e., a total number of tosses N . But the true likely number of heads is $N_0 + \Delta N$. (We are assuming that every toss gives either heads or tails, and no "other" result.) So, it seems "most likely" there is an excess of heads to tails. On the basis some assumes this is the case (or "as good"), we might give a "theoretical" for heads and tails. So we might set $N_0 = N_1 = N_0$. It follows that $N_0 = N_1 = N/2$, or $P(N_0) = P(N_1) = 0.5$.

We can generalize our reasoning to any situation in which there are no restrictions to "hypothetical" (that is, equally likely) possibilities of making a hit. Our particular example is to probability that a "single draw" from a scuttled deck of 52 cards will show the ten of clubs in \mathcal{J}_0 . The probability of finding a card with a picture is $1/52$.

If there are several different colored balls in an opaque box and we pick one out, "at random" (that is, without looking), the probability of getting a ball of a particular color is c . To probability that a "single draw" from a scuttled deck of 52 cards will show the ten of clubs is \mathcal{J}_0 . The probability of finding a card with a picture is $1/52$.

In Chapter 5 we described the idea of a "series" in various ways and one of them went like this: "When we did so we were talking about probabilities." When we drew a single card, probability then was known. There is no chance that it will pass right through and come through that hole with a cushion. Once the needle is in one, that's our control over it, we probably are right at 10 cards. We just don't know if it will be a diamond, a heart, and the rest of the cards have a cross-hatch, or not, then the card will "pass through" by the needle, etc. In a large number of random draws, we expect certain "number" of hits, N_0 , of course, will be in the neighborhood of the total number of the cards in the deck.

$$N_0/N = n_0/d$$

(6.2)

We note n_0/d , that is, that the probability that a single projectile particle will make a collision with a target atom is n_0/d .

$$P_0 = \frac{n}{d} \cdot d$$

(6.3)

where n/d is the number of atoms per unit cross section.

6.3 FLUCTUATIONS

We would like now to remember what probability we consider to consider all same gender detail can be asked: "How many heads do I really expect to get if I flip a coin $N = 10$?" Before answering the question, however, it is good to what does happen in such an "experiment". Figure 6-1 shows the results obtained in the first class "first" of 1000 experiments in which $N = 10$. The sequences of "heads" and "tails" are known just as they were recorded. The first game gave 11 heads; one second also had the same 11. In three trials we did not get 10 heads. Should we begin to suspect the coin? Or even, knowing in thinking that the most likely number of "heads" in such a game is 15? Many tests made 7023 were made to obtain a total of 100 experiments of 100 tosses each. The results of the experiment are given in Table 6-1.

Table 6-1
Number of heads in successive trials of 100 tosses of a coin

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	12	17	17	15	21	23	14	16	17	17	17	17	17	17	17	17	17	17	17	
2	17	17	15	15	15	17	17	17	17	17	17	17	17	17	17	17	17	17	17	
3	17	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	
4	12	11	11	12	12	21	12	12	12	12	12	12	12	12	12	12	12	12	12	
5	10	13	13	14	16	16	15	13	13	13	13	13	13	13	13	13	13	13	13	
6	14	13	16	15	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	
7	11	16	16	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	
8	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	
9	17	17	17	14	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	

H	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
-2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	

Fig. 6-1. Observed sequence of heads and tails in 100 games of 30 moves each.

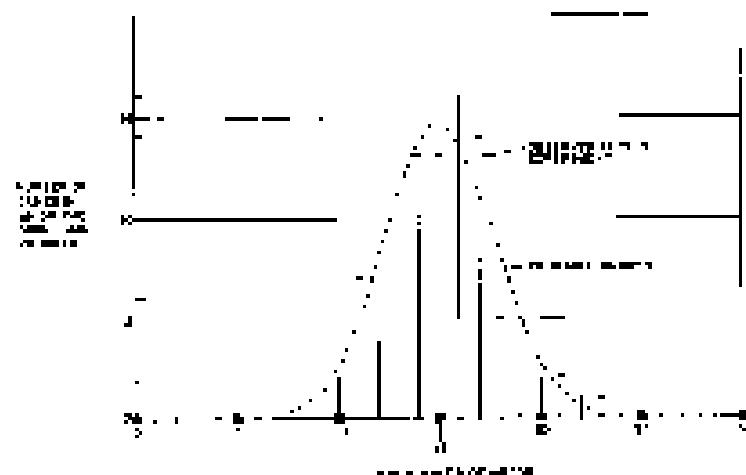


Fig. 6-2. Summary of the results of 1000 games of 100 moves each. The vertical bars show the number of games in which a score of H was obtained. The dashed curve shows the expected number of games with the score H obtained by a probability theory.

Looking at the numbers in Table 6-1, we see that some of the results are "surprised"; indeed they are between 12 and 13. We can gain a better feeling for the details of their results if we plot a graph of the distribution of the results. We start the "number of games" in which a score of H was obtained, and put the numbers for each H . Such a graph is shown in Fig. 6-2. A score of 14 heads was obtained in 11 games. A score of 15 heads was also obtained 11 times. Scores of 16 and 17 were each obtained 12 times. At 12, 13, and 18 we see 10 games. Is there something unusual? Was our "luck" not good enough?"

* After the first three games, the game theory was actually lost by taking 27 photons to exactly 100, but these forming the number of heads that showed.

we conclude now that the "most likely" score for a game of 10 tosses is not 50% heads? Strangely, in all the games in our together, there were 100 tosses. And the mean number of heads obtained was 14.92. The fraction of tosses that gave heads is 0.497, very nearly, but slightly less, than half. We should certainly not assume that the probability of obtaining heads is greater than 0.5. The fact that one particular set of characters just happened most often is no argument. We still expect that the most likely number of heads is 5.

We may ask the question, "What is the probability that a game of 10 tosses will yield 5 heads—or 6 or any other number?" We know well that in a game of one toss, the probability of obtaining one head is 0.5, and the probability of obtaining no head is 0.5. In a game of 10 tosses there are now possible outcomes: HHH, HTT, and so on. Since each of these sequences is equally likely, we conclude that (a) the probability of a score of two heads is $\frac{1}{5}$, (b) the probability of a score of one head is $\frac{2}{5}$, (c) the probability of zero heads is $\frac{1}{5}$. There are two ways of obtaining one head, but only one of obtaining either zero or two heads.

Consider next a game of 5 tosses. The same toss is equally likely to be heads or tails. There is only one way to obtain 3 in 5, we must have obtained 2 heads or 2 tails and two tails, and then, since on the last. Therefore, there are 10 ways of obtaining 2 heads. We know three odds after having thrown 100 heads (see page 20); we could have heads after the 50th only one head in the first 50 tosses (one way), for scores of 1-50, 2-50, 3-50, 4-50, 5-50 we have then the number of equally likely scores 5, 1, 3, 3, 1, with a total of 16 different possible sequences. The probabilities are in Fig. 6.3.

The argument we have been going over by diagrammatical analogy like that in Fig. 6.3. It is clear how the diagram should be extended for games with a larger number of tosses. Figure 6.4 shows such a diagram for a game of 6 tosses. The number of "ways" to any point on the diagram is just the number of different "path" sequences of heads and tails which can be taken from the starting point. The vertical position gives us the total number of heads thrown. The set of numbers which appears in such a diagram is known as Pascal's triangle. The numbers are also called the binomial coefficients because they also appear in the expansion of $(a + b)^n$. If we count the number of tosses and the number of heads thrown, the 16 numbers in the diagram are usually denoted by the symbol $\binom{n}{k}$. We may note that in passing that the binomial coefficients can also be computed from

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad (6.4)$$

where $n!$, or "factorial," represents the product $(n)(n-1)(n-2)\dots(2)(1)$.

We are now ready to compute the probability $P(k)$, or of throwing k heads in n tosses, using our definition Eq. (6.1). The total number of possible sequences is 2^n (since there are 2 outcomes for each toss) and the number of ways of getting k heads is $\binom{n}{k}$, all equally likely, so we have

$$P(k, n) = \frac{\binom{n}{k}}{2^n}. \quad (6.5)$$

Since $P(k, n)$ is the fraction of games which we expect to yield k heads, then in 100 games we should expect to find 5 heads 100 $\cdot P(5, 100)$ times. The actual curve in Fig. 6.2 passes through the points computed from $100 \cdot P(k, 100)$. We see that we never in about a score of 15 heads in 14 in 15 games, whereas the score was observed in 13 games. We expect a score of 16 in 15 or 16 games, but we obtained that score in 16 \pm 6. Such fluctuations are "part of the game."

This method we have used can be applied to the most general situation in which there are many possible outcomes of a single observation. Let us designate the n -outcomes by M (for "multi;" and L , "less")¹. In the general case, the probability of M or L in a single event need not change. Let p_M be the probability of obtaining the result M . Then, if the probability of L , is necessarily



Fig. 6.3. A diagram for showing the number of ways to score 0, 1, 2, or 3 heads in a game of 5 tosses.

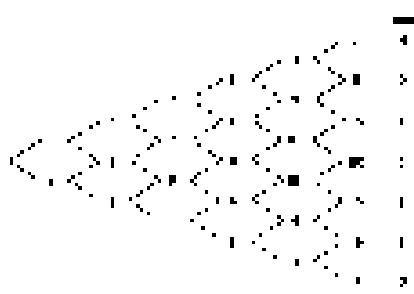


Fig. 6.4. A diagram like that of Fig. 6.3, for a game of 6 tosses.

(1) $\forall i$: In a set of events, the probability $P(E_i)$ that E_i will be observed is given by

$$P(E_i) = \text{exp}(E_i^2 \tau)^{-1} \quad (6.6)$$

This probability function is called the *exponential distribution probability*.

6-3 The random walk

There is another interesting problem in which the idea of probability p is required. Let's consider a "one-dimensional walk." In its simplest version, we imagine a "game" in which a "player" starts at the point $x = 0$ and at each "turn" is required to take a step either forward (forward $+x$) or backward (backward $-x$). The driving force behind such a walk is, for example, the presence of a cloud. You can't see through the foggy weather. In its general form, the random walk is related to the motion of atoms (or other particles) in a gas (cf. Brownian motion), and can be the random walk of some measurements. You will see that the random walk problem is closely related to the coin tossing problem we have already discussed.

First let us look at a few examples of a random walk. We may characterize the walker's progress by the net distance D traveled in N steps. We show in the graphs of Fig. 6-5 three examples of typical D 's a random walker (We have used four colors to distinguish the results of the coin tosses shown in Fig. 6-1)

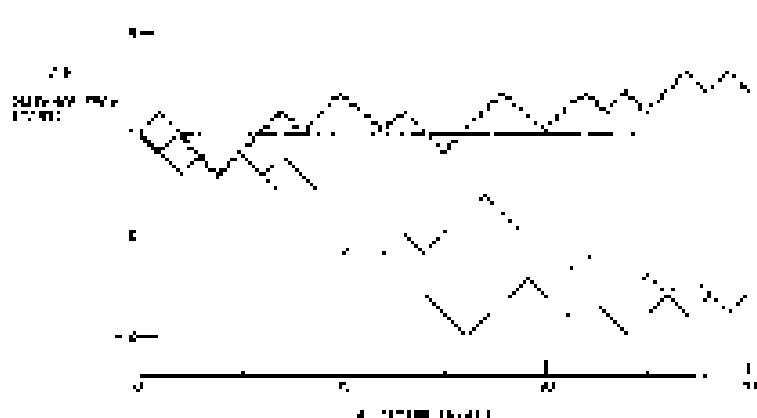


Fig. 6-5. The progress made in a random walk. The horizontal coordinate D is the total number of steps taken in the walk. Eccentric ERINI's last night he moved from his starting position.

What can we say about such a motion? We might first ask: "How far does he get on the average?" We may expect that the average progress will be zero, since he is equally likely to go either forward or backward. But we have the following situation: if, instead, he is more likely to have a good flight from the start up coast, we might, therefore, ask what is his average distance travelled in objective miles. I.e., is $\langle D \rangle$ the average of $|D|$? I.e., because it is convenient to deal with another measure of "progress," namely one of "the distance" D^2 is positive for either positive or negative motion, and is therefore a reasonable measure of such random walking.

We can thus use the expected value of D^2 as just $\langle D^2 \rangle$, the number of steps taken. By "expected value" we mean the probable value (our best guess), which we can think of as the expected average behavior in many repeated experiments. We represent $\langle D^2 \rangle$ as an expected value by $\langle D^2 \rangle$, and may refer to it also as the "mean squared distance." A relationship $\langle D^2 \rangle = \langle D \rangle^2 + \langle D^2 - D^2 \rangle$ is often approximately ($D^2 - D^2 \rangle \ll \langle D^2 \rangle$) true. Distances will be measured in terms of units of distance. We shall not concern here the units of distance.)

The expected value of D_N^2 for $N > 1$ can be obtained from D_{N-1} . If, after $(N-1)$ steps, we have $D_{N-1} = k$, then after N steps we have $D_N = D_{N-1} + 1$ or $D_N = D_{N-1} - 1$, with equal probability.

$$D_N^2 = \begin{cases} D_{N-1}^2 + 1 & \text{if } D_{N-1} = 1, \\ & \dots \\ D_{N-1}^2 - 2D_{N-1} + 1 & \text{if } D_{N-1} = N-1. \end{cases} \quad (6.7)$$

In a random of independent sequences, we expect to obtain each value one half of the time, so our average expectation is just the average of the two possible values. The expected value of D_N^2 is $\langle D_N^2 \rangle = 1$. In general, we should expect for D_{N-1}^2 the "expected value" $\langle D_{N-1}^2 \rangle$ by definition. So

$$\langle D_N^2 \rangle = \langle D_{N-1}^2 \rangle + 1 \quad (6.8)$$

We have already shown that $\langle D_0^2 \rangle = 1$; it follows that

$$D_N^2 = N, \quad (6.9)$$

(*per definitionem* $\langle D_N^2 \rangle = N$)

If we wish a number like a distance, rather than a distance squared, to represent the "progress made toward the origin" in a random walk, we can use the "root-mean-square distance," D_{rms} :

$$D_{\text{rms}} = \sqrt{\langle D^2 \rangle} = \sqrt{N}. \quad (6.10)$$

We have pointed out that the random walk is closely similar in its mechanics to the coin-tossing game we considered at the beginning of the chapter. Let's suppose the direction of each step is in correspondence with the appearance of heads or tails in a coin toss. If $D > 1$ just $N_0 = N_1$, the 2 times in N number of head-tail pairs. Since $N_0 = N_1 = N/2$, the total number of steps (two tosses), we have $D = 2N_0 = N$. We have derived earlier an expression for the expected distance, $\langle N_0 \rangle$, from the origin N_0 and, recall, the result of $\langle D \rangle$ (6.5). Since N is just a constant, we have the corresponding expression for D , namely for every head-tail pair, $N/2$ there is a tail-tail, "return" of 2 between N_0 and D). The graph of $\langle D \rangle$ vs N represents the distribution of distances, or "right get in 10 random steps until $D \approx 15$ is to be read $D \approx 3$; $N = 12$, $D = 2$, etc.)

The variance of D_N from its expected value $N/2$ is

$$M_{DD} = \frac{N}{2} = \frac{D}{2}, \quad (6.11)$$

so the correlation is

$$\left(\frac{D_N - \langle D \rangle}{\sqrt{M_{DD}}} \right)_{\text{rms}} = \frac{1}{2}\sqrt{N} \quad (6.12)$$

According to our result for D_{rms} , we expect that the "typical" distance is 10 steps, roughly $\sqrt{N} = \sqrt{30} = 5.5$, so a typical k should be about $5.5/2 = 2.8$ units from 15. We see that the "tail" of the curve in Fig. 6.2, measured from the center, is just about 3 units, in agreement with this result.

We are now in a position to consider a question we have avoided up to now. "How shall we tell whether a walk is "random" or "biased"?" We can give here at least a partial answer. For an honest walk, we expect the fraction of the times heads appears, $\langle N_0 \rangle/N$, to be,

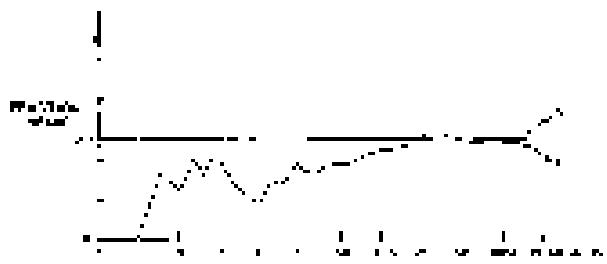
$$\frac{\langle N_0 \rangle}{N} = 0.5. \quad (6.13)$$

We also expect an overall N_0 to deviate from $N/2$ by about $\sqrt{N}/2$, or 2 or more to deviate by

$$\frac{\sqrt{N}}{N/2} = \frac{2}{\sqrt{N}}.$$

The larger N is, the closer we expect the fraction N_0/N to be to 0.5000.
6-3

Fig. 6-6. The fraction of the times that gave heads in a particular sequence of N heads or tails.



In Fig. 6-6 we have plotted the fraction N_H/N for the coin results reported earlier in this chapter. We see the tendency for the fraction of heads to approach 0.5 for large N . Unfortunately, for any given finite combination of runs, there is no guarantee that the observed frequency will be exact, even the expected deviation. There is always the chance that a large fluctuation, a "long run," of heads or tails will give an abnormally large deviation. All we can say is that if the deviation is near the expected value (as was often the case of 2 or 3), we have no reason to suspect the honesty of the coin. If it is much larger, we might begin to suspect, but cannot prove that the coin is loaded for tails (the fraction is lower).

We have also not considered how we should treat the case of a "short" or even single "long" run that always lands in either of two positions than we have just assumed: heads or tails. We have defined $P(H) = N_H/N$. How will we know what to expect for N_H ? In some cases, the best we can do is to measure the number of heads occurring in long numbers of tosses. In a way of anything better, we must set $N_H = P_H(\text{observed})$. (How could we expect anything else?) We must understand, however, that in such a case a different type of error, called the antecedent, might result in statistical significance. One would expect, however, that the various answers should agree within the certainty limit if $P(H)$ is near one half). All regions in which a result actually occurs that an "asymmetrical distribution" probability has an "error," one writes

$$P(H) = \frac{N_H}{N} \pm \frac{1}{\sqrt{N}}. \quad (6.14)$$

There is an implication in such an expression that there is a "true" or "natural" probability which could be compared if we knew enough, and that the observation may be in "error" due to a fluctuation. There is, however, no way to make such thinking logically consistent. It is probably better to realize that the probability concept is to a sense subjective, that it is always based on incomplete knowledge, and that its quantitative evaluation is subject to change as we obtain more information.

6-4 A probability distribution

Let us return now to the random walk and consider a modification of it. Suppose that in addition to a random choice of the direction (+ or -) of each step, the length of each step also varied in some unpredictable way, the only constraint being that on the average the step length was one unit. This case is more representative of something like the thermal motion of a molecule in a gas. If we let the length of a step be, then S may have any value, but nevertheless will be "near" 1. To be specific, we shall let $\langle S^2 \rangle = 1.0$, equivalently, $S_{\text{rms}} = 1$. Our definition for $\langle D^2 \rangle$ would proceed as before except that Eq. (6.8) would be changed now to read

$$\langle D^2 \rangle = \langle D_{\text{ave}}^2 \rangle + \langle S^2 \rangle = \langle D_{\text{ave}}^2 \rangle + 1.0. \quad (6.15)$$

We have, as before, that

$$\langle D_{\text{ave}}^2 \rangle = N. \quad (6.16)$$

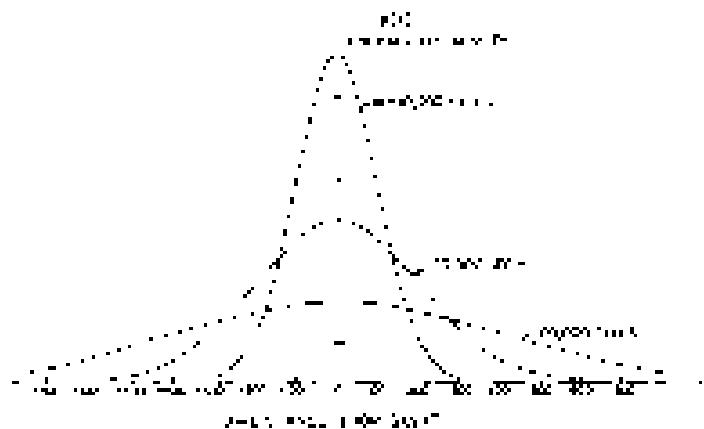


Fig. 6-7. The probability density for walking up or the distance D from the starting place in a random walk of N steps. (It is measured in units of the step length.)

What would we expect now for the distribution of distances D ? What is for example the probability that $D = 0$ after 40 steps? The answer is zero—the probability is zero that D will be an indefinite result, since there is no cause in all this, the sum of the successive steps (of varying lengths) would usually exceed the sum of forward steps. We cannot plot a graph like that of Fig. 6-2.

We can, however, obtain a representation similar to that of Fig. 6-2, if we ask, not what is the probability of obtaining D exactly equal to 0, 1, or 2, but instead what is the probability of obtaining D between 0, 1, or 2. Let us define $p(x, \Delta x)$ as the probability that D will lie in the interval Δx located at x (say from x to $x + \Delta x$). We expect that $p(x, \Delta x)$ is the fraction of D having D proportional to Δx , the width of the interval. So we can write

$$p(x, \Delta x) = p(x) \Delta x. \quad (6.13)$$

The function $p(x)$ is called the probability density.

The form of $p(x)$ will depend on N , the number of steps taken, and also on the distribution of individual step lengths. We cannot determine the $p(x)$ here, but for large N , $p(x)$ is the same for all reasonable distributions of individual step lengths, and depends only on N . We plot $p(x)$ for three values of N in Fig. 6-7. You will notice that the "bell-shaped" regions spread from $x = 0$ as these curves $\propto \sqrt{N}$, as we have shown it should be.

You may wonder that the value of $p(x)$ near zero is "curiously" proportional to \sqrt{N} . This occurs because the curves are all of a similar shape and ratio to one another. The areas must all be equal. Since $p(x) \Delta x$ is the probability of finding D in the interval Δx , we can determine the chance of finding D somewhere inside an arbitrary interval, from x_1 to x_2 , by cutting the interval in a number of small increments Δx and calculating the sum of the terms $p(x) \Delta x$ for each increment. The probability that D lies everywhere between x_1 and x_2 , which we may write $P(x_1 < D < x_2)$, is equal to the shaded area in Fig. 6-8. The smaller we take the increments Δx , the more correct is our result. We can write therefore,

$$P(x_1 < D < x_2) = \sum p(x) \Delta x = \int_{x_1}^{x_2} p(x) dx. \quad (6.14)$$

The area under the whole curve is the probability that D lies somewhere (that is, has some value between $x = -\infty$ and $x = +\infty$). That probability is surely 1. We must have that

$$\int_{-\infty}^{+\infty} p(x) dx = 1. \quad (6.15)$$

Stochastic curves in Fig. 5-7 go, while in proportion to \sqrt{R} , their heights must be proportional to $1/\sqrt{M}$ to maintain the total area equal to 1.

The probability density function we have been describing is one that is uncorrected mass moments. It is known as the normal or gaussian probability density. It has the mathematical form

$$p(v) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(v-\bar{v})^2}{2\sigma^2}}, \quad (6.20)$$

where σ is called the standard deviation and is given, in our case, by $\sigma = \sqrt{N}$ if the mass per gram is sufficient. Our $\bar{v} = \sqrt{RT}/M$.

We remember neither can the motion of a molecule, or of any particle, in a gas is like a random walk. Suppose we drop a small bit of organic compound and its mass of it vapor escape into the air. If there is no air currents, so that the air is circulating, the excess will also carry the vapor with them. But even in perfectly still air the vapor will gradually spread out until it has permeated throughout the room. We might detect it by its color or odor. The individual molecules of the organic vapor spread out in ∞ as because of the molecular collisions caused by all sorts of other molecules. If we know the average "step" size, say the number of steps is b in 10^{-10} sec., we can find the probability that one, or several molecules will be found at some distance from their starting point after any given time, t sec. As time passes, more steps are taken and the gas spreads out as in the successive curves of Fig. 5-7. In a later chapter, we shall find out how the step sizes and step frequencies are related to the temperature and pressure in a gas.

Earlier, we said that the pressure of a gas is due to the molecules bumping against the walls of the container. When we come later to make a more quantitative description, we will wish to know how fast the molecules are going. Why? because, since the impact they make will depend on that speed. We cannot, however, speak of the speed of the molecules. It is necessary to use a probability description. A molecule may have any speed, but some speeds are more likely than others. We describe what is going on by saying that the probability that any particular molecule will have a speed between v and $v + \Delta v$ is $p(v) \Delta v$, where $p(v)$ is probability density, a given function of the speed v . We shall see later how Maxwell, using common sense and the idea of probability, was able to find a mathematical expression for $p(v)$. The result of the form $p(v) \Delta v$ is shown in Fig. 6-9. Velocities may have any value, but are most likely to be near the true particle in expected value (\bar{v}).

We consider first of the curves of Fig. 6-9 in a somewhat different way. If we consider the molecules in a typical container (with a volume of, say, one liter), then there are a vast large number, N , of molecules present ($N \approx 10^{23}$). Since $p(v) \Delta v$ is the probability that one molecule will have its velocity in Δv by our definition of probability we mean that the expected number (ΔN) to be found with a velocity in the interval Δv is given by

$$\langle \Delta N \rangle = N p(v) \Delta v. \quad (6.21)$$

We call $\langle \Delta N \rangle$ the "fully situated in velocity." The area under the curve for two velocities v_1 and v_2 (for example the shaded area in Fig. 6-9, representing for the curve $p(v)$) the expected number of molecules with velocities between v_1 and v_2 . Since with a gas we are usually dealing with large numbers of molecules, we expect the deviations from the expected numbers to be small (like $1/\sqrt{N}$), so we often neglect to say the "expected" number, and say instead: "The number of molecules with velocities between v_1 and v_2 is the area under the curve." We should remember, however, that such statements are always about probable numbers.

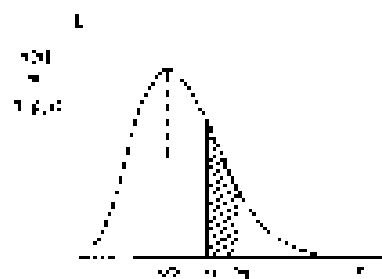


Fig. 6-9. The distribution of velocities of the molecules in a gas.

* Maxwell's expression is $\langle \Delta N \rangle = C v^2 / \sqrt{\pi R T}$, where C is a constant related to the mass density and C is chosen so that the total probability is one.

6-5 The uncertainty principle

The theory of probability was originally used in describing the behavior of the 10^{23} or so molecules in a sample of a gas, for it is clearly unphysical even to attempt to write down the position or velocity of each molecule. When probability was first applied to such problems, it was considered to be a convenience, a way of dealing with very complex situations. We now believe that the ideas of probability are essential in a description of atomic properties. According to quantum mechanics, the mathematical theory of particles, there is always some uncertainty in the specification of positions and velocities. We can, in fact, say that there is a certain probability that any particle will have a position near some coordinate x .

We can give a probability density $p_1(x)$ such that $p_1(x)dx$ is the probability that the particle will be found between x and $x + dx$. If the particle is reasonably well localized, say near x_0 , the function $p_1(x)$ might be given by the graph of Fig. 6-1(a). Similarly, we can specify the velocity of the particle by means of a probability density $p_2(v)$, with $p_2(v)dv$ the probability that the velocity will lie between v and $v + dv$.

It is one of the fundamental tenets of quantum mechanics that the two functions $p_1(x)$ and $p_2(v)$ cannot be chosen independently since, in particular, x and v must inherently interact. If we call the "width" of the $p_1(x)$ curve Δx , and that of the $p_2(v)$ curve Δv , the factor in the figure, $\Delta x \Delta v$, denotes that the product of the two will be at least as big as the number $\hbar/2\pi$, where \hbar is the mass of the particle and Δ is a Greek letter δ which means "and" or that "there is". We may write this basic relationship as

$$(\Delta x)(\Delta v) \geq \hbar/2\pi. \quad (6-22)$$

This equation is a statement of the Heisenberg uncertainty principle that we mentioned earlier.

Since the right-hand side of Eq. (6-22) is a constant, the equation says that if we try to "pin down" a particle by trying to fix its position or place it goes up by having a high speed. Or if we try to force it to go very slowly, or at a precise velocity, it "spreads out" so that we do not know very well just where it is. Particles behave in a funny way!

The uncertainty principle describes an inherent fuzziness that just can't, in any attempt, be described away. Our most precise description of nature must be in terms of probabilities. There are some people who do not like this way of describing nature. The best example that I have could only tell $\hbar/2\pi$ to myself, you will never live; they would know its speed and position him however. In the early days of the development of quantum mechanics, Heisenberg was quite worried about his publisher. He went to Ehrenfest and said, "He, I simply find that I don't know how to determine how electrons should go!" He worried about this problem for a long time and he really never really committed himself to the fact that this is the best description of nature that we can give. There are still one or two physicists who are working on the subject who have an intuitive conviction that it is possible somehow to describe the world in a definite way and that all of this uncertainty about the way things are set to removed. He has not been successful.

The necessary uncertainty in our specification of the position of a particle becomes most important when we wish to describe the structure of atoms. In the hydrogen atom, which has a nucleus of one proton with one electron outside of the nucleus, the uncertainty in the position of the electron is so large, as the atom itself is. We cannot, therefore, properly speak of electron buzzing in some "orbit" around the proton. The result we can say is that there is a certain chance per unit volume of finding the electron in an element of volume ΔV at the distance r from the proton. This probability density $\rho(r)$ is given by quantum mechanics. For an undisturbed hydrogen atom $\rho(r) = 4\pi r^2 n$, which is a bell-shaped function like that in Fig. 6-5. This is known as the "hydrogen" atom, where the function is decreasing rapidly. Since there is a solid probability of finding the electron at the mass of a proton.

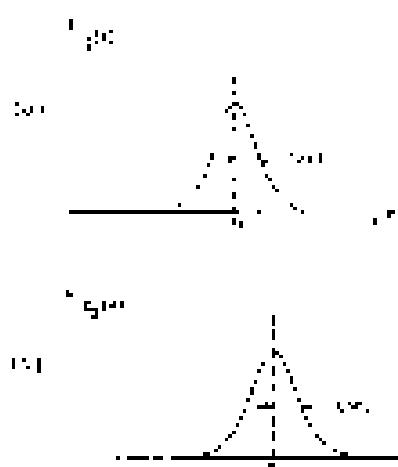
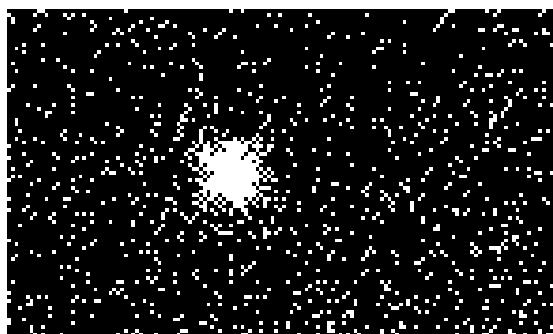


Fig. 6-1. Probability densities for observation of the position and velocity of a particle.

Fig. 6.11. A way of visualizing a hydrogen atom. The density (whiteness) of the cloud represents the probability density for observing the electron.



from the nucleus which greater than a , we may think of a as "the radius of the atom," about 10^{-10} meters.

We can form an image of the hydrogen atom by calculating a "cloud" whose density is proportional to the probability density for observing the electron. A sample of such a cloud is given in Fig. 6.11. This our best "picture" of a hydrogen atom is a nucleus surrounded by an "electron cloud" (although we really mean a "probabilistic cloud"). The electron is there somewhere, but "where" permits us to know only the chance of finding it at any particular place.

In its efforts to learn as much as possible about the atom, nuclear physics has found that certain things can never be "known" with certainty. Much of our knowledge must always remain uncertain. The word we can know is in terms of probabilities.

The Theory of Gravitation

7-1 Planetary motion

In this chapter we shall extend one of the most far-reaching generalizations of the human mind. While we are extending the human mind, we should take some time now to recall in case of anyone that could follow with sufficient interest and generality even an elementary science principle as the law of gravitation. What is the law of gravitation? It is that every object in the universe attracts every other object with a force which the two bodies in question¹ to the mass of each other varies inversely as the square of the distance between them. This statement can be expressed mathematically by the equation

$$F = G \frac{Mm}{r^2}$$

Now the we add the fact that an object responds to a force by accelerating in the direction of the force by an amount that is inversely proportional to the mass of the object, we shall have seen everything you need, for a sufficiently brilliant mathematician could then deduce all the consequences of these two principles. However, since you are not expected to be sufficiently brilliant as yet, we shall discuss the consequences in more detail, and not just leave you with only these two bare principles. We shall briefly relate the story of the discovery of the law of gravitation and discuss some of its consequences, its effects on history, the masses can prove a law easily, see some references of the law—say my favorite: we shall also discuss the relationships of the law to the other laws of physics. All this cannot be done in one chapter, but these subjects will be treated in due time in subsequent chapters.

The story begins with the ancients inserting the motions of planets among the stars, and finally showing that they were around the sun, a fact that was rediscovered later by Copernicus. Exactly how the planets were around the sun, with exactly what motion, took a little more work to discover. In the beginning of the 17th century there were great debates over whether they really went around the sun or not. Tycho Brahe had an idea that was different from anything proposed by the ancients. His idea was that these debates about the nature of the motions of the planets would be unnecessary if predicted positions of the planets in the sky were measured sufficiently accurately. If measurement showed exactly how the planets moved, then perhaps it would be possible to establish one or another viewpoint. This was a tremendous idea—that to find something out, is better to make some accurate experiments than to carry on deep philosophical arguments. Pursuing this idea, Tycho Brahe studied the position of the planets for many years in his observatory in the island of Hven, near Copenhagen. He made voluminous tables, which were then studied by the mathematician, Kepler, after Tycho's death. Kepler discovered from the data some very beautiful and remarkable relationships regarding planetary motion.

7-2 Kepler's laws

First of all, Kepler found that each planet goes around the sun in a curve called an ellipse, with the sun at a focus of the ellipse. An ellipse is not just one thing; there is a very specific and precise curve that can be obtained by using two forces, one at each focus, a force of引力 and a force of repulsion; mathematically, it

7-1 Planetary motion

7-2 Kepler's laws

7-3 Development of dynamics

7-4 Newton's law of gravitation

7-5 Universal gravitation

7-6 General relativity

7-7 What is gravity?

7-8 Gravity and relativity

is defined as equal to the sum of whose distances from two fixed points (the foci) is a constant. Or if you will, it is a function of distance (Fig. 7-1).

Kepler's second observation was that the planets do not go around the sun in uniform speeds, but move faster when they are nearer the sun, and move slowly when they are farther from the sun, in precisely this way. Suppose a planet is observed at any two successive times, let us say a week apart, and that the radius vector is shown in the figure. In each time its position. The white arc is covered by the planet during the week, and the two radii require, since a certain plane now, the shaded area shows it (Fig. 7-2). If two similar observations are made a week apart at a pair of the orbit, farther from the sun, then the planet moves more slowly, the primary bounded area is greater than that in the first case. So, in accordance with the second law, the orbital speed of each planet is such that the same "average" equal areas in equal times.

Finally, a third law was discovered by Kepler much later; this law is of a different character. From the other two, however, it does not with only a single short jump relate all the planets to one another. This law says that when the orbital periods and orbit sizes of any two planets are compared, the periods are proportional to the $3/2$ power of the orbit size. In this statement the period is the time interval it takes a planet to go completely around its orbit, and the size is measured by the length of the greatest diameter of the elliptical orbit, technically known as the major axis. More simply, if the planets went in circles, as they nearly do, the time required to go around the circle would be proportional to the $3/2$ power of the diameter (or radius). Thus Kepler's three laws are:

- I. Each planet moves around the sun in an ellipse, with the sun at one focus.
- II. The areas swept from the sun to the planet, whether real or imaginary, are equal intervals of time.
- III. The squares of the periods of any two bodies are proportional to the cubes of the semimajor axes of their respective orbits: $T^2 \propto a^{3/2}$.

7-2 Development of dynamics

While Kepler was discovering these laws, Galileo was studying the laws of motion. The problem was, what makes the planets go around? (In those days one of the theories proposed was that the planets went around because holding them were invisible angels, keeping them and driving the planets forward. You will see that this theory is now modified. It is assumed that in order to keep the planets going around, the invisible angels take them in a different direction and they have no wings. Otherwise, it is a conceivable celestial theory!) Galileo was interested in very remarkable forces when he lived, which was unusual for understanding these laws. That is, the principle of inertia—of something moving with uniform speed, resists changing its uniformity, once started, it will go on forever, existing at a uniform speed in a straight line. (Why does it keep on curving? We do not know, but that is the way it is.)

Newton modified this idea, saying that the only way to change the motion of a body is to use force. If no body speeds up, a force has been applied to the vector of motion. On the other hand, if its motion is changed to a new direction, a force has been applied, also. Newton thus asked the question: If there is a force, it must change the speed or the direction of motion of a body. For example, if a stone is released to falling and is pulled around in a circle, it takes a force to keep it in the circle. We have to pull on the string. In fact, it is less so that the acceleration produced by the force is inversely proportional to the mass, or the force is proportional to the mass times the acceleration. The more massive a thing is, the stronger the force required to produce a given acceleration. (The mass can be measured by pulling other stones on the end of the same string and making them go around the same circle at the same speed. In this way it is found that more or less force is required, the more massive object requiring more force.)

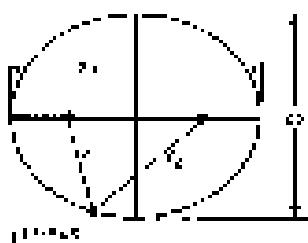


Fig. 7-1. An ellipse.

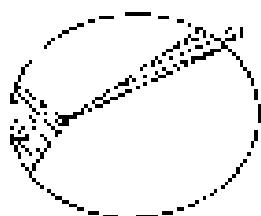


Fig. 7-2. Kepler's law of areas.

* A radius vector is a line drawn from the sun to the point in a planet's orbit.

The brilliant idea resulting from these considerations is that no mechanical force is needed to keep a planet in orbit (the *Leviathan* and *Principia* by sympatheticity) because the planet would move in that direction anyway—if there were nothing at all to distract the planet would go off in a straight line. But the actual motion deviates from the line on which the body would have gone if there were no force; this gives the body necessarily an angle relative to the moving line in the direction of its motion. In other words, because of the principle of inertia, the force needed to cause the motion of a planet around the sun is not a force around the sun but around the sun. (If there is a force toward the sun, the star might be the angel, *“because”*.)

2.4 Newton's law of gravitation

From his better understanding of the theory of motion, Newton appreciated that it was could be the attractive nature of forces that govern the motion of the planets. Newton wrote to himself (and perhaps we may be able to prove it soon) that “the very fact that celestial motions are swept out in equal times is a pretty argument of the composition of forces” (his italics) and “certainly motion—then. Below of course is a direct consequence of this fact, that all of the forces are directed exactly toward the sun.”

Next, by analyzing Kepler's third law it is possible to show that the further away the planet, the weaker the forces. If two planets of different distances from the sun are compared, the analysis shows that the forces are inversely proportional to the square of the respective distances. With the compilation of the laws, Newton concluded that there must be a force, inversely as the square of the distance, directed in a line between the two objects.

Bring in more of evidence in favoring for gravity. The Newton suggestion, of course, was this relationship applied more generally than just to the sun holding the planets. It was already known, for example, that the planet Jupiter had moons going around it as the moon of the earth goes around the earth, and Newton felt certain that each planet held its moons with a force. He already knew of the force holding a man to the earth, so he imagined that this was a universal force that everything held everything else.

The next problem was whether the pull of the earth on its people was the same as a pull on the moon, also inversely as the square of the distance. If an object on the surface of the earth falls 16 feet in the first second, then it is agreed from Galileo¹ that does the moon fall in the same time? We might say that the moon does not fall at all. But if the moon fell as the earth, it would go off in a straight line, rather than goes in a circle instead, so it really falls off from where it would have been if there were no force at all. We can calculate from the radius of the moon's orbit (which is about 240,000 miles and how long it takes to go around the earth (approximately 29 days), how far the moon moves in its orbit in 1 second, and can then calculate how far it falls in one second.^{*} This distance turns out to be, roughly 1/30 of a mile in a second. That fits very well with the inverse square law, because the earth's radius is 3000 miles, and if something which is 240,000 miles from the center of the earth lets 16 feet in a second, something 3000 miles, or 1/80 as far away, should fall only 1/80th of 16 feet, which also is roughly 1/30 of a foot. Trying to put this theory of gravitation to a test by a mile calculation, Newton made his calculations very carefully, and found a discrepancy so large that he hardly was satisfied by facts, and did not publish his results. Six years later a new measurement of the size of the earth showed that the astronomical had been using an incorrect distance to the moon, when they measured it, thus he made the calculations again, and the corrected figures, also obtained beautiful agreement.

This idea that the term “force” is somewhat misleading, because, as you see, it does not cause any change. This idea is sufficiently interesting to merit further

* That is how far the end of the object will fall below the straight line tangent at the point where the object was suspended before.

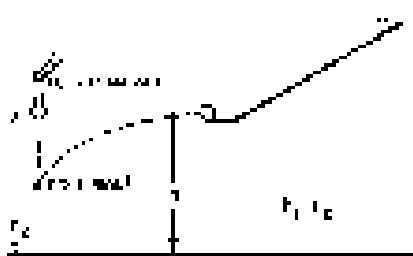


Fig. 7-3. Apparatus for showing the independence of distance and time interval.

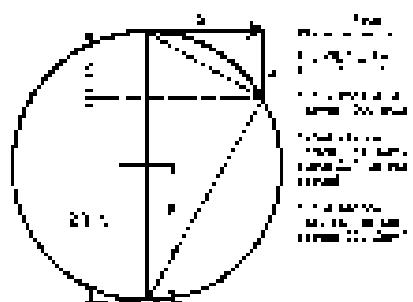


Fig. 7-4. Acceleration toward the center of a spherical earth. From Newton's geometry, $a/r = (G M_{\text{earth}})^{1/2}/r^3$, where M is the mass of the earth, 4000 miles, r is the distance "travelled horizontally" in one second, and G is the constant of gravitation, 6.67 times 10⁻¹¹.

explanation: the motion balls in the same path ought always fall straight down over the world, instead of falling out sideways. Let us take an example in the surface of the earth. An object, released near the earth's surface, will also fall 16 feet; even though it is moving horizontally, it still falls the same 16 feet, in the same time. Figure 7-3 shows an apparatus which demonstrates this. On the horizontal, t_0 is the time it takes for an object to be driven forward a little distance away. At the same instant is a ball which is going to fall vertically, and there is a curve, just switch enough so that it has the same time of fall as the first ball, comes to track. The second ball is released. Then they come to the same depth at the same time as measured by the fact that they fall in the same time. An object like a bullet shot horizontally, in a long enough way in one second—perhaps 1000 feet—he will still fall as far if it is aimed horizontally. What happens if we shoot a bullet lower and lower? Of course, longer than the earth's surface is curved. If we shoot it the same height, then when it falls 16 feet, it may be at just the same height above the ground as it was before. How can that be? It still falls, but the earth curves away, so it falls "backward" the earth. The question is, how far does it fall in one second? And the answer is 16 feet below the horizon! (See Fig. 7-4) we see the earth with us 1000 miles radius, and the angular, straight-line path that the bullet would take if there were no forces. Now, if we use some of Kepler's results in geometry, which says that the tangential to the mean proportionals between the two parts of the diameter cut by an angle of 1 radian, are exactly the horizontal distance travelled is the mean proportional between the 16 feet it has and the 1000-mile diameter of the earth. The square root of $(16/32800) \times 1000$ comes out very close to 5 miles. Thus we see that if the bullet runs out 5 miles a second, it then will continue to fall toward the earth at the rate of 16 feet each second, but will never get any closer, because the earth keeps curving away from it. This is what Galileo Galilei maintained in his "Dialogue Concerning the Two Chief World Systems," namely that the earth moves per second. (He mighta little longer because he was a little lighter.)

Another consequence of a new law is useful only if we can take more than we put in. Now, Newton used the second and third of Kepler's laws to derive his law of gravitation. What did he prefer? First, the analysis of the moon's motion was a predictive scheme. It connected the falling of objects on the earth's surface with that of the moon. Second, the question is, is the orbit an ellipse? We shall see in a later chapter how it is possible to rule this out theoretically, and in this one will prove that it should be an ellipse. So the first is needed by Kepler's first law. That Newton made his first guess full prediction.

Let us go to gravitation again. Many phenomena had previously undergone. For example, the pull of the moon on the earth causes the ocean tides. In mysterious the moon pulls the water up under it and makes the tides—people had thought of this before, but they were not as clear as Newton, and so they thought there ought to be only one tide during the day. The reasoning was that the moon pulls the water up uniformly making a high tide and a low tide, and since the earth spins under such a tide, the tide goes up and down in 12 hours. Astronomers thought that the high tide should be on the side out of the earth because as they argued, the moon pulls the earth away from the center. Most of these theories are wrong. It actually works like this—the pull of the moon for the earth and for the water is "backward" at the center. So, now, in which a place in the moon is pulled there *less* the average and the water which is farther away from it is pulled *more* than the average. Furthermore, the water can flow while the moon rotates, causing the tide picture to be something of this form:

What do we need to "balance"? What balances? Is the moon plus the whole earth toward it, why don't the earth fall right "up" to the moon? Because the earth does the same trick as the moon, it gives the earth around a point which is inside the earth but not at its center. The moon does the just as much the

* The sand is not present in this diagram.

earth, the earth and the moon both go around a central point, each following toward this common center, as shown in Fig. 7-5. This action around the common center is what balances the "pull" of each. So the earth is not going in a straight line toward the center, since it is moving. The word on the outside is "unbalanced" because the moon's attraction there is weaker than it is at the center of the earth, where it just balances the "centerugal force." The result of this unbalance is that the earth goes away from the center of the earth. On the near side, the attraction from the moon is stronger, and the unbalance is in the opposite direction, in space, the earth moves from the center of the earth. The net result is that we get an tidal bulges.

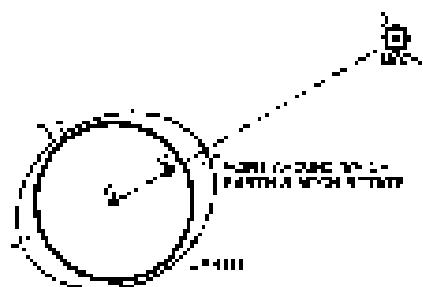


Fig. 7-5. The earth moves away with tides.

7-6 Erroneous gravitation

Who, then, can understand when we discuss gravity? Everyone knows the earth is round. Why is "heavier" toward? That is easy; it is due to gravitation. The earth can be understood to be round merely because everything about everything else and so it has attracted itself together so far as it can. If we go farther, the earth is no longer a sphere, when you are running, and this brings in additional effects which tend to oppose gravity near the equator. It is noted that the earth doesn't lie elliptical, and we can get the right shape for the ellipse. The error comes because now the sun, the moon, and the earth should be (nearly) spheres, just from the law of gravitation.

What does one do with the law of gravitation? If we look at the moons of Jupiter we can understand everything about the way they move around that planet. Incidentally, there was once a rather difficult problem with the moons of Jupiter that is worth remarking on. These four moons were studied very carefully by Galileo, who noted that the moons sometimes seemed to be ahead of schedule, and some times behind. (He calculated their orbits by waiting a very long time and finding out how long it took for the moons to go around.) Now they were ahead when Jupiter was particularly close to the earth and they were behind when Jupiter was further from the earth. This would have been a very difficult thing to explain according to the law of gravitation—it would have been, in fact, the death of the standard theory if there were no other explanation. If a law does not work even in one place where it ought to, it is just wrong. But the reason for the discrepancy was very simple and beautiful: it took a little while to see the moons of Jupiter because of the time it takes light to travel from Jupiter to the earth. When Jupiter is closer to the earth the time is 4½ hours, and when it is far from the earth, the time is more. This is why moons appear to us, on the average, a little earlier or a little behind, depending on whether they are closer to or farther from the earth. This phenomenon shows that light does not travel instantaneously, and destroyed the first estimate of the speed of light. This was done in 1656.

If all of the planets had had pull on each other, the force which controls, let us say, Jupiter in going around the sun is not just the force from the sun, there is also a pull from, say, Saturn. This force is not really strong, since the pull is much more massive than Sun—uh, but there is some pull, so the orbit of Jupiter should not be a perfect ellipse, and it is not, it is slightly off and "wobbly," around the exact elliptical orbit. Such a motion is a little more complicated. Attempts were made to analyze the motion of Jupiter, Saturn, and Uranus on the basis of the law of gravitation. The effects of each of these bodies on each other were calculated to see whether or not the right directions and integrations in these motions could be completely in accordance. It is one law. Go and behold, for Jupiter and Saturn, it was well, but Uranus was "wobbly." It behaved in a very peculiar manner. It was not traveling in an exact ellipse, but that was understandable because of the actions of Jupiter and Saturn. But, even if Uranus were made for these corrections, Uranus still was not going right on the laws of gravitation were it change, of being overruled, a possibility that could not be ruled out. Two men, Adama and Levenson, in England, at Paris, independently,

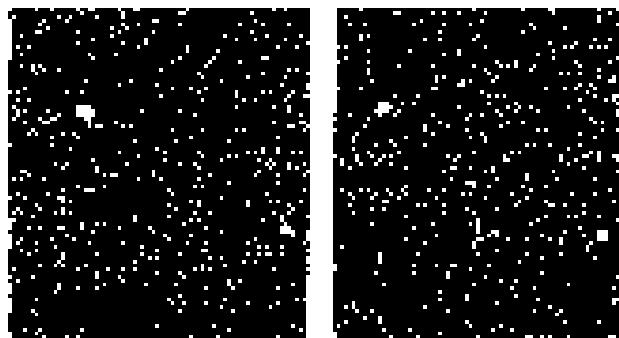


Fig. 7-6. A double star system.

entered at another possibility: perhaps there is another planet dark and invisible, which we have not seen. This planet, N , could put on charms. They calculate where such a planet would have to be in order to cause the observed perturbations. They send messages to the respective observatories, saying, "Gentlemen, point your telescopes in such and such a place, and you will see a new planet." It does depend on which whom you are working with, whether they say "they are sure it is not there" or not. They did pay attention to Newton; they looked, and there is no planet N now! The other observatory then also looked very quickly in the next few days and saw it not.

This discovery shows that Newton's laws are absolutely right in the solar system; not do they hold beyond the relatively small distances of the nearest stars! Two test best in the question, do stars affect each other as well as planets? We have definite evidence that they do in the double stars. Figure 7-6 shows a double star—two stars very close together (there is also a third star in the picture so that we will know that the photograph was not tampered). The stars are shown as they appeared several years later. We see that, relative to the "fixed" star, the two of the pair have rotated, i.e., the two stars are going around each other. Do they rotate according to Newton's law? Circular measurements of the true positions of one such double star system are shown in Fig. 7-7. There we see a beautiful ellipse, the minimum being in 1862 and going all the way around to 1904 (or now it must have gone around once more). Everything coincides with Newton's law, except that the star B has A near on the focus. Why should that be? Because the plane of the ellipse is not to be "perpendicular," say. We are not looking at right angles to the orbit plane, and when an ellipse is viewed at a tilt, it appears an ellipse but no longer a circle at the same place. Thus we can detect double stars, moving about each other, according to the requirements of the gravitational law.

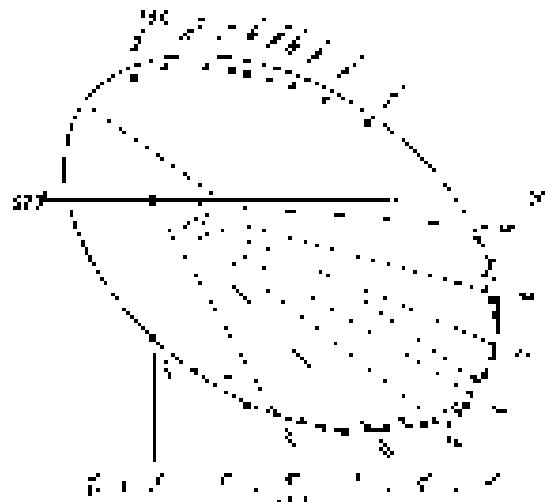


Fig. 7-7. Orbit of Star B with respect to Star A.

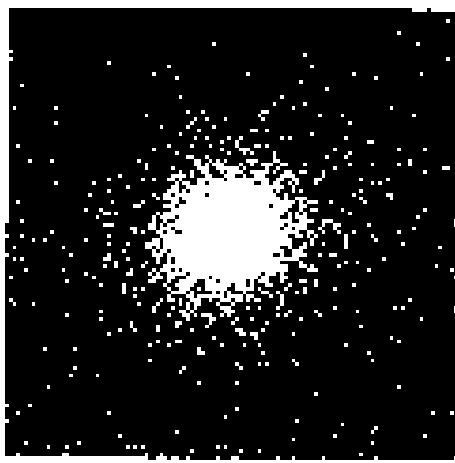


Fig. 7-8. A cluster of galaxies.

that the law of gravitation is true at even larger distances is indicated by Fig. 7-8. If one farther see gravitation acting here, he has no need. This figure shows one of the most beautiful things in the sky—a galactic star cluster. All of the galaxies cluster together, though they are packed so close toward the center. It is due to the pull of gravity of all the units. As a rule, the distances between even the nearest stars are very great and they very rarely collide. There are about 10^{11} stars in the cluster, and as we move outward there are fewer and fewer. It is obvious that due to the attraction among these stars it is clear that gravitation exists at these enormous dimensions, perhaps 100,000 times the size of the solar system. Let us now go further, and look at the galaxy photographs in Fig. 7-9. The shape of this galaxy indicates an obvious tendency for its matter to gather in a center. Of course we cannot prove that the law here is just as it is at square, only that there is at least some law in the enormous dimension, that holds the whole thing together. One may say, "Well, this is all very clever but who is it not just a bull?" Because it is obvious that the gravitation which a smaller gives up is transmitted, it must interact mostly in a plane. Accidentally, if you are looking for a good problem, the exact details of how the matter and forces here that determines the shape of these galaxies has not been worked out. It is, however, clear that the shape of the galaxy is due to gravitation; even though the complexities of the structure have not yet allowed

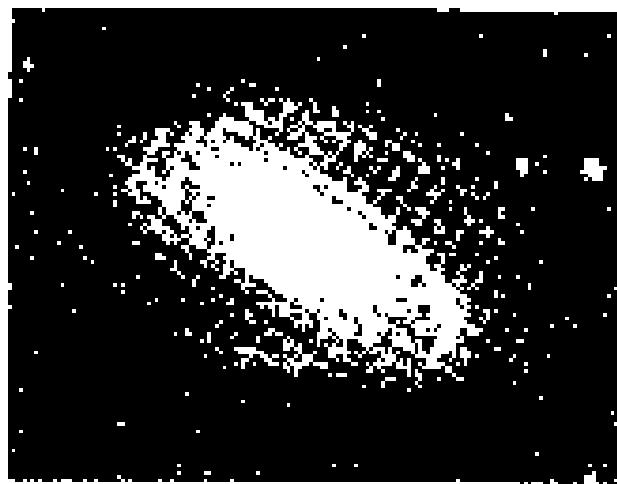


Fig. 7-9. A galaxy.

us to analyze it completely. In a galaxy we have a scale of perhaps 50,000 to 100,000 light years. The earth's distance from the sun is 8½ light minutes, so you can see how huge these dimensions are.

Gravity appears to act at these huge dimensions, as indicated by Fig. 7-10, which shows many "little" things clustered together. This is a cluster of galaxies, just like a star cluster. The galaxies attract each other so much that they form an agglomeration into clusters. Perhaps gravitation exists even over distances of over a million of light years, so far as we now know, gravity seems to go on forever regardless of the size of the system.

Not only can we understand the mechanics, but from the law of gravitation we can now get some idea about the origin of the stars. We have a big cloud of dust and gas, as indicated in Fig. 7-11, the gravitational attractions of the pieces of dust pull one another together until little lumps begin to form in the Figure (Fig. 7-12). These "lump" blocks which may be the beginning of the accumulation of dust and gases when due to their gravitation, begin, in turn, stars. Whether we have ever seen a star born or not is still open while Fig. 7-13 shows us the place of origin which suggests that we have. At the left is a picture of a region of gas with some stars in it taken in 1947, and at the right is another picture, taken only 7 years later, which shows two new bright spots. Has gravitation drawn the gravity center of each of them and reflected it into a lump enough that the stellar nucleus separation occurs in the manner and turns it into a star? Perhaps, and perhaps not. It is conceivable that in only seven years we should be so lucky as to see a star change its "face" visible to us; it is much less probable that we should see just

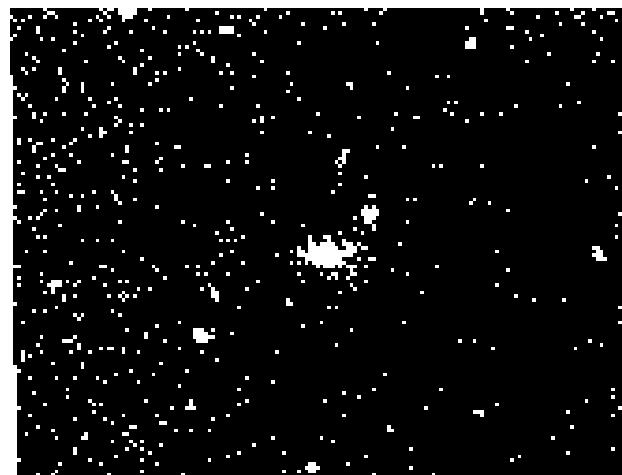


Fig. 7-10. A cluster of galaxies.

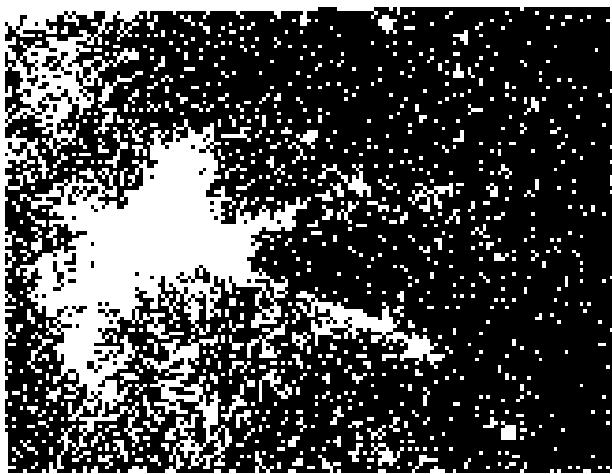


Fig. 7-11. An interstellar dust cloud.

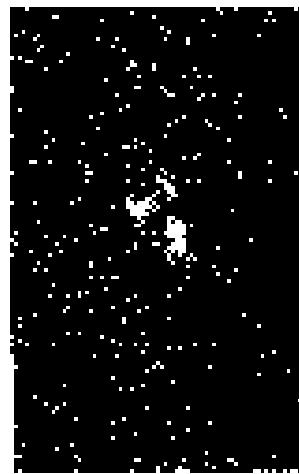
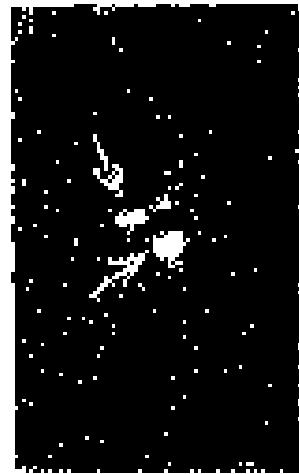


Fig. 7-12. The formation of new stars?



7-6 Cavendish's experiment

Gravitation, therefore, extends over enormous distances. But if there is a force between tiny particles, we might be led to believe it must be a force between our own objects. Instead of trying to watch the stars go around each other, why don't we make a pair of lead balls and make them orbit around the center of the ball of lead? The difficulty of this experiment when done in such a simple manner is the very weakness or deficiency of the force. It must be some sort of attractive force, which means something has to happen to keep the air molecules from flying off the Earth. And another: how the force can be measured. It was first measured by Cavendish, who, an apparatus which is schematically indicated in Fig. 7-13. The idea demonstrated the effect between two large, lead spheres of low density and another pair of lead spheres which were supported by a very fine fiber, called a torsion fiber. By measuring how much the fiber gets twisted, one can measure the strength of the force, so if the force is inversely proportional to the square of the distance, one can determine how strong it is. Thus, one may accurately determine the constant G in the formula

$$\delta \propto G \frac{m_1 m_2}{r^2}$$

All the masses and distances are known. You say, "We know G already for the earth." Yes, but we didn't know the mass of the earth. By knowing G , from this measurement and by knowing how strongly the earth attracts, we can indirectly know how great is the mass of the earth! This experiment has been called "weighing the earth." Cavendish claimed he was weighing the earth, but when he was measuring G , the coefficient of the gravity law, that is the only way in which the mass of the earth can be determined. G turns out to be

$$6.673 \times 10^{-11} \text{ newton} \cdot \text{m}^2/\text{kg}^2.$$

It is hard to exaggerate the importance of the effect on the history of science produced by this exactness of the theory of gravitation. Compare the situation, the scientific fulcrum, the incomplete knowledge, and the power in the earlier days when there were endless debates and disputes, with the courage and simplicity of this law. The fact that all the known and unknown and unknown parts of the sky were able to prevent them, and to make this man work undeterred if ever disaster hit the planets it could prove. This is the reason for the success of the sciences in following years, for it gave hope that the hidden phenomena of the world might also have such beautifully simple laws.

7-7 What is gravity?

But is this such a simple law? What about the machinery of it? All we have done is to deduce from the earth moves around the sun, but we have not said what makes it go. Newton made my hypothesis about this; he was so afraid to do that, that he did without getting into the controversy at all. No one has since given any machinery. There is no characteristic of the physical laws that they have this abstract character. The law of conservation of energy is, however, concerning gravitation, can have to be calculated and added together, two no mention of the machinery, and likewise the law of mechanics can not, that the mathematical laws for which no machinery is available. Why can we not manufacture to describe nature without a mechanism behind it? No one knows. We have to keep going because we find out more that way.

Many mechanisms for gravitation have been suggested. I am interested to consider one at least, which many people have thought of from time to time. A tree, for example, is quite subject to changes when he "falls over" it, but he does this "in his dream". It was first suggested about 1750. Suppose there were tiny vehicles moving in space at very high speed in a circular path and being only slightly attracted to gravity through mass. When they are disturbed, they give an impulse to the earth. However, given there are so many gravity and very no center, the

-----*

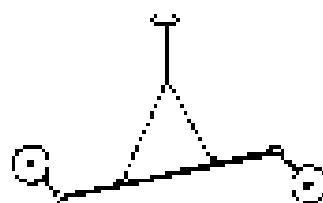


Fig. 7-13. A schematic diagram of the apparatus used by Cavendish to verify the law of universal gravitation by actually measuring the gravitational constant G .

imposes no balance. But when the sun is nearby, the particles coming toward the earth through the sun's rays are partially absorbed, so fewer of them are coming from the sun than are coming from the other side. Therefore, the earth feels a net impulse toward the sun, and it does not take much longer for the sun to impinge at 1% of the distance—because of the reduction of the resistance of the solid angle that the sun subtends as we vary the distance. What is wrong with that machinery? It involves some new consequences which you can prove. This particular view has the following merit over the earth, in circling around the sun, would impinge on more particles which are coming from the front side than from the hind side when you are in the sun. So even in your face is stronger than that on the back of your head. Therefore there would be more impulse given to the earth from the front, and the earth would feel a resistance to move and would be slowing up in its orbit. You can calculate how long it would take for the earth to slow as a result of this resistance, and it would not take long enough for the earth to still be in a orbit, so this mechanism does not work. You can very easily have an inverted "U" explaining gravity without also predicting some other phenomenon that does not exist.

But we still cannot yet make a relation of gravitation to other forces. This is an angle of view that in terms of other forces at the present time, it is not an angle of electricity or anything like that, so we have no explanation. However, gravitation and other forces are very similar and it is interesting to note analogies. For example, the force of attraction between two charged objects is exactly like the law of gravitation; the force of electricity is also similar with a minus sign. Times the product of the charges, and inversely to the square of the distance. It is in the opposite direction of a repel. But it is still not very remarkable that the two laws involve the same function of distance! Perhaps gravitation and electricity are just some closely related law, we think. Many attempts have been made to unify them: the so-called unified field theory is only a very elegant attempt to combine gravitation and magnetism. But in unifying gravitation and electricity, the most curious thing is to combine strength of the forces. Any theory that combines even both measurable deduces how strong the gravity is.

If we take, in some natural or in the caption of two electrons (i.e., it's a "natural" charge due to electricity), and the attraction of two electrons due to their masses, we can measure the rate of electrical repulsion to the gravitational attraction. The ratio is independent of distance and is a fundamental constant of nature. The ratio is shown in Fig. 7-11. The gravitational attraction relative to the electrical repulsion between two electrons is 1 divided by 4.17×10^{39} . The question is, who chose such a large number in some form? It is not accidental. Like this, is it the volume of the entire observable universe? We have considered two equal charges of the same mass, one electron. This "united number" is a natural constant, as it involves something deep in nature. When you say "united number" you mean something like "the ratio of the universe." When you say "united constant," you mean it is of the ratio of the ratio of the universe. It is very difficult to find an equation for which such a farcical number is a natural root. Other possibilities have been thought of: one is to relate it to the age of the universe. Certainly we have to find another "united number": But the age is 10¹⁰ years. The ratio is 10¹⁰ years? No, because years are not "natural"; they were devised by man. As an example of something natural, let us consider the time it takes light to go across a proton, 10^{-22} second. Two components have to do with the age of the universe? 2×10^{10} years, the sequence 2×10^{10} ? It is about the same number of zeros going off it, so it has been suggested that the gravitational constant is related to the age of the universe. If this is true, the gravitational constant would change over time, because as the universe got older the ratio of the age of the universe to the time which it takes for light to go across a proton would be gradually increasing. In this case, the gravitational constant is changing with time! Of course the changes would be so small that it is quite difficult to be sure.

One test which we can think of is to determine when we shall have the $\sqrt{2}$ of the change: during the past 10¹⁰ years, which is approximately the age from the earliest life on the earth to now, a thousandth of the age of the universe. In this case, the gravity constant would have increased by about 10 percent. D 7-3

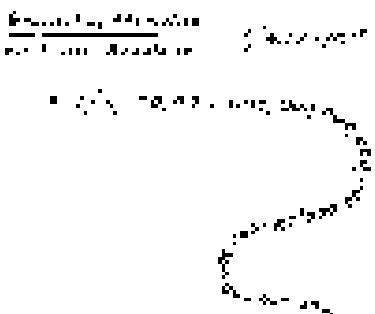


Fig. 7-4. The relative strength of electrical and gravitational interactions between two electrons.

but if we do, if we consider the structure of the sun—the balance between the weight of its material and the tension which holds it together—is maintained inside it—we can deduce that if the gravity were 10 percent stronger, the sun would be much more massive, 10 percent brighter, by the fifth power of the gravity constant! If we calculate what happens to the orbit of the earth when the gravity is changing, we find that the earth was then closer in. Anywhere the sun would be stronger, the gravitational acceleration, and all of the states, would not have been in the sun, but separate the air or life would not have to live in the sun. So we know now believe that the gravity constant is changing with the age of the universe. But such arguments as the one we have just given are not very convincing, and the subject is not completely closed.

It is a fact that the force of gravitation is proportional to the mass. PROPORTIONAL means it holds exactly a measure of inertia. If one hand it is to hold something which is going around in a circle. Therefore two objects, one heavy and one light, spin around the same axis at the same speed because of gravity, will stay together because they form a circle requires a force which is enough for a bigger mass. And as the gravity is stronger for a given mass in just the right proportion so that the two objects will go around together. If one object were made the other would always move; it is a perfect balance. Therefore Galileo in Time would find things "weightless" inside a space ship, as they happened to let go of a piece of it. For example, it would go around the earth, instead the earth goes to the whole space ship, and you would appear to move "weightless" because the ship spaces. It is very interesting that this force is exactly proportional to the mass with great precision, because if it were not exactly proportional there would be some effect by which inertia and weight would differ. This absence of such an effect has been checked with great accuracy by an experimental done first by Eötvös in 1909 and more recently by Dicke. For a kilometer-sized, the masses and weights are exactly proportional within 1 part in 100,000,000 or less. This is a famous cubic experiment.

7.3 Gravity and relativity

Another topic regarding gravitation is Einstein's modification of Newton's law of gravitation. I skip all the excitement it created, Newton's law of gravitation is very good! It was modified by Einstein to take into account the theory of relativity. According to Einstein, the gravitational effect is instantaneous. That is, if we were to move a mass, we would feel a new force because of the new presence of the mass by such means we could send signals at infinite speed. Einstein advanced arguments which suggest that we cannot send signals faster than the speed of light so the law of gravitation must be wrong. By correcting it to make the change is a constant, we have the so-called Einstein's law of gravitation. One feature of this new law which is quite easy to understand is this. In the Einstein's relativity theory, anything which has energy has mass. Mass in the sense that it is affected gravitationally. Two lights which has an energy, has a "mass." When a light has a mass it comes out the sun must be attracted to it by the sun. Thus the light there is going straight but is deflected. Or along the surface of the sun, for example, the comet which approaches the sun could appear deflected. But where they would be if the sun were not there, and this has been observed.

Finally, let us compare gravitation with other theories. In recent years we have discovered that there are more than particles and that there are several kinds of interactions, one is nuclear forces and none of the sources in classical theory has yet been found to exhibit gravitation. The quantum-mechanical argument is still not yet been able to be used to gravitational. While the scale is so small that we can see the quantum effects, the gravitational effects are so weak that the usual form of general theory of gravitation has not yet developed. On the other hand for consistency in our physics, there is it would be important to see whether Newton's law modified to Einstein's law can be further modified to be consistent with the uncertainty principle. This has not been done yet but not yet completed.

Motion

B-1 Description of motion

In order to feel the laws governing the various changes that take place in bodies as time goes by, we must be able to measure the changes and have some way to record them. The simplest change to measure is a change in the position with time, which we call motion. Let us consider some solid object with a permanent mark, which we shall call a point, that we can observe. We shall consider motion of the kind marked, which is just the motion along an immobile or the motion of a falling ball, and shall try to describe the fact that it moves and how it moves.

The velocity is very simple indeed, but many subtleties enter into the description of change. Some changes are more difficult to describe than the motion of a point in a straight line, for example the speed of drift of a cloud that is drifting very slowly, the rapidly changing temperature, or the change of a woman's mind. We do not know a simple way to analyze a change of mind, but since the cloud can be represented or described by many molecules, perhaps we can do the motion of the cloud in principle by describing the motion of individual molecules. Likewise, perhaps even the changes in the sun may have a parallel in changes of the atoms inside the sun, but we have no such knowledge yet.

At any rate, that's why we begin with the motion of "points"; perhaps we should think of them as atoms, but it is probably better to be more rough in the beginning and simply to think of some kind of small objects—small, that is, compared with the distance moved. For instance, in describing the motion of a car that is going a hundred miles, we do not have to distinguish between the front and the rear of the car. To be sure, there are slight differences, but for rough purposes we say "the car," and likewise it does not matter that our points are not absolute points; for our present purposes it is not necessary to be absolutely precise. At any rate, we take a first look at this subject we are going to begin with the three dimensions of the world. We shall not concentrate on moving in one direction, as in a straight line. We shall return to three dimensions later. We are now to describe motion in one dimension. Now, you may say "There is a kind of trick," and indeed it is. How can we describe such a one-dimensional motion—intensity, of course? Nothing could be simpler. Among many possible ways, one would be the following: To determine the position of the car at different times, we measure its distance from the starting point and record all the observations. In Table B-1, I represent the distance of the car, in feet, from the starting point, and I suppose the road is smooth. The first 200 ft can also represent two dimensions and zero time—the car has not started yet. After one minute it has started and has gone 1200 ft. Then in two minutes, 1200 ft far from the starting point, it picked up more distance in the second minute—it has accelerated; cut something happened between 3 & 4, & 4 and even more so at 5—it stopped at a light perhaps? Then it speeds up again and goes 11,000 ft; it travel'd of 6 minutes 15,000 ft; in the end of 7 minutes, and 20,000 feet in 6 minutes; at 9 minutes it was advanced to only 24,000 feet because in the last minute it was stopped by a stop.

That is one way to describe the motion. And the way is by means of a graph. If we plot the time horizontally and the distance vertically, we obtain a curve something like that shown in Fig. B-1. As the time increases, the distance increases, at first very slowly and then more rapidly, and very slowly again for a little while at 4 minutes; then it increases again for a few minutes and finally, at 9 minutes, appears to have stopped increasing. These observations can be made from the

B-1 Description of motion

B-2 Speed

B-3 Speed as acceleration

B-4 Distance as an integral

B-5 Acceleration

Table B-1

t (min)	d (ft)
0	0
1	1200
2	4800
3	9600
4	9900
5	9600
6	13000
7	19000
8	23500
9	24000

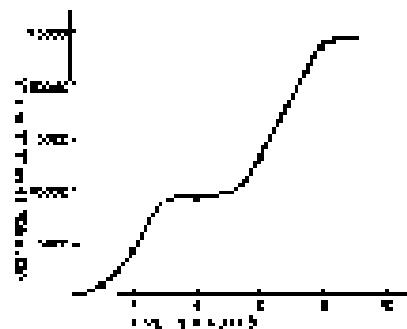


Fig. B-1. Graph of distance versus time for the car.

Table 8-3

t (sec)	s (ft)
0	0
1	16
2	64
3	144
4	256
5	400
6	576

graph, without a graph. Obviously, for a complete description one would have to know where the car is at any particular instant, but we suppose that the graph means something, that it is a function of all the intermediate times.

The motion of a car is complicated, but another example may help illustrating that s does not in general increase uniformly. Let us consider a falling body. At zero seconds the car starts out at zero feet, and at the end of 1 second it has fallen 16 feet. At the end of 2 seconds it has fallen 64 feet, at the end of 3 seconds 144 feet, and so on. If the individual numbers are plotted, every point lies on a parabola, as we showed in Fig. 4-2. The corresponding curve can be written as

$$s = 16t^2. \quad (8.1)$$

This formula enables us to calculate the distance at any time. You might say that ought to be a formula for the first graph too. Actually, one may write such a formula also, namely,

$$s = f(t). \quad (8.2)$$

meaning that s is some quantity depending on t , or, in mathematical language, a function of t . But we do not know what the function f is, there is no way we can write it in full, inside basic form.

We have now seen two examples of motion, and quite satisfied with very simple ideas, in substance. However, there are subtleties, several of them. In the first place who does mean he can call yourself? It turns out that these deep philosophical questions have to be analyzed very carefully in physics, and this is not so easy to do. The theory of relativity shows that our ideas of space and time are not as simple as one might think at first sight. However, for our present purposes, the accuracy that we need at best, we need not be very careful about defining things precisely. Perhaps you are: "That's a terrible thing. I learned that in school we have to define everything logically." We cannot define everything logically! If we attempt to do just that, analysis of thought will come to a standstill, when we express each concept again in the whole. "You don't know what you're talking about!" The person you say, "What do you mean by time?" What do you mean by nothing? What do you mean by your?" and so on. In order to handle all these subtle questions, we just have to agree that we are talking about roughly the same thing. That's how it is in the real world we live in for the present... but remember that there are some subtleties that have to be discussed, we'll discuss them later.

Again, quite logically involved, and strongly mentioned, is that it should be possible to integrate that the interval which we are observing is always between somewhere. (Of course, when we are looking up at the stars, but maybe when we look away, it isn't there.) It turns out that in the motion of objects, that idea also is to be—well, not just a matter of our own and suitable choices. That subtlety we shall have to get into in quantum mechanics. But we are first going to do it with the problems we select later during the considerations, and then we shall be in a better position to judge our estimate in the light of the more exact knowledge of the subject. You see, therefore, like a simple pair of shoes that, time and space we know what these concepts are in a rough way, and those were just driven in by kind of a general means.

8-3 Speed

Even though we know roughly what speed means, there are still some rather deep subtleties involved that the ancient Greeks were not able to adequately describe the sometimes varying velocity. This will be discussed when we try to compute exactly what is meant by "speed." The Greeks got very confused about this, and a new branch of mathematics had to be developed between the geometry and algebra of the Greeks, 400+ years. Babyish! As an illustration of the difficulty, try to solve this problem by sheer algebra! A solution is being indicated on p. 12.

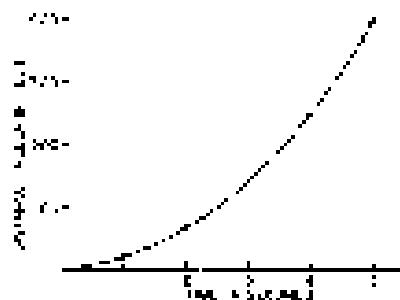


Fig. 8-2. Graph of distance versus time for free fall.

that the volume of the balloon is increasing at the rate of 100 cm³ per second; at what speed is the radius increasing when the volume is 1000 cm³? The Greeks were somewhat confused by such a dilemma, being helped, of course, by some very confusing Greeks. To show just how wrong was their concept of motion, let us consider speed in the time, Zeno proposed a large number of paradoxes, of which we shall mention one. This is to his credit, this is the earliest evidence of difficulty in thinking about motion. "Listen," he says, "to the following argument. Achilles runs 10 times as fast as a tortoise, nevertheless, he can never catch the tortoise. For suppose that they run in a race where the tortoise is 100 m ahead of Achilles; then when Achilles has run the 100 meters to the place where the tortoise was, the tortoise has proceeded 10 meters. During that same time Achilles has run another 10 meters to catch up with the tortoise, but by arriving at the end of each run, Achilles finds the tortoise is still 10 m ahead of him, so again Achilles runs. In this fashion Achilles' continuous motion, and motion, is self-defeating. Therefore, at any moment the tortoise is always ahead of Achilles and Achilles can never catch up with the tortoise." What is wrong with this? It is that all the amount of time can be divided into an infinite number of pieces, just as a long line can be divided into an infinite number of pieces by dividing repeatedly by two. And so, although there are an infinite number of steps (in the argument) in the path... at which Achilles reaches the tortoise, it doesn't mean that this is a finite segment of time. We can see from this example that there are hidden some subtleties in the way about speed.

In order to get to the subtleties in a clearer fashion, we remind you of a joke which will surely stir up some humor. At the point where the lady in the car is caught by a cop, she goes up to him and says, "Sorry, sir, I was going 60 miles an hour!" She says, "It's impossible, sir, I was traveling for only seven minutes. It's ridiculous! How can I go 60 miles an hour? What I wasn't going, sir?" Now, would you believe her? You say, "Oh, right? Of course, you were really the cop, then no stories are intended; it's very simple, go on, say, "I'll stick to the wages!" But let us suppose that we do not have that escape and we take a more honest approach and attack on the problem, and try to explain... Not only what we know by the time that she was going 60 miles an hour. Just what do we know? We say, "What we know, however, is this: if you kept on going the same way as you were going now in the next hour you would go 60 miles." You could say, "Well, my foot was off the accelerator and the car was slowing down, so if I kept on going like this, why it wouldn't go 60 miles." Or, instead, the ball is bounces up and we want to know, especially, at what height it bounces? The ball goes on going the way it is going. What does that mean? Stop or accelerate, just keep it. No, just carrying with it's own velocity. But that is what we are trying to define. For if the ball keeps on going the way it is going, it will just stop on going the way it is going. Thus we need to define the velocity better. What do we do? Just the same. That's all we do. Now, "If I keep on going like this I'm going in one more hour, I would run faster than I did at the end of the street." It is not so easy to see what we mean.

Many physicians think that measurement is the only definition of any thing. Obviously, then, we should use one instrument that measures the speed—the speedometer—and say, "Look, lady, your speedometer reads 60." So she says, "My speedometer is broken and didn't read at all." Does this mean the car is standing still? We believe that there is something to measure before we talk of a speedometer. Only, does our, for example, "The speedometer isn't working right," or "The speedometer is broken?" That would be a meaningful sentence if the velocity and so resulting independence of the speedometer. As we know it, we can make claim that is independent of the speedometer, since the speedometer is not the only to measure this idea. So, in case if we can get a better definition of the idea, we say, "Yes, of course, before you went an hour, you would do that well, but, if you are not measured, you would go 60 feet; but, you were going 60 for per second, and if you kept on going, the next second, you would be 60 feet, and the will down. There is far less than that." She says, "This is... I don't know how fast I am going 60 feet per second." There is only a few

against going 80 miles an hour." "But," we reply, "it's the same thing." If it is the same thing, it should not be necessary to go into the circumlocution about 96 feet per second. In fact, the falling ball would not keep going the same way even one second because it would be changing speed, and we shall have to define speed somehow.

Now we seem to be getting on the right track: it goes something like this: If the lady kept on going for another 1/1000 of an hour, she would go 1/1000 of 80 miles. To other words, if she did not have to keep on going for the whole hour, the speed is then for a moment she is going at that speed. Now what that means is that if she went just a little bit more in time, the extra distance she goes would be the same as the distance that you, at a constant speed of 80 miles an hour. Perhaps the idea of the 80 foot per second is right; we say how far she went in the last second, divide by 80 feet, and if it comes out 1/1000 of an hour was 80 miles an hour. In other words, we can find the speed in this way. We don't know how far we go in a very short time. We divide this distance by the time, and that gives the speed. But the time should be made as short as possible, the better the better, because some change would take place during that time. If we take the time of ϵ for my body as an hour, the idea is ridiculous. If we take ϵ as a second, the result is pretty good for a car, because the car's constant change is great. But not for a falling body; so in order to get the speed more and more correctly, we should take a smaller and smaller time interval. What we should do is take a millionth of a second, and divide the distance by a millionth of a second. The result gives the velocity per second, which is what we mean by the velocity, or we can define it that way. That is a suggested answer for the lady, or rather, that is the definition that we are going to use.

The foregoing definition involves a new idea, an idea that was not available to the Greeks in a general form. That idea was to take a. infinitesimal distance and the corresponding infinitesimal time from the ratio, and when it comes to that ratio as the time and the distance become smaller and smaller and smaller. In other words, take a "unit" of the distance traveled divided by the time required, as the times taken give smaller and smaller and infinitesimal distances. This idea was invented by Newton and by Leibnitz, independently, and is the beginning of a new branch of mathematics, called the differential calculus. Calculus was invented in order to describe motion, and its first application was to the problem of "finding what is meant by going 80 miles an hour."

Let us try to define velocity a little better. Suppose that in a short time, ϵ , the car or other body goes a short distance Δx ; then the velocity, v , is defined as

$$v = \Delta x/\epsilon$$

an approximation that however nearer and nearer to the v is taken smaller and smaller. If a mathematical expression is desired, we can say that the velocity equals the limit as ϵ is made larger and larger in the expression above:

$$v = \lim_{\epsilon \rightarrow 0} \frac{\Delta x}{\epsilon}. \quad (8.1)$$

We cannot do the same thing with the lady in the car, because the table is incomplete. We know only where she was at intervals of one minute; we can get a rough idea that she was going 8000 ft/min during the 7th minute, but we do not know, exactly the moment 7 min, or whether she had been speeding up and the speed was 9000 ft/min at the beginning of the 6th minute, and is now at 8 ft/min, or decelerating etc., because we do not have the exact details in between. So only if the table were completed with an infinite number of rows could we really calculate the velocity from such a table. On the other hand, when we have a complete mathematical formula as is the case of a falling body (Eq. 8.7), then it is possible to calculate the velocity, because we can calculate the position in any time whatsoever.

Let us take as an example the problem of determining the velocity of the falling ball at the particular instant t seconds. One way to do this is to see from

Table 4-2 was stated in the problem; it was $400 = 256 = 174$ ft, so it is going 144 ft/year; however, that is wrong, because the speed is changing. In the average it is 144 ft/year, using this is great, but the ball is speeding up and is really going faster than 144 ft/year. We want to find our velocity after four. The technique involved in this process is the following: We know where the ball was at 5 sec. At 5 sec we saw the distance that it has gone all together is $16(5)^2 = 400$ ft (see Eq. 4.3). At 5 sec it had already fallen 400 ft; in the last tenth of a second it fell $400 - 360 = 16.0$ ft. Since 16.0 ft in 0.1 sec is the same as 160 ft/sec, that is the speed more or less, but it is not exactly correct. Is that the speed at 5, or at 5.1, or the change between at 5.05 sec, or what is that the speed? Never mind — the problem was to find the speed at 5 seconds, and we do not know exactly what we have to do a better job. So, we take one-thousandth of a second more than 5 sec, or 5.001 sec, and calculate the total fall as

$$s = 16(5.001)^2 = 16(25.000001) = 400.00015 \text{ ft}$$

In the Sec. 6.2C we saw the ball fell 0.160006 ft, and if we divide this number by 0.001 sec we obtain the speed as 1600.6 ft/sec. That is closer, very close, but it is still not exact. It should now be evident what we must do to find the speed exactly. To perform the mathematics we state the problem: "Take more time to find the velocity at a specific time, in which in the original problem was 5 sec. Now the total distance, which we call s_2 , is 16 t^2 , or 400 ft, in $t > 5$ sec. In order to find the velocity, we ask 'At what time $t_0 + \Delta t$ falls s_2 , or $s_2 = s_1 + \Delta s$, where is the body?' The new position is $16(t_0 + \Delta t)^2 = 16t_0^2 + 32t_0\Delta t + \Delta s$. So it is falling along when it was before, but also before it was only 16 t_0 . This distance we shall call $s_3 = (\Delta t)$ little bit more, or $s_3 = x$ if you like extra x . Now if we subtract the distance at t_0 from the distance at $t_0 + \Delta t$ we get the extra distance $s_3 = 32t_0\Delta t + \Delta s = 16\Delta t^2$. Our first approximation to the velocity is

$$v = \frac{s}{t} = 32t_0 + \Delta s. \quad (4.4)$$

The true velocity is the value of the ratio, v/t , when Δt becomes infinitesimally small. In other words, after forming the ratio, we take the limit as Δt gets smaller and smaller, that is, approaches 0. The equation reduces to

$$\lim_{\Delta t \rightarrow 0} \frac{s}{\Delta t} = 32t_0.$$

In our problem, $t_0 = 5$ sec, so the solution is $s = 16(5)^2 = 400$ ft/sec. A few times later when we took $\Delta t = 0.1$ and 0.01 sec successively, the value we got for v was a little more than 160, but now we see that the actual velocity is probably 1600 ft/sec.

4-4 Speed as a Derivative

The procedure we have just carried out is performed whenever we calculate that for nonuniform (or special) motion have been assigned to our quantities s and v in this section, the cases above because s and v becomes s_1 . This is because "you can't hit of it" and don't be implying that s can be made smaller. The prefix Δ is not a multiplier, any more than can s means $s \cdot 1$; in Δs is simply added a time increment, and remains as of its special character. Δs has no analogous meaning for the distance s . Thus, Δs is not a factor. It cannot be cancelled in the ratio $\Delta s/\Delta t$ to give s/t , any more than the ratio s/t can be reduced to 1/2 by cancellation. In this situation, velocity is added to the unit of s/t as when Δt gets smaller or

$$v = \lim_{\Delta t \rightarrow 0} \frac{\Delta s}{\Delta t}. \quad (4.5)$$

This is really the same as the previous expression (Eq. 4.4), but it has the advantage of showing that something is changing, and a *slope* — that of what is changing.

Inertial "y," or a *local approximation*. We know another law, which says that the change of distance of a moving point is the velocity times the time interval $\Delta t = t - \tau$. This statement is true only if the value v is not changing during that time interval; and the condition is not very far from the truth as Δt goes to 0. Physicists like to state it as $\Delta s \approx v \Delta t$, because by Δs they mean an *infinitesimal* in which Δt is very small; while Δt is *meaningful*. In such case it is valid to a close approximation. If Δt is too long, the velocity might change during the interval, and the approximation would become less accurate. For a time Δt approximating $v \Delta t$ is *infinitesimally*. To this end here we can write $\Delta s \approx v \Delta t$.

$$v = \lim_{\Delta t \rightarrow 0} \frac{\Delta s}{\Delta t} = \frac{ds}{dt}.$$

The quantity ds/dt which we found here is called "the derivative of s with respect to t " (this language begins to keep track of what was computed), and the computational process of finding it is called finding a derivative. In the ordinary $f(x)$'s and $g(x)$'s which appear separately are called *arguments*. To familiarize you with the words, we say we found the derivative of the "function $f(x)$ ", or the derivative (with respect to x) of $f(x)$ is $f'(x)$. When we get used to the words, the f 's are more easily understood. The process, for us and the derivative of a more complicated function. We shall consider the function $s = At^2 + Bt + C$, which might describe the motion of a point. The letters A , B , and C represent constant numbers, as in the common general form of a quadratic equation. Starting from the formula for the motion, we wish to find the value v at any time. To find the velocity in the more elegant manner, we change s to $t + \Delta t$ and note that s has changed to $s + \Delta s$; then we find v at $t + \Delta t$. That is to say,

$$\begin{aligned} s + \Delta s &= A(t + \Delta t)^2 + B(t + \Delta t) + C \\ &= At^2 + Bt + C + 2At\Delta t + A(\Delta t)^2 + B\Delta t. \end{aligned}$$

But since

$$s = At^2 + Bt + C,$$

we find also

$$\Delta s = 2At\Delta t + B\Delta t + A(\Delta t)^2 \approx A(\Delta t)^2.$$

But we do not want Δs — we want s divided by Δt . We divide the preceding expression by Δt , getting

$$\frac{\Delta s}{\Delta t} = 2At^2 + B + B(\Delta t) + A(\Delta t)^2.$$

Table B-3. A Short Table of Derivatives

For the differentiable functions of t , s , v , and a arbitrary constants

Function	Derivative
$s = t^n$	$\frac{ds}{dt} = nt^{n-1}$
$s = ct$	$\frac{ds}{dt} = c$
$s = s_0 + v_0 t + \frac{1}{2}at^2 + \dots$	$\frac{ds}{dt} = \frac{dv}{dt} = v_0 + \frac{da}{dt}t = \dots$
$s = c$	$\frac{ds}{dt} = 0$
$s = e^{kt} + c$	$\frac{ds}{dt} = k \left(\frac{d}{dt}e^{kt} + \frac{d}{dt}c + \frac{d}{dt} \frac{dc}{dt} + \dots \right)$

As A goes toward 0 the limit of $\Delta x/\Delta t$ is dx/dt and is equal to

$$\frac{dy}{dt} = 3x^2 + 5.$$

This is the fundamental process of calculus, differentiating functions. The process is even more simple than it appears, however, because there are various ways to find dy/dx from which we can subtract terms or cube or any higher power of dx , since terms may be dropped when we take the limit as A goes to 0. When the limit is taken, A has a value greater than zero, so we know what x is going to. There are many rules for differentiating various types of functions. These will be summarized, or can be found in tables. A short list is found in Table 8-1.

8-1 Distances for Discrete

Now we have a discrete time-varying problem. Suppose the initial distance of distances we have a table of speeds at 5 discrete times, starting from zero. Just the falling ball, such speeds and times are shown in Table 8-4. A similar table could be made for the velocity of the car by measuring the speed each second every minute or half minute. If we know how fast the car is going at any time, can we determine how far the car is going? This problem is just the inverse of the one solved above; we are given the velocity and asked to find the displacement. How can we find the distance if we know the speed? If the speed of the car is constant, and the only goes very quickly but for a moment, then slows down, speeds up, and so on, how can we determine how far one has gone? That is easy. We use the same idea, and replace the Δt by Δx and v by a . Let us say, "In the first second her speed was v_1 and x_1 , and from the formula $x = v_1 t$ we can calculate x_2 for the car went the first second at that speed." Now at the next second he speed is v_2 , by the same but slightly different, we can calculate x_3 for the second the Δx is the Δx_2 by taking the new speed times the Δt . We generate Δx 's for each second, to the end of the car. We now have a number of little distances, and the total distance will be the sum of all these little pieces. That is, the distance will be the sum of the velocities times the times, or $\sum v_i \Delta t$, where the Greek letter Σ (Sigma) is used to denote addition. To be more precise, it is the sum of the velocity at a certain t . Let us say the car is interrupted at t_0 ,

$$x = \sum v_i \Delta t. \quad (8.6)$$

The rule for the times is that $t_{i+1} = t_i + \Delta t$. However, the distance we obtain by the method will not be exactly, because the velocity changes during the time interval Δt . If we take the times short enough, the sum is precise, or we take them smaller and smaller until we obtain the desired continuum. That is, it is

$$x = \lim_{\Delta t \rightarrow 0} \sum v_i \Delta t. \quad (8.7)$$

These terms allow have invented a symbol for this limit, analogous to the symbol for the derivative. The symbol \int is also required so that the time is as small as we can be, the velocity is then called v at the time t , and the addition is written as a sum with a Greek Σ (from sigma), which has become distance and is now unfortunately just called an integral sign. Thus we write

$$x = \int v(t) dt. \quad (8.8)$$

The process of adding all these terms together is called integration, and \int is the opposite process to differentiation. The derivative of the integral is v , or the operator d/dt makes the other (\int). One can get experience for integrals by taking the formulas for derivatives and turning them backwards, because they are integrating each other inversely. Thus you can work out the own value of integrals by differentiating all sorts of functions. For every function that is differentiable, we get an integral: "Guru's if we turn it around."

Table 8-1

Velocities of a Falling Ball

t (sec)	v (ft/sec)
0	0
1	32
2	64
3	96
4	128

Every function can be differentiable analytically, i.e., the process can be carried out algebraically, and leads to a definite function that it is not possible in a simple manner to write an analytical value for any integral of $\vec{F}(t)$. You can calculate it, for instance, by taking a derivative, and then multiplying again with a function $\vec{v}(t)$ and again with a other function until you have it nearly right. In general, given some particular function, it is not possible to find, analytically, what the integral is. One may always try to find a function which, when differentiated, gives some desired function, but one may not find it, and it may not exist, at the sense of being expressible in terms of known, but not already been given names.

8-5 Acceleration

The next step in developing the equations of motion is to introduce another term which goes beyond the concept of velocity in the of change of velocity, and we now ask: "How does the velocity change?" In previous chapters we have discussed cases in which force produce changes in velocity. You may have worked with your environment where you can think about from a car to fall miles an hour, or even seconds. From such a performance we can say how fast the speed changes, but only on the average. What we do, *i.e.*, discuss is the next level of complexity, which is how fast the velocity is changing. In other words, by how many feet per second does the velocity change in a second, maybe, how many feet per second does *s* change? We have already derived the formula for the velocity of a falling body as $v = \sqrt{2gt}$, which is charted in Table 8-4, and now we want to find out how much the velocity changes per second; this quantity is called the acceleration.

Acceleration is defined as the ratio $\frac{\Delta v}{\Delta t}$ of change of velocity. From the preceding discussion we *already* enough already to state the acceleration as the derivative $\frac{dv}{dt}$; in the same way that the velocity is the derivative of the position if we *now* differentiate the formula $v = \sqrt{2gt}$ we obtain, for a falling body,

$$a = \frac{dv}{dt} = \frac{d}{dt}(\sqrt{2gt}) \quad (8-9)$$

[In differentiating the term $\sqrt{2gt}$ we can utilize the result obtained in a previous problem, where we found that the derivative of \sqrt{x} is simply $\frac{1}{2\sqrt{x}}$ times unity. So by letting $B = 32$, we have at once that the derivative of $\sqrt{2gt}$ is $\frac{B}{2}\sqrt{2g}$.] This means that the velocity of a falling body is changing by 16 ft/s per second, per second changes. You may ask, from logic, that the velocity increases by 16 ft/s in each second. This is a very simple case, for most objects are usually not constant. The reason the mass is not *so* constant is, is that the force on the falling body is constant, and Newton's law says that the acceleration is proportional to the force.

As a further example, *i.e.*, to find the acceleration in the problem we have already solved for the velocity, bringing up:

$$s = 2t^2, \quad \vec{s} = \vec{v}, \quad \vec{v}$$

and differentiating $s = 2t^2$ with respect to time,

$$v = 2t^2 = \vec{v}$$

After acceleration is the derivative of the velocity with respect to the time, we need to differentiate the last expression above. Recall the rule that the derivative of the sum of terms on the right equals the sum of the derivatives of the individual terms. To differentiate the first of the terms, instead of going through the fundamental process again we note that we have already differentiated a quadratic term when we *f* learned "*to f*", and the effect was to double the initial coefficient and change t^2 to t^3 in order to get the corresponding derivative of the linear, and you can check this result yourself. The derivative of $3t^2$ will then be $6t$. Next we differentiate \vec{v} , a sine wave, which has been treated previously. The derivative of \vec{v} is zero, because this term contributes nothing to the acceleration. So final result, therefore, $a = \frac{dv}{dt} = 6t$.

For reference, we state two very useful formulas which can be obtained by integration. If a body starts from rest and moves with a constant acceleration a , its velocity v at time t is given by

$$v = at.$$

The distance it covers in this same time is

$$x = \frac{1}{2}at^2.$$

Various modifications of these are used in solving derivatives. Since velocity is *displacement* and acceleration is the time derivative of the velocity, we can also write

$$a = \frac{d(v)}{dt} = \frac{dv}{dt}, \quad (8.10)$$

which are common ways of writing a second derivative.

We have noted how total velocity is equal to the integral of the acceleration. This is just the opposite of $v = at$ — if you integrate the acceleration, the result is the integral of the velocity, so distance can be found by twice integrating the acceleration.

In the foregoing discussion of motion was in only one dimension, and experiments only in one dimension were done in three dimensions. Consider a particle P which moves in three dimensions in a very narrow window. As we began our discussion of the three-dimensional case of a moving car by observing the distance of the car from its starting point in $1/1000$ th miles. We then discussed velocity in terms of three distances, x , y , z , and acceleration in terms of changes in velocity. You can see the three-dimensional motion analogously. It will be simpler to illustrate the motion on a two-dimensional diagram, and then extend the ideas to three dimensions. We establish a pair of axes at right angles to each other, and determine the position of the particle along each axis by measuring how far it is from each of the two axes. Thus each position is given in terms of x -distance and y -distance, and the motion can be seen as that by constructing a table in which both these distances are given as functions of time. Extension of this process to three dimensions requires only one additional right angle to the first two, and measuring a third distance, the z -distance. (The distances are *not* measured from coordinate planes instead of lines.) Having established a table with x - and y -distances, we can determine the velocity. We then find the components of velocity in each direction. The horizontal part of the velocity, or *horizontal component*, is the derivative of the x -distance with respect to the time, or

$$v_x = \frac{dx}{dt}. \quad (8.11)$$

Similarly, the vertical part of the velocity, or *vertical component*, is

$$v_y = \frac{dy}{dt}. \quad (8.12)$$

As for circumference,

$$v_z = \frac{dz}{dt}. \quad (8.13)$$

Now given the components of velocity, how can we find the velocity along the actual path of motion? In the two-dimensional case, consider two successive positions of the particle separated by a short distance ds and a short time interval $dt = t_2 - t_1$. The horizontal distance traveled horizontally is distance $\Delta x = v_x dt$, and vertically a distance $\Delta y = v_y dt$. (The symbol " \sim " is read "is approximately.") The total distance traveled is approximately

$$ds \sim \sqrt{(\Delta x)^2 + (\Delta y)^2}, \quad (8.14)$$

as shown in Fig. 8-1. The approximate velocity during this interval can be obtained by dividing by dt and by letting $ds \rightarrow 0$, or at the beginning of the interval.

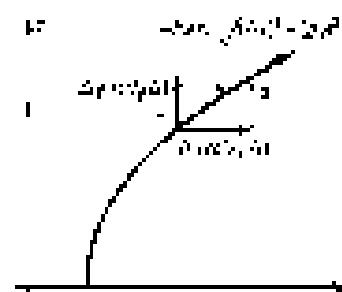


Fig. 8-1 Description of the motion of a body in two dimensions and the computation of its velocity.

We then get the relation 35

$$\tau = \frac{dr}{dt} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} = \sqrt{1 - r^2} \quad (3.5)$$

In three dimensions the result is

$$\tau = \sqrt{V^2 + \frac{d\theta}{dt} + \frac{d\phi}{dt}}. \quad (3.6)$$

To this same 3D as we defined velocities, we can define accelerations; we have an x -component of acceleration \ddot{x}_x , which is the derivative of \dot{x}_x , the x -component of the velocity (that is, $\ddot{x}_x = d\dot{x}_x/dt$), the second derivative of x with respect to t , and so on.

Let us consider one simple example of projectile motion in a plane. We shall take a z -axis to where a body moves horizontally with a constant velocity v_0 , and at the same time under gravity it falls vertically with a constant acceleration $-g$; what is its motion? We can say $\dot{x}_x = v_0$ and $\ddot{x}_x = 0$. Since the velocity s is constant,

$$s = \omega, \quad (3.7)$$

and since the downward acceleration $-g$ is constant, the distance r the object falls can be written as

$$r = -\frac{1}{2}gt^2. \quad (3.8)$$

What is the curve of its path, i.e., relation between y and x ? We can eliminate t from (3.7) and (3.8), since $t = \omega/\dot{x}_x$. When we make this substitution we find that

$$y = -\frac{g}{2\omega^2}x^2. \quad (3.9)$$

This relation between y and x may be considered as the equation of the path of the falling body. When this equation is plotted we obtain a curve now called a parabola, any freely falling body that is shot out in any direction will travel in a parabola, as shown in Fig. 3-4.



Fig. 3-4. The parabola described by a falling body with initial horizontal velocity.

Newton's Laws of Dynamics

9-1 Momentum and force

The discovery of the laws of dynamics, or the laws of motion, was a cornerstone moment in the history of science. Before Newton's time, the motions of things like the planets were, say, roughly, but far from clear and complete understanding. Even the slight deviations from Kepler's laws, due to the perturbations of the planets, were unpredictable. The kinetics of pendulums, oscillators with springs and so on, in their smallness, could all be analyzed completely by Newton's laws were it not for one. So it is with this chapter. Before this chapter we could not calculate how a mass on a spring would move, much less predict the orbit of a planet on the year. This is not true now. After this chapter we will be able to calculate not only the motion of the sun, "planetary motion," but also the motion of the planet Venus, just as in Chapter 4. In Chapter 9 we will learn about the planet Venus, just as in Chapters 4 and 8, and so on.

Galileo made a great advance in the understanding of motion when he discovered the principle of inertia. An object in a straight line, if not disturbed, continues to move with a constant velocity in a straight line. "It was originally moving, and it continues to move still; it has just been moving still." Of course this was quite unlike the common sense (or "folklore") that people back then believed, but "that's how it is and it's in itself" - "nothing against the law." A logical argument therefore had the right rule, and that argument was supplied by Galileo.

Of course, the law thing which is needed to rule for finding how an object changes its speed if something is affecting it. That is the formulation of Newton's law, or we call it three laws. The First Law was a mere restatement of the Galilean principle of inertia just mentioned. The Second Law gives a specific way of doing this, how the weight, change in direction influences called force, can change velocity and hence the motion of, and we shall discuss that in another time. Here we shall discuss only the Second Law, which asserts that the motion of an object is changed by forces in this way: the greater influence of a gravity related movement is proportional to the force. We shall do this more qualitatively, but let us first explain the law.

Momentum is not the same as weight. A lot of words are used in physics, and they all have precise meaning in physics, although they may not have such precise meanings in everyday language. Momentum is an example, and we must define it precisely. There is a certain part of the motion of an object that is larger, if you're pushing it, pushing it than another object. But it's not because in the usual sense, the one object has more mass. Actually, we must do it the same, "from 'light'" and "heavy" for two masses and two velocities. There is a difference to be understood between the weight of an object and its weight. "How hard it is to get it going again, how much force is needed to accelerate it, weight and inertia are proportional" and so on. Gravity's influence is often taken to be numerically equal, which causes a certain confusion to the students. On Mars, weights would be different but the amount of force needed to accelerate would be the same.

We use the term mass as a quantitative measure of inertia, and we rely primarily on, for example, how much an object is accelerated at a certain speed and measuring how much force we need to keep it in the circle. In this way we find a certain quantity of mass for every object. Now the mass of an object is a product of its mass in increasing its velocity. That Newton's Second Law says

9-1 Momentum and force

9-2 Speed and velocity

9-3 Components of velocity, acceleration and force

9-4 What is the force?

9-5 Meaning of the dynamical equations

9-6 Numerical solution of the equations

9-7 Planetary motion

be written in this form, namely this way:

$$F = -\frac{d}{dt}(mv), \quad (9-1)$$

Now there are several points we have to consider: In writing down very low orbits so far, we have made intuitive sense, simplifications, and assumptions which are at best combined approximations into our "form." Like we may have to come back and study in greater detail exactly who such assumptions, but if we try to do this too soon we shall get confused. Thus at the beginning, we have several things to consider. First, that the mass of an object, is however, it isn't really, but we don't care so much with the Newtonian approximation to many systems, we ignore all the mass, and that, further, when we put two objects together, their masses add. These ideas were of course applied by Newton when he wrote his equations, for otherwise it's meaningless. For example, suppose the most general formula for the velocity; then the momentum would never change. In any case, consider on the low orbits nothing unless you know how the mass changes with velocity. At first we say, it does not change.

Then there are some implications concerning forces. As a useful approximation we think of force as a kind of push or pull that we make work on a body, but we can define it more accurately now that we have the law of motion. The most important thing to realize is that this relationship involves not only changes in the magnitude of the momentum or the velocity but also in their direction. If the mass is constant, then Eq. (9-1) can also be written as

$$\vec{F} = m \frac{d\vec{v}}{dt} = m\vec{a}, \quad (9-2)$$

The acceleration \vec{a} is the rate of change of the velocity, and Newton's Second Law says more than can the effect of a given force varies inversely as the mass; it says also that the derivative of the change in the velocity is of the direction of the force not the same. Thus we must understand that a change in a velocity, or an acceleration, has a wider meaning than its common language. Let velocity of a moving object change by its speeding up, slowing down either a slowdown, we say it accelerates with a negative acceleration, or changing its direction of motion. An acceleration at right angles to the velocity was discussed in Chapter 7. There we saw that an object moving in a circle of radius r with a certain speed v along the circle falls away from a straightline path by a distance equal to $(1/2)(v^2/r)$, or v^2/r . Thus the formula for acceleration at right angles to the motion is

$$a = v^2/r, \quad (9-3)$$

and a force at right angles to the velocity will cause an object to move in a curved path whose radius of curvature can be found by dividing the force by the mass to get the acceleration, and then using (9-3).

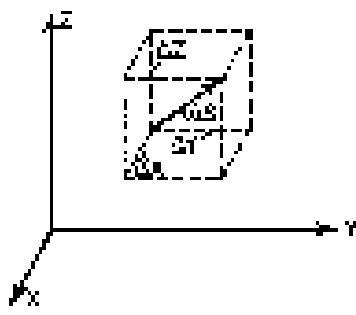


Fig. 9-1. A small displacement of an object.

9-2. Speed and Velocity

In order to make our language more precise, we shall make one further distinction at one use of the words speed and velocity. Ordinarily we think of speed and velocity as being the same, and in ordinary language they are the same. But in physics we have taken coverage of the term v ; there are two words and we've chosen to use them to distinguish each other. We carefully distinguish velocity, which has both magnitude and direction, from speed, which we choose to mean the magnitude of the velocity, but which does not include the direction. We can further elaborate this more precisely by defining three, the x , y , and z -components of an object's velocity with time. Suppose, for example, that at a certain instant an object is moving as shown in Fig. 9-1. In a given small interval of time it will cover certain distances due to the translation, dx in the x -direction, and dy in the y -direction. The resultant net of these three coordinate changes is a displacement as along the diagonal of a parallelepiped whose sides are dx , dy , and dz . In terms

of the velocity. The displacement Δx is the x -component of the velocity times Δt , and similarly for Δy and Δz :

$$\Delta x = v_x \Delta t, \quad \Delta y = v_y \Delta t, \quad \Delta z = v_z \Delta t. \quad (9.5)$$

9.3 Components of velocity, acceleration, and force

In Eq. (9.4) we have recorded the velocity in components by taking first that the object is moving in the x -direction, the y -direction, and the z -direction. The velocity is completely specified, but in an unimpressive direction. It is good to summarize velocity in its three-dimensional components:

$$v = \sqrt{v_x^2 + v_y^2 + v_z^2} \quad v_x = v \cos \theta_x \quad v_y = v \cos \theta_y \quad (9.6)$$

On the other hand, the speed of the object is

$$|v| = \sqrt{v_x^2 + v_y^2 + v_z^2}. \quad (9.7)$$

Now suppose that, because of the action of a force, the velocity changes in some particular direction at a different time. Take, as shown in Fig. 9.2, the change in the x , y , and z components of velocity. The change in the x -component of the velocity in the x -direction is $\Delta v_x = v_x - v_0$; it is v_0 in what we call a component of the acceleration. Similarly, we see that $\Delta v_y = v_y - v_0$, and $\Delta v_z = v_z - v_0$. In these terms, we can put Newton's Second Law, $F = m a$, in two dimensions as the second law in the y -direction, or in the z -direction, so the component of the force in the y -, or z -direction is equal to the mass times the rate of change of the corresponding component of velocity:

$$\begin{aligned} F_y &= m a_{y,y} = m (\Delta v_y / \Delta t) = m v_0, \\ F_z &= m a_{z,z} = m (\Delta v_z / \Delta t) = m v_0, \\ F_x &= m a_{x,x} = m (\Delta v_x / \Delta t) = m v_0. \end{aligned} \quad (9.8)$$

Just as the velocity and acceleration have been resolved into components by projecting onto axes, representing one quantity along each of three coordinate axes, so, in the same way, a force is resolved. F itself is represented by just a magnitude F . That is, F is a scalar:

$$\begin{aligned} F &= F \cos (\theta_x, F_x) \\ F_y &= F \cos (\theta_y, F), \\ F_z &= F \cos (\theta_z, F) \end{aligned} \quad (9.9)$$

where F is the magnitude of F ($F = \sqrt{F_x^2 + F_y^2 + F_z^2}$) and θ is the angle between the vector and the direction of F .

Newton's second law is given in complete form as Eq. (9.8). If we know the forces on an object and resolve the forces in the x , y , and z components, then we can find the motion of the object from one moment in time to the next. Let us consider a simple example. Suppose there are no forces in the x and y directions; the only force acting is the gravitational, say, vertically. Equation (9.8) tells us that there would be changes in the velocity in the vertical direction, but no changes in the horizontal direction. This was demonstrated in a special apparatus in Chapter 7 (see Fig. 7.1). A falling body moves horizontally without any change in the horizontal motion, while it moves vertically the same way as it would move in the horizontal motion were zero. In other words, motions in the x , y , and z -directions are independent if they are in the horizontal.

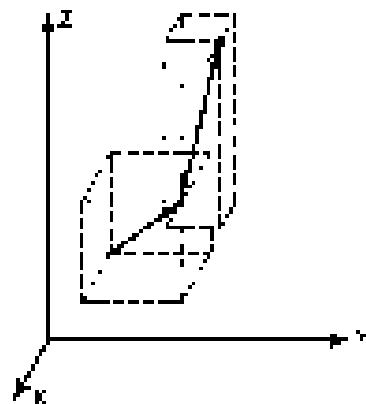


Fig. 9.2. A change in velocity is also both the magnitude and direction of a_{\perp} .

9.4 What is the Force?

In order to use Newton's laws, we now need some formulae for the forces. These laws are independent of the forces. If an object is undergoing some motion, it is always, find it. Our program for the future of dynamics, as set by Isaac Newton,

law for the force. Newton himself was not to give such examples. In the case of gravity he gave a specific formula for the force, i.e. the law of other forces he gave even less of the information. In his Second Law, which we will study in the next chapter, having to do with the causality of motion and reaction.

Extending our previous example, what are the forces on objects near the earth's surface? Near the earth's surface, the force of the vertical direction due to gravity is proportional to the mass of the object and is nearly independent of height, for heights small compared with the earth's radius R : $F = GmM/R^2 = mg$, where $g = GM/R^2$ is called the acceleration of gravity. Thus the law of gravity tells us that weight is proportional to mass; the force is in the vertical direction and is the mass times g . Again we find that the motion in the horizontal direction is at constant velocity. The interesting motion is in the vertical direction, and Newton's Second Law tells us

$$mg = m(gx'/dt^2) \quad (9.9)$$

Dividing both sides we find that the acceleration in the vertical motion is constant and equal to g . This is of course the well-known law of free fall under gravity, which leads to the equation

$$\begin{aligned} x' &= v_0 t + \frac{1}{2} g t^2 \\ x &= x_0 + v_0 t + \frac{1}{2} g t^2. \end{aligned} \quad (9.10)$$

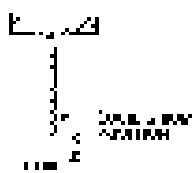


Fig. 9.3. A mass on a spring.

As another example, let us suppose that we have here able to pull a weight (Fig. 9.3) which applies a force proportional to the distance and directed oppositely to x' . If we start about x_0 , which is at one we balanced out by the mutual stretch of the spring, and apply a uniform force, we see that if we pull the mass down, the spring pulls up, while if we push it up the spring pulls down. This machine has been designed exactly so that the mass is periodic, the more we pull it up, in exact proportion to the displacement from the balanced condition, and the force applied is similarly proportional to how far we pull down. Is the whole the dynamics of this machine, we see a rather beautiful motion—up, down, up, down, ... The question is, will Newton's equations correctly describe this motion? Let us see how we can exactly calculate how it moves with this periodic oscillation, by applying Newton's law (9.7). In the present instance, the equation is

$$-kx = m(gx'/dt^2) \quad (9.11)$$

From which we see that when the velocity in the x -direction changes at a rate proportional to x . Having this to go on by repeating numerous oscillations, we shall imagine either that the scale of time has changed or that there is an accident in the name, to this we compare to have $v_0 = 1$. But we shall try to solve the equation

$$gx'/dt^2 = -x \quad (9.12)$$

provided, we must know what x is, but of course we know that the velocity is the rate of change of the position.

9.5 Meaning of the dynamical equations

Now let us try to analyze just what Eq. (9.12) means. Suppose that at a given time t the object has a certain velocity v_0 and position x_0 . What is the velocity and what is the position at a slightly later time $t + \Delta t$? If we can answer this question a problem arises, for then we can start with the given condition and compute how it changes for the first instant, the next instant, the next instant, and so on, and in this way fully analyze the motion. To be specific, let us assume that at the time $t = 0$ we are given that $v_0 = 1$ and $x_0 = 0$. Why does the object move at all? Because there is a force in. Unless it is at rest initially except $x = 0$, if $v_0 = 0$ no force is applied. Therefore the velocity which is zero is not to change, because of the law of motion. Once it starts to build up some velocity it cannot start to move less, and so on. Now at any time t , if Δt is very small,

as may happen. In position, we think in terms of the position x . $x(0) = 100$ and the velocity at time t is a very good approximation as

$$x(t) \approx x(0) + v_0 t \quad (9.10)$$

The problem is, the more accurate this expression is, the less useful it becomes because v_0 is not constantly zero. Now when when "the velocity" is zero, in order to get the velocity later, say exactly at the time $t + \Delta t$, we need to know how the velocity changes, the acceleration. And how are we going to find the acceleration? That is where the law of dynamics comes in. The law of dynamics tells us what the acceleration is. If, like the acceleration, $a = 0$,

$$v(t + \Delta t) = v(t) + a \Delta t \quad (9.11)$$

$$= v(0) + a t \Delta t. \quad (9.12)$$

Equation (9.12) is merely *definitional*; it says that a velocity changes because of the presence of acceleration. But Eq. (9.11) is *dynamical*; that is, it relates the acceleration to the velocity. So that at the particular time for the particular problem, you can replace the acceleration by $-a(t)$. Therefore, if we knew both the $v(0)$ and $a(t)$ at given time, we know the acceleration, which tells us the new velocity, and we know the new position—this is how the mechanics works. The velocity changes a little bit because of the force, and the position changes a little bit because of the velocity.

9.6 Numerical solution of the equations

Now let us really solve the problem. Suppose that we take $\Delta t = 0.100$ sec. At the end of the first Δt we find that this is not small enough we may have to go round and round you with $v = 0.010$ sec. Starting with our initial value $v(0) = 1.00$ m/sec $v(0.100) = 1.00$. It is the x position $x(0)$; and the velocity which is zero at $t = 0.100$ sec. Thus $v(0.1)$ is still 1.00, because it has not yet started to move. But the new velocity at 0.100 sec will be the old velocity $v(0)$ + 0 plus the current acceleration. The acceleration is $a(0) = -1.00$. Thus

$$v(0.1) = 0.00 + -0.10 \times 1.00 = -0.10$$

Now at 0.200 sec:

$$\begin{aligned} v(0.2) &= v(0.1) + a(0.1) \\ &= -0.10 + 0.10 \times 0.10 = 0.00 \end{aligned}$$

and

$$\begin{aligned} v(0.2) &= v(0.1) - a(0.1) \\ &= 0.00 - 0.10 \times 0.10 = -0.10. \end{aligned}$$

And so, so and so and on, we can calculate the rest of the motion, and that is just what we shall do. However, for practical purposes, there are some little tricks by which we can increase the accuracy. For example, in this case, after we have started it, we would find the motion only moves slightly because $a = -0.100$ sec $k = 1.0$ is only, and we would have to go through a very long Δt to move, say, $x = 0.01$. Then to go through a reasonable total time interval would take a lot of cycles of computation. So we shall examine the trick in a way that will increase the precision of our calculations using the same time interval $\Delta t = 0.100$ sec. This can be done if we make a small improvement in the technique of the analysis.

Notice that the new position x is defined just implies the time it, *exactly*, bands the velocity v . But the velocity $v(0)$? The velocity at the beginning of the time interval is one velocity and the velocity at the end of the time interval is another velocity. Our improvement is to use the velocity *midway* between. If we know the speed now, but the speed is changing, then we are not going to get the right answer by going at the same speed all time. We would use some speed between the "true" speed and the "fast" speed at the end of the interval. The same considerations also apply to the velocity to compute the velocity changes. So

Table 9.1

solution of $v' = v - at$ Interval, $\Delta t = 0.10 \text{ sec}$

t	v	v'	a
0.0	1.000	0.000	-1.000
0.1	0.991	-0.100	0.995
0.2	0.980	0.120	-0.990
0.3	0.962	0.158	-0.982
0.4	0.942	0.193	-0.972
0.5	0.917	0.224	-0.957
0.6	0.887	0.251	-0.932
0.7	0.852	0.273	-0.897
0.8	0.813	0.291	-0.854
0.9	0.768	0.305	-0.803
1.0	0.718	0.315	-0.744
1.1	0.662	0.321	-0.677
1.2	0.600	0.323	-0.607
1.3	0.532	0.321	-0.534
1.4	0.458	0.315	-0.459
1.5	0.380	0.315	-0.384
1.6	0.302	0.311	-0.309
1.7	0.223	0.304	-0.232
1.8	0.143	0.294	-0.152
1.9	0.062	0.281	-0.071
2.0	-0.039	0.265	-0.029
2.1	-0.179	0.247	-0.019
2.2	-0.379	0.227	-0.012
2.3	-0.629	0.205	-0.007
2.4	-0.840	0.181	-0.004
2.5	-0.991	0.155	-0.002
2.6	-1.000	0.128	-0.001
2.7	-0.991	0.099	0.002
2.8	-0.980	0.068	0.004
2.9	-0.962	0.036	0.007
3.0	-0.942	0.003	0.010
3.1	-0.917	-0.031	0.014
3.2	-0.887	-0.062	0.018
3.3	-0.852	-0.088	0.022
3.4	-0.813	-0.108	0.026
3.5	-0.768	-0.124	0.030
3.6	-0.718	-0.136	0.034
3.7	-0.662	-0.144	0.038
3.8	-0.600	-0.150	0.041
3.9	-0.532	-0.153	0.044
4.0	-0.458	-0.154	0.046
4.1	-0.380	-0.153	0.047
4.2	-0.302	-0.150	0.048
4.3	-0.223	-0.145	0.048
4.4	-0.143	-0.138	0.047
4.5	-0.062	-0.129	0.045
4.6	-0.039	-0.118	0.043
4.7	-0.179	-0.105	0.040
4.8	-0.379	-0.090	0.036
4.9	-0.629	-0.073	0.031
5.0	-0.840	-0.055	0.026
5.1	-0.991	-0.035	0.020
5.2	-1.000	-0.014	0.014
5.3	-0.991	0.005	0.008
5.4	-0.980	0.025	0.004
5.5	-0.962	0.045	0.002
5.6	-0.942	0.064	0.001
5.7	-0.917	0.081	0.000
5.8	-0.887	0.096	0.000
5.9	-0.852	0.109	0.000
6.0	-0.813	0.121	0.000
6.1	-0.768	0.131	0.000
6.2	-0.718	0.139	0.000
6.3	-0.662	0.145	0.000
6.4	-0.600	0.150	0.000
6.5	-0.532	0.153	0.000
6.6	-0.458	0.155	0.000
6.7	-0.380	0.155	0.000
6.8	-0.302	0.154	0.000
6.9	-0.223	0.151	0.000
7.0	-0.143	0.146	0.000
7.1	-0.062	0.139	0.000
7.2	-0.039	0.130	0.000
7.3	-0.179	0.119	0.000
7.4	-0.379	0.106	0.000
7.5	-0.629	0.091	0.000
7.6	-0.840	0.075	0.000
7.7	-0.991	0.057	0.000
7.8	-1.000	0.038	0.000
7.9	-0.991	0.017	0.000
8.0	-0.980	-0.005	0.000
8.1	-0.962	-0.025	0.000
8.2	-0.942	-0.045	0.000
8.3	-0.917	-0.064	0.000
8.4	-0.887	-0.081	0.000
8.5	-0.852	-0.096	0.000
8.6	-0.813	-0.109	0.000
8.7	-0.768	-0.131	0.000
8.8	-0.718	-0.145	0.000
8.9	-0.662	-0.153	0.000
9.0	-0.600	-0.155	0.000
9.1	-0.532	-0.155	0.000
9.2	-0.458	-0.154	0.000
9.3	-0.380	-0.153	0.000
9.4	-0.302	-0.151	0.000
9.5	-0.223	-0.146	0.000
9.6	-0.143	-0.139	0.000
9.7	-0.062	-0.130	0.000
9.8	-0.039	-0.119	0.000
9.9	-0.179	-0.106	0.000
10.0	-0.379	-0.091	0.000
10.1	-0.629	-0.075	0.000
10.2	-0.840	-0.057	0.000
10.3	-0.991	-0.038	0.000
10.4	-1.000	-0.017	0.000
10.5	-0.991	0.005	0.000
10.6	-0.980	0.025	0.000
10.7	-0.962	0.045	0.000
10.8	-0.942	0.064	0.000
10.9	-0.917	0.081	0.000
11.0	-0.887	0.096	0.000
11.1	-0.852	0.109	0.000
11.2	-0.813	0.131	0.000
11.3	-0.768	0.145	0.000
11.4	-0.718	0.153	0.000
11.5	-0.662	0.155	0.000
11.6	-0.600	0.155	0.000
11.7	-0.532	0.154	0.000
11.8	-0.458	0.153	0.000
11.9	-0.380	0.151	0.000
12.0	-0.302	0.146	0.000
12.1	-0.223	0.139	0.000
12.2	-0.143	0.130	0.000
12.3	-0.062	0.119	0.000
12.4	-0.039	0.106	0.000
12.5	-0.179	0.091	0.000
12.6	-0.379	0.075	0.000
12.7	-0.629	0.057	0.000
12.8	-0.840	0.038	0.000
12.9	-0.991	0.017	0.000
13.0	-1.000	0.005	0.000
13.1	-0.991	-0.005	0.000
13.2	-0.980	-0.025	0.000
13.3	-0.962	-0.045	0.000
13.4	-0.942	-0.064	0.000
13.5	-0.917	-0.081	0.000
13.6	-0.887	-0.096	0.000
13.7	-0.852	-0.109	0.000
13.8	-0.813	-0.131	0.000
13.9	-0.768	-0.145	0.000
14.0	-0.718	-0.153	0.000
14.1	-0.662	-0.155	0.000
14.2	-0.600	-0.155	0.000
14.3	-0.532	-0.154	0.000
14.4	-0.458	-0.153	0.000
14.5	-0.380	-0.151	0.000
14.6	-0.302	-0.146	0.000
14.7	-0.223	-0.139	0.000
14.8	-0.143	-0.130	0.000
14.9	-0.062	-0.119	0.000
15.0	-0.039	-0.106	0.000
15.1	-0.179	-0.091	0.000
15.2	-0.379	-0.075	0.000
15.3	-0.629	-0.057	0.000
15.4	-0.840	-0.038	0.000
15.5	-0.991	-0.017	0.000
15.6	-1.000	0.005	0.000
15.7	-0.991	0.025	0.000
15.8	-0.980	0.045	0.000
15.9	-0.962	0.064	0.000
16.0	-0.942	0.081	0.000
16.1	-0.917	0.096	0.000
16.2	-0.887	0.109	0.000
16.3	-0.852	0.131	0.000
16.4	-0.813	0.145	0.000
16.5	-0.768	0.153	0.000
16.6	-0.718	0.155	0.000
16.7	-0.662	0.155	0.000
16.8	-0.600	0.155	0.000
16.9	-0.532	0.154	0.000
17.0	-0.458	0.153	0.000
17.1	-0.380	0.151	0.000
17.2	-0.302	0.146	0.000
17.3	-0.223	0.139	0.000
17.4	-0.143	0.130	0.000
17.5	-0.062	0.119	0.000
17.6	-0.039	0.106	0.000
17.7	-0.179	0.091	0.000
17.8	-0.379	0.075	0.000
17.9	-0.629	0.057	0.000
18.0	-0.840	0.038	0.000
18.1	-0.991	0.017	0.000
18.2	-1.000	0.005	0.000
18.3	-0.991	-0.005	0.000
18.4	-0.980	-0.025	0.000
18.5	-0.962	-0.045	0.000
18.6	-0.942	-0.064	0.000
18.7	-0.917	-0.081	0.000
18.8	-0.887	-0.096	0.000
18.9	-0.852	-0.109	0.000
19.0	-0.813	-0.131	0.000
19.1	-0.768	-0.145	0.000
19.2	-0.718	-0.153	0.000
19.3	-0.662	-0.155	0.000
19.4	-0.600	-0.155	0.000
19.5	-0.532	-0.154	0.000
19.6	-0.458	-0.153	0.000
19.7	-0.380	-0.151	0.000
19.8	-0.302	-0.146	0.000
19.9	-0.223	-0.139	0.000
20.0	-0.143	-0.130	0.000
20.1	-0.062	-0.119	0.000
20.2	-0.039	-0.106	0.000
20.3	-0.179	-0.091	0.000
20.4	-0.379	-0.075	0.000
20.5	-0.629	-0.057	0.000
20.6	-0.840	-0.038	0.000
20.7	-0.991	-0.017	0.000
20.8	-1.000	0.005	0.000
20.9	-0.991	0.025	0.000
21.0	-0.980	0.045	0.000
21.1	-0.962	0.064	0.000
21.2	-0.942	0.081	0.000
21.3	-0.917	0.096	0.000
21.4	-0.887	0.109	0.000
21.5	-0.852	0.131	0.000
21.6	-0.813	0.145	0.000
21.7	-0.768	0.153	0.000
21.8	-0.718	0.155	0.000
21.9	-0.662	0.155	0.000
22.0	-0.600	0.155	0.000
22.1	-0.532	0.154	0.000
22.2	-0.458	0.153	0.000
22.3	-0.380	0.151	0.000
22.4	-0.302	0.146	0.000
22.5	-0.223	0.139	0.000
22.6	-0.143	0.130	0.000
22.7	-0.062	0.119	0.000
22.8	-0.039	0.106	0.000
22.9	-0.179	0.091	0.000
23.0	-0.379	0.075	0.000
23.1	-0.629	0.057	0.000
23.2	-0.840	0.038	0.000
23.3	-0.991	0.017	0.000
23.4	-1.000	0.005	0.000
23.5	-0.991	-0.005	0.000
23.6	-0.980	-0.025	0.000
23.7	-0.962	-0.045	0.000
23.8	-0.942	-0.064	0.000
23.9	-0.917	-0.081	0.000
24.0	-0.887	-0.096	0.000
24.1	-0.852	-0.109	0.000
24.2	-0.813	-0.131	0.000
24.3	-0.768	-0.145	0.000
24.4	-0.718	-0.153	0.000
24.5	-0.662	-0.155	0.000
24.6	-0.600	-0.155	0.000
24.7	-0.532	-0.154	0.000
24.8	-0.458	-0.153	0.000
24.9	-0.380	-0.151	0.000
25.0	-0.302	-0.146	0.000
25.1	-0.223	-0.139	0.000
25.2	-0.143	-0.130	0.000
25.3	-0.062	-0.119	0.000
25.4	-0.039	-0.106	0.000
25.5	-0.179	-0.091	0.000
25.6	-0.379	-0.075	0.000
25.7	-0.629	-0.	

the planet times the rate of change of its velocity in the x-direction. Thus we find the following steps:

$$\begin{aligned} m(\dot{x}_x, \ddot{x}_x) &= -GMm_x/r^2, \\ m(\dot{x}_y, \ddot{x}_y) &= -GMm_y/r^2, \\ r &= \sqrt{x^2 + y^2}. \end{aligned} \quad (3.17)$$

This, then, is the set of equations we must solve. Again, in order to simplify the numerical work, we shall suppose that the unit of time, or the mass of the sun, are both so adjusted (or fixed in units) that $GM = 1$. For our specific example we shall suppose that the initial position of the planet is $x = \pi = 0.500$ and $y = 0.000$, and that the velocity is $\dot{x} = 1$ in the x -direction at the start, and is of magnitude 1.600. Now how do we make the calculation? We again make a table with columns for the time, the x -position, the x -velocity x_x , and the x -acceleration \ddot{x}_x , each separated by a double line, three columns for y -position, y -velocity x_y , and acceleration \ddot{x}_y in the y -direction. In order to get the accelerations we begin by need Eq. (3.17); it tells us that the acceleration in the x -direction is $-\dot{x}_x/r$, and the acceleration in the y -direction is $-\dot{x}_y/r$, and therefore the square root of $x^2 + y^2$. Thus, given x and y we just do a little calculating on the side. Taking the square root of the sum of the squares to find r and then, to get ready to calculate the two accelerations, it is useful also to evaluate $1/r^2$. This work can be done rather easily by using a slide rule of logarithmic values and reciprocals; then we need only multiply x by $1/r^2$, which we do on a slide rule.

Our calculation thus proceeds by the following steps, taking time intervals $\epsilon = 0.001$: Initial velocities $\dot{x} = 1$

$$\begin{aligned} x(0) &= 0.500 & y(0) &= 0.000 \\ \dot{x}(0) &= 0.000 & \dot{y}(0) &= 1.600 \end{aligned}$$

From these we find

$$\begin{aligned} r(0) &= 0.500 & 1/r^2(0) &= 8.000 \\ \dot{x}_x &= -4.000 & \dot{r}_x &= 0.070 \end{aligned}$$

Thus we may calculate the velocities $-\dot{x}_x/r$ and $-\dot{x}_y/r$:

$$\begin{aligned} \dot{x}_x(0.1) &= 0.000 - 4.000 \times 0.070 = -0.280; \\ \dot{x}_y(0.1) &= 1.600 + 0.000 \times 0.070 = 1.600. \end{aligned}$$

Now our main calculation begins:

$$\begin{aligned} x(0.1) &= 0.500 - 0.00 \times 0.1 = 0.500 \\ \dot{x}(0.1) &= 0.0 - 1.60 \times 0.1 = 0.160 \\ r &= \sqrt{0.500^2 + 0.160^2} = 0.507 \\ \dot{x}_x(0.1) &= 0.480 \times 0.507 = 0.24 \\ \dot{x}_y(0.1) &= -0.160 \times 0.507 = -0.126 \\ \dot{x}_x(0.2) &= 0.000 - 0.04 \times 0.1 = -0.008 \\ \dot{x}_y(0.2) &= 1.600 - 0.08 \times 0.1 = 1.508 \\ \dot{x}(0.2) &= 0.440 - 0.508 \times 0.1 = 0.422 \\ \dot{x}_x(0.3) &= 0.160 - 0.00 \times 0.1 = 0.160 \end{aligned}$$

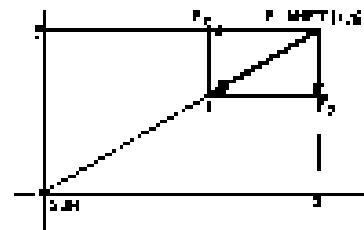


Fig. 9-5. The force of gravity on a planet.

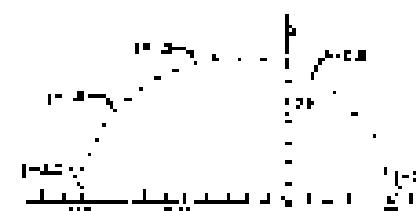


Fig. 9-6. The calculated motion of a planet around the sun.

In this way we obtain the values given in Table 9-2, and in 20 steps or so we have closed the planet initially around the sun. In Fig. 9-6 are plotted the x - and y -coordinates given in Table 9-2. The dots represent the positions of the planet at times t (unit: 0.001). We see that at $t = 0$ start the planet moves rapidly

Table 9-2
Solutions of $d^2\theta/dt^2 + \omega_0^2/r^2 \sin^2\theta d\phi/dt = -\omega_0^2/r^2$, $r = \sqrt{\mu^2 - \epsilon}$
Interval $\epsilon \in [0, 10]$

Orbit	ω_0	$\theta = 0$				$\theta = \pi/2$		$\theta = \pi$	
		a_r	\dot{r}_r	$\dot{\theta}_r$	$\dot{\phi}_r$	\ddot{r}_r	$\ddot{\theta}_r$	$\ddot{\phi}_r$	
1.0	0.300	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.1	-0.480	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.2	-0.421	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.3	0.397	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.4	-0.332	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.5	0.313	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.6	-0.285	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.7	-0.247	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.8	-0.208	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1.9	-0.169	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.0	-0.129	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.1	-0.089	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.2	-0.049	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.3	-0.009	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.4	0.031	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.5	0.079	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.6	0.127	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.7	0.175	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.8	0.223	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
2.9	0.271	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.0	0.319	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.1	0.367	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.2	0.415	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.3	0.463	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.4	0.511	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.5	0.559	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.6	0.607	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.7	0.655	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.8	0.703	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
3.9	0.751	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.0	0.799	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.1	0.847	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.2	0.895	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.3	0.943	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.4	0.991	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.5	1.039	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.6	1.087	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.7	1.135	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.8	1.183	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
4.9	1.231	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.0	1.279	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.1	1.327	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.2	1.375	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.3	1.423	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.4	1.471	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.5	1.519	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.6	1.567	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.7	1.615	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.8	1.663	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
5.9	1.711	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.0	1.759	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.1	1.807	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.2	1.855	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.3	1.903	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.4	1.951	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.5	1.999	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.6	2.047	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.7	2.095	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.8	2.143	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
6.9	2.191	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.0	2.239	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.1	2.287	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.2	2.335	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.3	2.383	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.4	2.431	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.5	2.479	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.6	2.527	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.7	2.575	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.8	2.623	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
7.9	2.671	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.0	2.719	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.1	2.767	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.2	2.815	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.3	2.863	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.4	2.911	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.5	2.959	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.6	3.007	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.7	3.055	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.8	3.103	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
8.9	3.151	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.0	3.199	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.1	3.247	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.2	3.295	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.3	3.343	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.4	3.391	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.5	3.439	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.6	3.487	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.7	3.535	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.8	3.583	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
9.9	3.631	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
10.0	3.679	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000

5-8

5-9

5-10

5-11

5-12

5-13

5-14

5-15

5-16

5-17

5-18

5-19

5-20

5-21

5-22

5-23

5-24

5-25

5-26

5-27

5-28

5-29

5-30

5-31

5-32

5-33

5-34

5-35

5-36

5-37

5-38

5-39

5

and at the end it increases slowly, and so the shape of the curve is determined. Thus we see that we really do know how to calculate the motion of planets.

Now let us see how we can calculate the position of Neptune, Jupiter, Mars, etc., in any year, given the planet. If we have a great number of planets, and let the sun move slow, can we do the same thing? Of course we can. We can take as reference a particular planet, let us say planet number i , when last at position (x_1, y_1, z_1) ; $i = 1$ may represent the sun, $i = 2$ Mercury, $i = 3$ Venus, and so on. We must know the positions of all the bodies. The force acting on i is due to all the other bodies which are located, let us say at positions (x_j, y_j, z_j) . Therefore the equations are

$$\begin{aligned} M_1 \cdot \frac{\partial \phi_1}{\partial t} + \sum_{j=1}^n &= G(m, m_1)(x_j - x_1), \\ M_2 \cdot \frac{\partial \phi_2}{\partial t} + \sum_{j=1}^n &= G(m, m_2)(x_j - x_2), \\ M_3 \cdot \frac{\partial \phi_3}{\partial t} + \sum_{j=1}^n &= G(m, m_3)(x_j - x_3). \end{aligned} \quad (9.13)$$

For the next few days, we will be dedicating time to the planning and design phase.

$$s_1 = \sqrt{(\bar{s}_1 - s_1) + (\bar{(s_{1-1})}(\bar{Y}_{1-1} - \bar{s}_1))} \quad (217)$$

Also, $\sum_{i=1}^n \omega_i > 0$ since over all masses of i , ω_i other bodies' orbital velocities, for $i = 1$. Thus all we have to do is to make many iterations, few more estimates, we never run out time, for the motions of Jupiter, time for the motions of Saturn, and so on. Each time we have all initial positions and velocities we can calculate all the accelerations from Eq. (9.18) by first calculating all the distances using Eq. (9.19). How long will it take to do this? If you do it at home, it will take a very long time! But in modern times we have machines which do astronomical calculations at very good computing machine speed. I mentioned earlier that it is a millionth of a second, to do an addition. To do a multiplication takes longer, say 10 microseconds. It may be that in one cycle of iteration, according to our problem, we may have 300 multiplications, or something like this, so one cycle will take 300 microseconds. This means that we can do 3000 cycles of computation per second. In other words, if we were up, of say, one part in a billion, we would need 3×10^9 cycles to correspond to one revolution of a satellite around the sun, that corresponds to 1.6 minutes. One of 100 seconds or about two minutes. Thus it takes only two minutes to follow Jupiter around the sun with all the pre-estimates of all the planets correct to one part in a billion, by this method! It turns out that the error comes down as the square of the time scale. If we make the interval a thousand times smaller, it is a million times more accurate. So, let us trace the motion of Mars again starting

So, as we see, the basic idea of the method is to calculate even the motion of a massless spring. Now, armed with the machinery of Newton's laws, we can not only calculate such simple motions but also, given sufficient time, the motion of every tiny particle in one of the planets, to as many a degree of precision as we wish.

Conservation of Momentum

10-1 Newton's Third Law

On the basis of Newton's second law of motion, which gives the relation between the acceleration of a body and the force acting on it, any problem in mechanics can be solved in principle – or requires, to summarize the motion of a few particles, one can use the numerical methods developed in the preceding chapter. But there are just reasons to make further study of Newton's laws. First, there are quite simple cases of motion which can be analyzed not only by numerical methods, but also by direct mathematical analysis. For example, although we know that the acceleration of a falling body is $g = 9.8 \text{ m/s}^2$, and from this fact could calculate the motion by numerical methods, it is much easier and more satisfactory to analyze the motion and find the general solution, $x = x_0 + v_0 t - \frac{1}{2} g t^2$. In fact, however, although we can work out the problem of a harmonic oscillator by numerical methods, it is also possible to show analytically that the potential is given by a trigonometric function. This solution is necessary to prove all that nothing is difficult when there is a simple and more accurate way to get the result. In the same manner, although the motion of one body around the sun, determined by gravitation at a fixed point by solving the equations of motion of Lagrange, which show the general shape of the orbit, it is necessary to get the exact shape when analysis reduces to a perfect ellipse.

Thus far, however, very few problems which can be solved exactly by any means. In the case of the harmonic oscillator, for example, if the spring force is not proportional to the displacement, but is something more complicated, one must fall back on the numerical method. Or, if there is, say, a varying mass of the sun, so that the total number of bodies is three, such analysis cannot produce a simple formula for the motion, and in practice the problem is still done numerically. This is the famous three-body problem, which is long and longer than powers of analysis; it is very interesting how long it took people to appreciate the fact that perhaps the power of mathematical analysis was limited and it might be necessary to use the numerical methods. Today an enormous number of problems that cannot be done analytically are solved by numerical methods, and the old three-body problem, which was supposed to have solved, is solved in a manner of solving in exactly the same manner and was described in the preceding chapter, namely, as being a tough nut to crack. However, there are also situations where both methods fail: the single problem can be solved by analytic and the inexpressibly difficult problem by numerical, approximate methods, but the very same disease will always be met by either method. An complicated problem is, for example, the collision of two antinuclei, or a formation of microcosmics of a gas. There are countless particles in a cubic centimeter of gas, and it would be ridiculous to try to make calculations with so many variables (about 10^{23} – a hundred million trillion). Anyways like the majority of problems the mass of a gas or a bunch of atoms, or the motion of the stars in a galaxy cluster, instead of just two or three stars, using standard classical mechanics we cannot do directly, so we have to give up for me.

In the solutions of these we cannot follow details, we need to know some general properties, that is, general theorems or insights which are independent of Newton's laws. One of these is the principle of conservation of energy, which was discussed in Chapter 4. Another is the principle of conservation of momentum, the subject of this chapter. And he can see everything mechanics is when we find there are certain patterns of motion that are repeated in many different circum-

10-2 Newton's Third Law

10-3 Conservation of momentum

10-4 Momentum & conserved!

10-5 Momentum and energy

10-6 Relativistic mechanics

shocks, so it is good to study these problems in some particular circumstances. For example, we shall study oscillations, different kinds of oscillations such as harmonic. In the theory of fluids it does not make much difference what "fluid" is, the laws of the flow are similar. Other problems that we shall study are vibrations in classifications and, in particular, the peculiar phenomena of mechanical waves—sound, vibrations of solids, and so on.

In our discussion of Newton's laws it was explained that these laws are a kind of saying on fact, and "Pay attention to the forces," and that Newton could hardly give things other than the nature of forces. In the case of gravitation, he gives us the complete law of the force. In the case of the very complicated forces between atoms, he was not able to give the right laws for the forces; however, he did give us one rule, one general property of forces, which is expressed in his third law, and that is the total knowledge that Newton had about the nature of forces. The law of gravitation and this principle, but no other.

This principle is the *conservation of momentum*.

What is meant by something of this kind? Suppose we have two small bodies, say particles, and suppose that the first one exerts a force on the second one, pushing it with a certain force. Then, simultaneously, according to Newton's third law, the second particle will push on the first with an equal force, in the opposite direction; otherwise, these forces effectively act at the same time. This is the hypothesis, in fact, that Newton assumed, and it seems to be quite accurate, though not exact (see also [this article](#) for the errors). For the moment we shall take it to be true that action equals reaction. Of course, if there is a third particle on the same line, another law, the law of the third, says that the total force on the first one is equal to the total force on the second, since the third particle, for this law, exerts its own pull on each of the other two. The result is that the total effect on the first two is in some intermediate form, and the forces on the first two particles are, in general, neither equal nor opposite. However, the forces on each particle can be resolved into parts, there being one parallel to the particle to each other interacting particle. Then each pair of particles has corresponding components of mutual interaction that are equal in magnitude and opposite in direction.

10-2 Conservation of momentum

Now what are the interesting consequences of the above relationships? Suppose, for simplicity, that we have just two interacting particles, possibly of different masses, and a mutual force due to the forces between them, the equal and opposite what we call antiparticles? According to Newton's Second Law, since in the case of a change of the momentum, we have nothing but the rate of change of momentum v , if particle 1 is equal to minus the rate of change of momentum p_2 , or exactly

$$\frac{dp_1}{dt} = -\frac{dp_2}{dt} \quad (10-1)$$

Now if the rate of change is always equal and opposite, it follows that the total change is the summation of particle 1 to zero, and opposite to the total change in the momentum of particle 2. This means that if we add the momentum of particle 1 to the momentum of particle 2, the rate of change of the sum of these, due to the mutual forces (called internal forces) between particles, is zero; that is

$$(p_1 + p_2)/dt = 0 \quad (10-2)$$

This is sometimes called the law of the system. If the rate of change of this sum is always zero, there is just another way of saying that the quantity $p_1 + p_2$ does not change. This quantity is also written p , \mathbf{p} , \mathbf{p}_{tot} , and is called the *total momentum* of the two particles. We have now obtained the result that the total momentum of the two particles does not change because of any mutual interactions between them. This statement expresses the law of conservation of momentum.

momentum in that 2-particle example. We conclude that if there is any kind of force, no matter how complicated, between two particles, and we measure or calculate $p_{1,2} = p_{2,1}$, that is, the sum of the two momenta both before and after the forces act, the results should be equal, i.e., the total momentum is a constant.

If we extend the argument to three or more interacting particles in some complicated situation, it is evident that as long as no net force is experienced, the total momentum of all the particles stays constant, since each particle's momentum is constant, due to another, it exactly compensated by the decrease of the second, due to the first. That is, all the internal forces will balance out, and therefore cannot change the total momentum of the particles. (That is, there are no forces from the outside (other than grav.) that are too large to change the total momentum, since the total momentum is a constant.)

It is worth describing what happens if there are forces that do not add up to the total momentum of the particles. In quantum physics, we realize that this is typical. If there are only mutual forces, then, as before, the total momentum of the particles does not change, so maybe less complicated than forces. On the other hand, suppose there are other forces coming from the particles outside the looks of a group. Any force exerted by one particle on another particle, we call an external force. We can then deduce that the sum of all external forces equals the rate of change of the total momentum of all the particles inside, a very useful theorem.

The conservation of the total momentum of a number of interacting particles can be expressed as

$$m_{1,1} + m_{1,2} + \dots + m_{1,n} = \text{a constant}. \quad (10.2)$$

There are no two external forces. There are masses and corresponding velocities of the particles who are not listed 1, 2, 3, 4, The general statement of Newton's Second Law for each particle,

$$F = \frac{d}{dt}(mv), \quad (10.4)$$

is also specified by the components of the second column, in any given dimension; thus, the component of the force on a particle is equal to the component of the rate of change of momentum of that particle, Δ .

$$F_x = \frac{d}{dt}(mv_x), \quad (10.5)$$

and similarly for the y - and z -directions. The other F_y , (10.3) is nearly the same form, and F_z is also similar.

In addition to the law of conservation of momentum, there is another interesting consequence of Newton's Second Law, to be proved later, but merely stated now. This principle is that no laws of physics will be different whether we are moving x - or moving with a uniform speed in a straight line. For example, a child bounces a ball in one direction; this is the "law." But this is the same as though the child has nothing to do but ground. Even though the airplane is moving with a very high velocity, but at a constant v velocity, the laws look the same to the child as they do when the airplane is standing still. This is however not always principle. As we see it here we shall call it "Galilean relativity" to distinguish it from the more general analysis, called by Einstein, which we shall study later.

We have just studied the law of conservation of momentum from Newton's laws, and we could prove from here to that the speed laws that describe impacts are equivalent. But for the sake of variety, and also as an illustration of a kind of reasoning that can be used in physics in other circumstances, we say, for example, one might not know Newton's laws etc. might take a different approach. We shall consider the laws of physics and collisions from a completely different point of view. We can have a discussion on the principle of Galilean relativity, stated above, and shall end up with the law of an unbroken constant.

We shall start by assuming that we would look at us: "we are along in a certain state and move it as it would if we were standing still." Before our

pushing collision in which two bodies move and stick together, or come together and bounce apart. We shall first consider two bodies that are held together by a spring or something else and are then suddenly released and pushed by the spring or perhaps by a little explosion. In either case, the initial motion is only one direction. First, let us suppose that the two bodies are exactly the same and not symmetrical objects, and then we have a simple explosion to start them. After the explosion, one of the bodies will be moving, let us say toward the right, with a velocity v . Then it appears reasonably that the other body is moving now at the left with a velocity v , because if the objects are different in mass or in center of gravity to be preferred one so the bodies would do something that is somewhat like this. This is an illustration of a kind of thinking that is very useful in many problems you will often be brought up to the just started with the following.

The first result from our experiment is that equal objects will have equal speed. But now suppose that we have two objects made of different materials, say copper and aluminum, and we make the two masses equal. We shall now suppose that if we do the experiment, with two masses that are equal, even though the objects are not alike, the velocities will be equal. Suppose, weight object "But you know, you could do it backwards, you do not have to suppose that. You could define equal masses to mean that they have equal velocities in the experiment." We follow his suggestion and make a simple extension between the copper and a very large piece of aluminum, so heavy that the copper flies out and the aluminum in all angles. That is too much of a strain on our nerves and mind to do. Let us just take a very tiny piece, such that when we make the extension the aluminum goes flying away and the copper hardly budges. That is interesting aluminum. Evidently there is some effect there in returning to the beginning again the time in which the velocities come out equal. Very well then, let us take it around, and say that when the velocities are equal, the masses are equal. This appears to be just as difficult, and it seems reasonable that we can transform physical laws into laws of numbers. Nevertheless, there are some physical laws involved, and this necessary this definition of equal means we immediately find out of the laws of physics.

Suppose we know from the foregoing experiment that two mass velocities, v_1 and v_2 for example, of aluminum, may be equal masses, and we compare a third body, say a piece of gold, with the copper in the same manner as above, making sure that its mass is equal to the mass of the copper. If we now make the experiment between the aluminum and the gold, there is no change in experiment except the mass of the copper however, the aluminum shows that they actually are. So again, by experiment, we have found a new law, a statement of it is now probably that two masses are equal if their mass velocities measured by equal velocities in the experiment are the same as each other. (The statement does not follow to all from a similar statement of a principle regarding measurement of quantity.) From this example we learn to take more seriously what things if we are careless. It is very easy a distinction to say the masses are equal when the velocities are equal, because we may be inclined to say it is implying the same kind of equality which in turn may be a misleading about an experiment.

As a second example, suppose that A and B are found to be equal by using the experiment of the other path of registration, which gives a certain velocity, if we then use another experiment will it be true or not true that the velocities now obtained are equal? Again, at least there is nothing that can decide this question, but experiment of course that it is true. So here is another law which might be stated: If two bodies have equal masses, as measured by equal velocities at one velocity, they also have equal masses when measured at another velocity. From these examples we see that appears to be only a preliminary result involved some laws of physics.

In the development of this follows we find that as is true for equal masses have equal and opposite velocities when an explosion occurs between them. We can make another assumption in the inverse case. If two unequal objects, moving in opposite directions with equal velocities, collide and stick together by some kind of glue, then which body will carry the moving after the collision? This is again a law.

symmetrical situation, with no preference between right and left, so we assume that they stand still. We shall also suppose that any two objects of equal mass, even if the objects are made of different materials, will collide and stick together when moving with the same velocity. As opposite direction's will come to rest at the collision.

10-3 Momentum & momentum

We can readily the above assumptions experimentally: first, if two identical objects of equal mass, separated by an explosion may well move apart with the same speed, and second, if two objects of equal mass, moving toward each other with the same speed, collide and stick together they will stop. This we can do by means of a explosive, invented as follows (Fig. 10-1). When you fire at another, the bullet will commonly penetrate. So does (Fig. 10-1). We could not do experiments by firing this gun because there is not only "fatty, but, by adding a large force, we can easily get rid of friction. Any object will slide without difficulty, even on at a constant velocity, as advocated by Galileo. This is done by suspending the objects on strings. Because air has very low friction, an object glides along with practically constant velocity when there is no applied force. Thus, we use two glide blocks which have been made carefully to have the same weight, or mass (their weight was measured results), but we know the frictional drag is proportional to the mass, and we place a small explosive cap to a close cylinder between the two blocks (Fig. 10-2). When all start, the blocks fly apart at the center point of the track and stay them apart by exploding the cap with an electric spark. Why should happen? If the speeds are equal, when they fly apart, they should arrive at the ends of the trough at the same time. On reaching the ends they will both bounce back with practically equal elasticity, and re-occur together and stop at the center. When they started it is a good test; when it is actually done the result is just as we have described (Fig. 10-3).

Now, the next thing we might like to agree on is what happens in a less simple situation. Suppose we have two equal masses one moving with velocity v_1 and the other standing still, and they collide and stick - what's going to happen? There is a nice allegory when we are discussed, driving with an unknown velocity. What velocity? That is the problem. To find the answer, we make the assumption that if we ride along in a car, passing over, look the same as if we are standing still. You start with the knowledge that two equal masses moving in opposite directions with equal speeds, will stop dead when they collide. Now suppose you're with the humans, you are riding in an automobile at a velocity v_1 . What total does it look like? Since we are riding along a lane of the two masses which are coming together, don't you appear to me to have zero velocity. The other mass, however, going past us with velocity v_2 , will appear to be coming toward us at a velocity $-v_2$ (Fig. 10-4). Evidently, the combined velocities the collision will seem to be going by with velocity v_1 . We therefore conclude that an object with velocity v_1 , hitting an object, one at rest, will end up with velocity v_1 , or what is mathematically speaking the same. An object with velocity v_1 hitting and sticking to another, now at v_2 , produces an object moving with velocity $v_1 + v_2$. Note that if we multiply the mass and the velocity separately and add them together, we [1/2] we get the same answer as when we multiply the mass and the velocity of everything it contains. The times we do this, allows when happens when a mass of velocity v_1 and v_2 is head-on.

Exactly the same manner we can deduce what happens when equal objects having any two velocities, fire each other.

Suppose we have two equal bodies with velocities v_1 and v_2 , respectively. Each collide and stick together. What is their velocity after the collision? Again we ride by in a car which is going at velocity v_1 , so that one body appears to be at rest. The other runs appears to have a velocity $-v_2$, v_2 , and we have the same result as we had before. When it is all finished they will be moving at $(v_1 - v_2)$ with respect to the car. What then is the actual speed on the ground?



Fig. 10-1. Top view of linear air bag test.



Fig. 10-2. Sectional view of gliders with explosive detonator cylinder attachment.

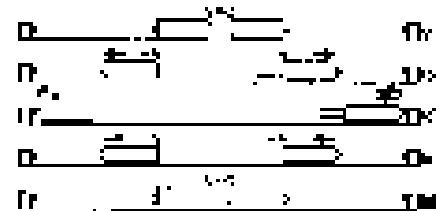


Fig. 10-3. Schematic view of collision-reaction experiments with equal masses.

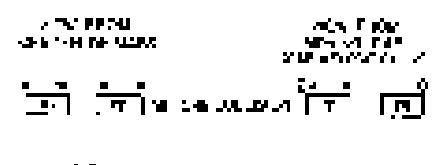


Fig. 10-4. Two views of an inelastic collision between equal masses.

$$E_{\text{kin}} + E_{\text{pot}} = m_1 v_1^2 + m_2 v_2^2 + \frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 \quad (\text{Again we neglect})$$

the potential energy due to the motion of the center of mass.





Fig. 10-6. Two views of another reaction between equal masses.

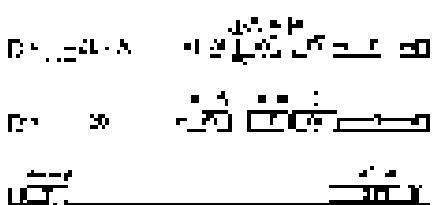


Fig. 10-6. An experiment to verify that a reaction with relatively small kinetic energy gives the same velocity $v/2$.



Fig. 10-7. Two views of another collision between m_1 and m_2 .

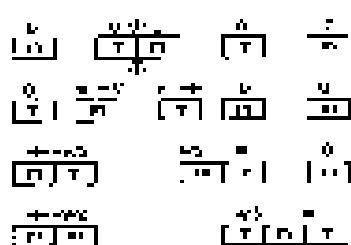


Fig. 10-8. Another one reaction between m_1 and m_2 .

$$m_1 v_1 + m_2 v_2 = m_1 v_1' + m_2 v_2' \quad (10-9)$$

Thus, using this principle, we can analyze any kind of collision to which two types of objects meet, hit each other, and stick. To see it through we have worked many of the collisions, we conducted quite a few experiments, made some simplified calculations by assuming that the collision happened in one dimension. The principle is the same, but the details get somewhat complicated.

In order to test experimentally whether an object moving with velocity v collides with 2 unequal masses m_1 and m_2 , let us consider a reaction between two objects of mass m_1 and m_2 moving with equal initial velocities $v/2$. We place in the trough two equally massive objects, two of which are initially joined together with their respective initial velocities, the third being very near to but still slightly separated from these and positioned such as nicely to approach but not yet stick to either object when hit. Now, a moment after the collision, we have two objects of mass m_1 moving with equal and opposite velocities v . A moment after that, one of these collides with the third object and makes an angle of θ with the incoming, so we calculate with velocity $v/2$. How do we test whether it is really $v/2$? By comparing the initial positions of the masses on the trough so that the distances to the ends are not equal, but are in the ratio 2:1. Using our tape, which continues to move with velocity v , we can now travel a much distance to a given position on the trough and stop by pulling the string to the smallest slot available by the second object before it collides with the third. The mass m_1 and the mass $2m_1$ should reach the end at the same time, and when we try it we find that they do (Fig. 10-9).

The next problem that we want to work out is what happens if we have two different masses. Let us take a mass m_1 and a mass $2m_1$ and apply our exclusive interaction. What will happen then? In a situation of the collision, we know m_1 releases m_2 to the left; what velocity does m_2 attain? The experiment we have just done may be repeated with an experiment between the second and third masses and when m_2 moves, we get the same result, namely, the resulting masses m_1 and $2m_1$ attain velocities $-v$ and $v/2$. Thus the direct reaction between m_1 and $2m_1$ gives the same result as the symmetric collision between m_1 and m_2 because the sum between m_1 and $2m_1$ is a third mass m_3 which they stick together. Furthermore, we find that m_2 moves toward $2m_1$ starting from the end of the trough, but when velocities (simply) inversely reversed stop and if they stick together.

Now the new question we may ask is this: What will happen if a mass m_1 with velocity v , say, hits an object that is at rest, $2m_1$, at rest? This is very easy to answer using our principle of collision reciprocity, for we simply note that when m_1 and m_2 have just separated from each other with velocity $-v/2$ (Fig. 10-9). From the law of velocities we:

$$v' = v - \text{velocity of } m_2 = v - v/2 = v/2$$

and

$$v_2' = -v/2 - \text{velocity of } m_1 = -v/2 + v/2 = 0.$$

After the collision, the mass m_1 appears to us to be moving with velocity $v/2$, thus we have the law (i.e., the law of velocities before and after collision is $v_1' + v_2' = 0$) of conservation of momentum, with a rather simple proof since m_1 is the only thing moving on, stuck together, with a velocity $v/2$ as much. The general consequence is that the sum of the velocities of the masses and the velocities stays the same, and this is because $v/2$, or m_1 , is simply holding up the "center of the conservation of momentum," piece by piece.

Now we have one last item. Using the same arguments, we can repeat the result of the previous three, we ignore them, i.e., the case of m_1 against three unequal mass rect as shown in Fig. 10-9.

In every case we find that the mass of the first object times its velocity, plus the mass of the second object times its velocity, is equal to the total mass of the final object times its velocity. These are all examples that, of course, the conservation of

momentum. Starting from simple astronomical cases, we have demonstrated that for most simple cases, we can in fact do it for any rational mass ratio, and since every case is exceedingly close to a rational ratio, we can banish errors due to irrationality as well.

III-4 Momentum and energy

All the foregoing examples are simple cases where two bodies collide and stick together, or were initially stuck together and later separated by an explosion. However, there are collisions in which the bodies do not stick, as, for example, two bodies of equal mass often collide with equal speeds and then rebound. For a brief moment, they are in contact and then are separated. At the instant of maximum compression they both have zero velocity and energy is stored in the elastic bodies, as in a compressed spring. This storage is derived from the kinetic energy that had been lost before the collision, which however comes at the instant their velocity is zero. The loss of kinetic energy is only temporary, however. The unbalanced condition is analogous to the cap that releases energy in an explosion. The body is now "unstable" decomposes in a kind of explosion, and fly apart again; but we already know that this—the bodies may appear with equal speeds. However, this speed C , should be less, if possible, i.e., the initial speed, because until "kinetic energy" is available for the explosion, depending on $m_1 = m_2$. If the material is purely no kinetic energy is recovered. But if it is something more rigid, some kinetic energy is usually regained. In the collision the rest of the kinetic energy is transformed into heat and vibration, energy—the bodies are hot and stirring. The vibrational energy also is sent transferred to us. It is possible to make the colliding bodies from highly elastic materials, such as steel, with carefully designed spring numbers, so that the collision passes the very little heat and vibration. In these circumstances the velocities of rebound are practically equal to the initial velocities and a collision is called elastic.

That the velocities before and after an elastic collision are equal is not a matter of conservation of momentum, but a matter of conservation of kinetic energy. That the speeds of the bodies remaining after a non-elastic collision are equal to each other, however, is a matter of conservation of momentum.

We might similarly consider collisions between bodies of different masses, different initial velocities, and compare degrees of elasticity, and determine final velocities and the loss of kinetic energy, but we shall not go into the details of these processes.

There collisions are especially interesting for systems that have no natural "years, months, or years." Even when there is a collision there is no rules for the change to be implemented, because the objects that move apart can in the same conditions as when they collided. Therefore, between very elementary objects, the collisions are always elastic or very nearly elastic. For instance, the collisions between atoms molecules or protons would be perfectly elastic. Although this is an excellent approximation, even such collisions are not perfectly elastic otherwise one could not break hand bone energy to the form of light or heat emission could come out of a pair. This is while, in a collision, a large energy is lost if it is emitted, but this occurrence is very rare and the energy emitted is very small. So for most purposes collisions of molecules or atoms can be considered as perfectly elastic.

As an interesting example, let us consider an elastic collision between two objects of equal mass. If they move together with the same speed, they would come apart with same speed, by symmetry. Let now look at this in another circumstance, in which one of them is moving with velocity v and the other one is at rest. What happens? We have been through this before. We watch the system and collision from a car moving along with one of the objects, and we find that if a stationary body is struck elastically by another body of exactly the same mass, the moving body stops and the one that was standing still now moves away with the same speed that the other one had, the bodies simply exchange velocities. This is what is usually demonstrated with a billiard ball apparatus. More

generally, if both bodies are moving with different velocities, they simply exchange velocity or impact.

Another example of an almost elastic interaction is magnetism. If we arrange a pair of U-shaped magnets in our glide blocks, so that they repel each other, when one glide goes up against the other, it pushes it up too, and the glide falls, and now the other goes along. Remarkably,

The principle of conservation of momentum is very useful, because it enables us to solve many problems without knowing the details. We can not know the details of the gas reactions in the cup explosion, yet we could predict the velocities with which the bodies came apart, for example. Another interesting example is rocket propulsion. A rocket of mass M , carries a small amount of mass m with constant velocity V relative to the rocket. After this is thrown, it moves originally steadily still, with the same velocity V . Using the principle of conservation of momentum, we can calculate this velocity to be:

$$v = \frac{m}{M} \cdot V$$

So long as m/M is less than one, the rocket continues to pick up speed. Rocket propulsion is essentially the same as the result of a gun: there is no need for any air to push against.

10.5 Relativistic momentum

In mechanics the law of conservation of momentum has undergone certain modifications. However, this is not Γ , the today's mechanics, but only mainly in the definition of things. In the theory of relativity it turns out that we do have conservation of momentum. The particle has mass and the momentum is still given by $m v$, the mass times the velocity, but the mass changes with the velocity, hence the momentum also changes. The mass varies with velocity according to the law

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}}. \quad (10.7)$$

where m_0 is the mass of the body at rest and c is the speed of light. It is easy to see from the formula that there is negligible difference between m and m_0 , unless v is very large, so that in ordinary velocities the correction for momentum reduces to the old formula.

The components of momentum for a single particle are written as

$$p_x = \frac{m v_x}{\sqrt{1 - v^2/c^2}}, \quad p_y = \frac{m v_y}{\sqrt{1 - v^2/c^2}}, \quad p_z = \frac{m v_z}{\sqrt{1 - v^2/c^2}}, \quad (10.8)$$

where $v^2 = v_x^2 + v_y^2 + v_z^2$. If the components are summed over all the interacting particles, both before and after a collision, the sums are equal; that is, momentum is conserved in the collisions. The same holds true in any directions.

In Chapter 1 we saw that the law of conservation of energy is an axiomatic, we recognize that energy appears in different forms, kinetic energy, mechanical energy, radiant energy, heat energy, and so on. In some of these cases, like a soap for example, the energy might be said to be "hidden." This example might suggest the question, "Are there also hidden forms of momentum — perhaps heat momentum?" The answer is that it is very hard to hide momentum for the "visible" reasons.

The sufficient condition of the nature of a body (variable measure of heat energy, if the square of the velocities are summed). This sum will be a positive result having no direction, etc. later. The fact is there, whether or not the body moves as a whole, and conservation of energy is the form of this conservation. On the other hand, if one sums the velocities, which carry direction, and finds a result that is not zero, that means that there is a drift of the entire body in some particular direction, and such a gross momentum is readily observed. Thus there is no reason to hide momentum, because the body has "no memory" not only in F

when it moves to a where. This form momentum, as a "mechanical" quantity, is without a body. Nevertheless, momentum can be added—in the Electromagnetic Field, for example. This is another effect of relativity.

One of the propositions of Newton was that if a charge in a chargeless environment—let us say our car is not the case; in situations involving electrical fields, for instance if an electrical charge in one location is suddenly moved, interactions on another charge, at another place, do not appear instantaneously. There is a time delay. In short, characteristics, such as the forces are equal the momentum will not react until there will be a short time during which the car will be trouble, because first while the first charge will feel a certain reaction force, and will pick up some momentum, but the second charge has not yet come and has not yet changed its momentum. It takes time for the influence to cross the intervening distance, where it goes at 300,000 miles a second. In that case the change from a charge to another is not experienced. Of course if the second charge has felt the effect of the first charge's motion then the momentum is then still check out—right, but during that short interval, momentum is not conserved. We say, well, it is by saying that during this interval there is another kind of momentum besides that of the particle, m_1 , m_2 . That is momentum in the electromagnetic field. If we add the field momentum to the momentum of the particles, then momentum is conserved at any moment of time. The fact that the electromagnetic field can possess momentum and energy makes that field very real, and so the law of relativity, described here that there are just the forces between particles has to be modified so that the two particles make a field, and a field makes another particle, and the field itself has such a similar properties of energy and momentum just as particles can have. To take another example: an electromagnetic field has waves, which we call light; it turns out that light also has the momentum well, so when light interacts on an object it carries in a definite amount of momentum per second. This is equivalent to a force, because if the immovable object is picking up a certain amount of momentum per second, is momentum, by definition and the situation is exactly the same as if there were a force on it. Light can exert pressure by bombarding an object. This pressure is very small, but with sufficiently intense appurtenance it is measurable.

Now in quantum mechanics it turns out that momentum is a different thing—it is no longer true. It is hard to define exactly what is meant by the velocity of a particle, but momentum still exists. In quantum mechanics the difference is that when the particles are represented as waves, the momentum is measured by the number of wave-packets per meter. The greater the number of waves, the greater the momentum, in spite of the differences. The use of construction of waves can help us to understand momentum. Even though the car—thing is false, and all the implications of Newton were wrong, the conservation of momentum, a quantum mechanics nevertheless, is true and the particle has momentum itself.

Werner

11.1 Symmetry in physics

In this chapter we introduce a subject that is technically known in physics as *symmetry in physical laws*. The word "symmetry" is used here with a special meaning, and therefore needs to be defined. What is a *symmetrical law*? In a definition like this we have to picture the law as symmetrical, and this is somehow to go on the other side. Professor Hermann Weyl has given this definition of symmetry: a thing is symmetrical if one can submit it to a certain operation, and it appears exactly the same after the operation. For instance, if we look at a view that is left-right-symmetrical, then turn it 180° around the vertical axis, it looks the same. We do not expect the definition to be similar in Weyl's more general form, and in that form we shall discuss symmetry of physical laws.

Suppose we built a complex machine in a certain place, with a lot of complicated internal parts, and with forces to move them, and so on. Now suppose we built exactly the same kind of equipment at some other place, matching part by part, with the same conditions and the same acceleration, everything else being only displaced laterally by some distance. Then, if we put the two machines in the same initial circumstances, in exact correspondence, we ask: Will one machine behave exactly the same as the other? Will it follow all the motions in exact parallelism? Of course, the answer may well be no, because if we choose the wrong place for our machine it might be inside a wall and influences from the wall would make the machine not work.

All of our laws in physics assume a certain amount of symmetry even in their application; they are not purely mathematical or abstract ideas. We have to understand what we mean when we say that the phenomenon is the same as in the case the apparatus is in a new position. We mean that we move everything that we believe to be relevant, if the phenomenon is not the same, we suggest that something important has not been moved, and we proceed to look for it. Now, however, if there are claim that the laws of physics do not have this symmetry. On the other hand, we have said it: we expect to find it, if the laws of physics do have this symmetry: looking around, we may discover, for instance, that the wall is pushing on the apparatus. The basic question is, if we define things well enough, if >1 the bounded forces are included inside the apparatus, will the forces not be moved from one place to another, will they keep the same? Will the machinery work the same way?

It is clear that what we want to do is to move all the equipment and external influences, but for simplicity let us consider, first, and all—let us if we do this, we have the same phenomena again for the trivial reason that we are right back where we started. Now, we cannot move everything. That is true, and in practice that with a certain amount of intelligence about where to move, the machinery will work. In other words, if we do not go mad about it, if we know the origin of the outside forces, and arrange that these are, moreover, then the machine will work the same in one location as in another.

11.2 Translations

We shall limit ourselves to just one law, for which we now have sufficient knowledge. In previous chapters we have seen the translation of variables and the sum related by a set of three equations for each particle:

$$\pi_0(\theta^0_{\mu}/\theta^{00}) = p_\mu \quad \pi_0(\theta^0_{\nu}/\theta^{00}) = p_\nu \quad \pi_0(\theta^0_{\lambda}/\theta^{00}) = p_\lambda \quad (11.1)$$

11.1 Symmetries in physics

11.2 Translations

11.3 Rotations

11.4 Forces

11.5 Factor algebras

11.6 Newton's laws in vector notation

11.7 Scalar product of vectors

Now this means that there exists a way to measure x_1 , y_1 , and z on three perpendicular axes, and the force along these directions will be the same and one doesn't need to measure from some origin, but where should we origin? Well, Joe would tell us at least it's OIC; there is some place that we can measure from, perhaps because of the universe, such that these forces are constant. But we can also immediately that we can never find the center, because if we use some other origin, it would make an difference. To take notes, anyone that there are two people, say, who has an origin \vec{r} in one place, and Max, who has a parallel system where origin is somewhere else (the \vec{r}'). Now when Joe measures the distance of \vec{r} from \vec{r}' he finds it to be a and it's well-knownly known that, because it is the defining of a (area of a parallelogram). After one a for \vec{r} , when measuring the same point \vec{r}' he finds a different a (in order to distinguish it). We will call it a' , and in principle it's different a . Though in our example they are numerically equal, so we can

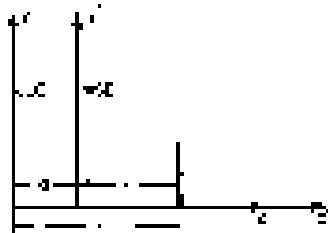


Fig. 11-1. Two parallel coordinate systems.

Now in order to complete our analysis we must know what Max would obtain for the forces. The force is supposed to be along parallel to and by the form in the introduction or mean the part of the total which is in the x -direction, which is the magnitude of the force times the cosine of its angle with the x -axis. Now we see that Max would get exactly the same projection as Joe did just, as we have a set of equations:

$$F_x = F_{x_1} - F_{x_2}, \quad F_y = F_{y_1} - F_{y_2}, \quad (11.1)$$

Let's look at the relationships between these like as given by the first line.

The question is, if Joe knows Newton's laws, and it has to be with when Newton's laws, and they didn't know Farhi's? Then it makes no difference from where origin we measure the points? In other words, assuming that equations (11.1) are true and the Eqs. (11.2) and (11.3) give the relationship of the components, is it or is it not true that

$$\begin{aligned} (11.2) \quad m_1(x_1^2 + y_1^2)^{1/2} &= F_{x_1}, \\ (11.3) \quad m_2(x_2^2 + y_2^2)^{1/2} &= F_{x_2}, \\ (11.4) \quad m_1(x_1^2 + y_1^2)^{1/2} &= F_{x_1}, \end{aligned}$$

In order to use these equations we shall calculate the formula for x_1 twice. First of all

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}}, \quad x = \dot{x}t, \quad \frac{dx}{dt} = \frac{dx}{dt}.$$

Now we still assume that Max's origin is fixed (not moving) relative to Joe's; therefore x is a constant and $dx/dt = 0$, we will find the

$$dx/dt = dx/dt'$$

and therefore:

$$d^2x/dt^2 = d^2x/dt'^2,$$

therefore we know that Eq. (11.4a) becomes

$$m_1(d^2x/dt^2)^{1/2} = F_{x_1},$$

(We also suppose that forces measured by Joe and Max are equal.) Thus the acceleration times the mass is the same as the other follows. We have also found the formula $F_x = F_{x_1}$, i.e., substituting from Eq. (11.1), we find that

$$F_x = F_{x_1}.$$

The other two forces will be Max appear the same; he can prove Newton's law but with different constants m and they will still be right. That proves that (11.4)

There is no unique way to define the origin of the world; because the laws will appear the same, from one user position, they are observer.

This is also true: if there is a piece of equipment in one place and a certain kind of machinery in it, the same equipment in another place will behave in the same way. Why? Because one machine, when analyzed by Joe, has exactly the same equations as the other one, analyzed by Joe. Since the equations are the same, the phenomenon happen. That is, the behavior of equipment in a new position behaves the same as it did in the old position, as the proof that the equations when replaced in space does not change themselves. Therefore we say that the law of physics are invariant for translational displacements, symmetrical in the sense that the laws do not change when we make a translation of our coordinates. Of course it is quite obvious intuitively. That is to say, it is interesting and enlightening to discuss the mathematics of it.

11-1 Rotations

The above is the first of a series of ever more complicated propositions concerning the symmetry of a physical law. The next proposition is: let us again make no difference in where we choose the axes. In other words, if we build a piece of equipment in some place and we do it properly, and nearby we build the same kind of apparatus but turn it up at an angle, will it operate in the same way? Obviously it will not if it is a Grandfather clock, for example. If a pendulum clock stands upright, it works fine, but if it is tilted, the pendulum falls against one side of the case and nothing happens. The theorem is even false in the case of the pendulum clock, unless we include the earth, which is pulling on the pendulum. Therefore we can make a proposition about pendulum clocks if we believe in the symmetry of physical law for rotation: something else is involved in the operation of a pendulum clock besides the mechanics of the clock, something outside it that we should look for. We may say perhaps no pendulum in space will not work the same way when located at different places relative to the mysterious source of gravitation, perhaps the earth. Indeed, we know that a pendulum clock will not work in space. For example, would not tick either, because there is no effective force, and on Mars it would go at a different rate. Pendulum clocks do involve something more: but just the mechanics itself may involve something on the outside. Once we recognize this factor, we see that we must turn the earth along with the apparatus. Otherwise we do not have to worry about Earth, it's best to do the one simply with a moment of rotation, so you can't bring into the problem, does now again in the new position the same as it did before. When we are rotating in space our bodies, a change changing, obviously. To change does not seem to be necessary more, but in the new position we seem to be in the same condition as in the old. This has a certain tendency to continue one, because it is true that in the new initial position the laws are the same as in the original position, but it is not true that we can a song to change the same laws as it does when we are not carrying it. If we perform sufficiently delicate experiments, we can tell that the earth is rotating, but not how it is going around. In other words, we cannot locate the angular position, but we can tell that it is changing.

Now we must discuss the effect of angular orientation upon physical laws. Let us find out who has the same symmetries with Joe and Mrs. Works-ager. This time, to avoid needless complication, we shall suppose that Joe and Mrs. Mac had the same initial position (we have already discussed). However, he moved by translation in another plane. Assume that Mrs.'s new axes rotated relative to Joe's by an angle θ . Two coordinate systems are shown in Fig. 11-2, which is restricted to two dimensions. Consider any point P having coordinates (x, y) in Joe's system and (x', y') in Mrs.'s system. We shall begin, as in the previous case, by expressing the coordinates x' and y' in terms of x , y , and θ . To do this we first drop perpendiculars from P to all four axes and draw all perpendiculars to xy . Inspection of the figure shows that x' can be written as the sum of two lengths along the x axis and y' as the difference of two longitudes along xk . All these lengths are expressed

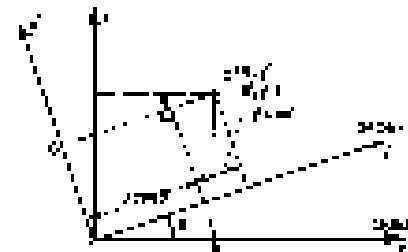


Fig. 11-2. Two coordinate systems having different angular orientation.

in terms of α , β , and δ in equation (11.5), at which we have added an equation for the third dimension:

$$\begin{aligned} x' &= x \cos \theta + z \sin \theta \\ y' &= y \cos \theta - z \sin \theta \\ z' &= z \end{aligned} \quad (11.5)$$

The next step is to analyze the relationship of forces as seen by the two observers, choosing the same generalized coordinates as before, i.e. x and z , which are already been analyzed as having components F_x and F_z (as seen by \mathbf{O}), by setting up a parallel of motion, located at point P in Fig. 11-2. For simplicity, let us move both sets of axes so that the origin is at P , as shown in Fig. 11-3. This sees the components of \mathbf{F} along his axes as $F_{x'}$ and $F_{z'}$. F_x has components along both the x' and z' axes, and F_z likewise has components along both these axes. To express F_x in terms of $F_{x'}$ and $F_{z'}$, we shall then consider F_x along the x axis, and in a like manner we can express F_y in terms of $F_{x'}$ and $F_{z'}$. The results are

$$\begin{aligned} F_x &= F_{x'} \cos \beta - F_{z'} \sin \beta \\ F_y &= F_{x'} \sin \beta - F_{z'} \cos \beta \\ F_z &= F_{z'} \end{aligned} \quad (11.6)$$

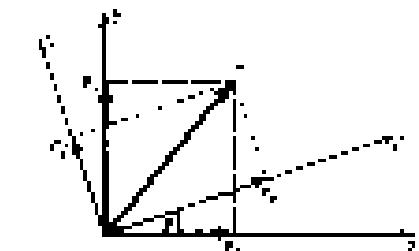


Fig. 11-3. Components of a force in the two systems.

It is interesting to note an accident of sorts, which is of extreme importance: the formulas (11.5) and (11.6) for components of \mathbf{F} and components of \mathbf{F}_x , respectively, are of identical type.

As before, Newton's laws are assumed to be true in both systems, and are repeated in equations (11.7). The question, again, is whether \mathbf{F} can satisfy Newton's laws—will the results be correct for the system of rotated axes? In other words, if we compare (1) Eqs. (11.5) and (11.6) give the right magnitude of the components, (2) is it of right direction?

$$\begin{aligned} m(\ddot{x}'^2/\partial t^2) &= F_{x'} \\ m(\ddot{y}'^2/\partial t^2) &= F_y \\ m(\ddot{z}'^2/\partial t^2) &= F_z \end{aligned} \quad (11.7)$$

To do this comparison, we calculate the left-hand sides independently, and compare the results. To calculate the left sides, we multiply equations (11.5) by m , \ddot{z} , and the product twice with respect to time, assuming the angle β to be constant. This gives

$$\begin{aligned} m(\ddot{x}'^2/\partial t^2) &= m(\dot{x}^2/\partial t^2)\cos^2 \beta - m(\dot{z}^2/\partial t^2)\sin^2 \beta \\ m(\ddot{y}'^2/\partial t^2) &= m(\dot{x}^2/\partial t^2)\cos^2 \beta - m(\dot{z}^2/\partial t^2)\sin^2 \beta \\ m(\ddot{z}'^2/\partial t^2) &= m(\dot{z}^2/\partial t^2) \end{aligned} \quad (11.8)$$

We take out the right sides of equations (11.7) by multiplying equation (11.7) into equations (11.6). This gives

$$\begin{aligned} F_{x'} &= m(\dot{x}^2/\partial t^2)\cos^2 \beta - m(\dot{z}^2/\partial t^2)\sin^2 \beta \\ F_y &= m(\dot{x}^2/\partial t^2)\sin^2 \beta - m(\dot{z}^2/\partial t^2)\cos^2 \beta \\ F_z &= m(\dot{z}^2/\partial t^2) \end{aligned} \quad (11.9)$$

Remark: The left sides of Eqs. (11.8) and (11.9) are identical, so we conclude that if Newton's laws are correct in one set of axes, they are also valid in another set of axes. This result, which has long been established for bivariate kinematics of axes, however, is not surprising. That means that all axes in a particular space are unique, but of course they can be more convenient for certain practical problems. For example, it is useful to have gravity along one axis, but this is not physically necessary. Second, a massive thin ring piece of equipment which is completely self-contained, with all the forces generating its motion completely inside the enclosure, would make the same when turned on its side.

11-4 Vectors

Newton's second law, because the other laws of physics so far we know today have the two properties which we call *invariance* (or *covariance*) under translation of coordinates and of time. These properties are so important that a mathematical technique has been developed to take advantage of them by writing and using physical laws.

The foregoing may yet appear considerably “weird” mathematical work. To test out the details of it in more detail, let us consider questions 2 and 3, more difficult mathematics machinery has been devised. This system, called “vector analysis,” simplifies the task of this chapter; strictly speaking, however, this is a chapter on the symmetry of physical laws. By the methods of the vector analysis we want to do everything required for achieving the results that we sought, but in practice we should like to do things more easily and rapidly, so we discuss the vector techniques.

We begin by noting some characteristics of two kinds of quantities that are important in physics. (Actually there are more than two, but we'll start out with two.) One of them, like the number of people in a flock, we call an *intensive quantity*, or an *independent quantity*, or a *scalar*. Temperature is an example of such a quantity. Other quantities that are independent of places do take a value, for instance velocity; we have a step back of which way a bird is going, not just its speed. Measuring the force also over direction, across displacement when, for example, going from one place to another in space, we can take track of time between, even if we only also in time where it went, we need to specify a coordinate.

All quantities that carry a dimension, like a step in space, are called *vectors*.

A vector is given by numbers. In order to represent a step in space say from the origin to some particular point, b , whose location is (x, y, z) , we really need three numbers, but we are going to invent a single mathematical symbol, \mathbf{v} , which is to be any other mathematical symbol we have on hand.¹ It is not a single number, it is systematically three numbers (x, y, z) . The three free numbers \mathbf{v} not really only three numbers, because it is better to use a different coordinate system. In these numbers \mathbf{v} could be changed to (x', y', z') . However, we must be able to our mathematics to jump around so we are going to make the same move to represent \mathbf{v} , the three numbers (x, y, z) and the three numbers (x', y', z') . Last is, we use the same move to represent the first set of three numbers for one coordinate system, into the second set of three numbers if we are going to other coordinate system. This has the advantage that when we change the coordinate system, we do not have to change the letters of our equations. That is, if we equate \mathbf{v} to (x, y, z) , and then use another system, we have to change to (x', y', z') . We shall just write \mathbf{v} , with the understanding that it represents (x, y, z) if we use one set of axes, or (x', y', z') if we use another set of axes, and so on. The three numbers which describe the quantity in a given coordinate system are called the *components* of the vector in a direction of the coordinate axes of that system. That is, we use the same symbol for the three letters that is a symbol for the same object, or even from different ways. The way that we can say “ \mathbf{v} is a quantity,” implies a physical reality and the reality of “ \mathbf{v} ,” x , y , z , x' , y' , z' depend on the coefficients in terms of which we measure \mathbf{v} , the symbol \mathbf{v} will represent the something no matter how we put the axes.

Now suppose there is another direction of pointing \mathbf{v} , say, any other coordinate system, also three numbers associated with \mathbf{v} , like (x_1, y_1, z_1) . Then these numbers change to three other numbers by a certain, mathematically true, as we change the axes. It must be the same rule that relates (x, y, z) to (x_1, y_1, z_1) . In other words, any physical quantity agrees with transformations which transform us to the components of a step in space is a vector. An equation like

$$\mathbf{F} = \mathbf{v}$$

would thus be true in any coordinate system if it were true in one. This is what

¹ In type code this is represented by boldface characters but I don't know if it works?

of course, counts for the other components.

$$F_x = a, \quad F_y = b, \quad F_z = c.$$

or, alternatively, for

$$F_x = a', \quad F_y = b', \quad F_z = c'.$$

The fact that a physical relationship can be expressed as a vector equation tells us the relationship is unchanged by a mere rotation of the coordinate system. That is, the reason why vectors are so useful in physics.

Now let us examine some of the properties of vectors. As examples of vectors we may mention velocity, momentum, force, and acceleration. For many purposes it is convenient to represent a vector quantity by an arrow that indicates the direction in which it is going. Why can we represent that way? Because it has the static mechanical transformation properties of a "step in space." We then represent it as a diagram as if it were a step, using a scale such that one unit of force, or one Newton, corresponds to a certain convenient length. Once we have done this, a force can be represented as length, because an equation like

$$\mathbf{F} = k\mathbf{r},$$

where k is some constant, is a perfectly legitimate equation. Thus we can always represent forces by lines, which is very convenient, however, since we've drawn the line we no longer need the axes. Of course, we can quickly calculate the three components as they change from time to time. However, because force is not a covariant quantity,

1.1-4 Vector algebra

Now we must describe the laws, or rules, for combining vectors to form new ones. The first such combination is the addition of two vector quantities. But \mathbf{a} is a vector object in some particular coordinate system, but the three components (a_x, a_y, a_z) , and that is another vector which has the three components (b_x, b_y, b_z) . Now let us invent three new numbers, $a_1 = b_x, a_2 = b_y, a_3 = b_z$. Do these form a vector? "Well," we might say, "they are three numbers, and every three numbers form a vector!" No, not every three numbers form a vector! In order for it to be a vector, not only must there be three numbers, but these must be associated with a coordinate system in such a way that if we turn the coordinate system, the three numbers "relocate" on each other, get "mixed up" in each other, by the precise laws we have already described. So the question is, "Can we make the coordinate system so that (a_x, a_y, a_z) become (a_1, a_2, a_3) and (b_x, b_y, b_z) become (b_1, b_2, b_3) ?" What do $(a_1 + b_1, a_2 + b_2, a_3 + b_3)$ become? Do they become $(a_x + b_x, a_y + b_y, a_z + b_z)$ or not? This exercise, of course, you remember the coordinate transformations of Eq. (11.2) except that we call it a *Möbius transformation*. If we apply these transformations to a_1, a_2, a_3 and b_1, b_2, b_3 to get $a_1 + b_1$, we find that the transformed $a_1 + b_1$ is *indeed* the same as $a_x + b_x$. When a and b are "added together" in this sense, they will form a vector which we may call c . We could write this as

$$\mathbf{c} = \mathbf{a} + \mathbf{b}$$

Now comes the interesting property:

$$\mathbf{c} = \mathbf{b} + \mathbf{a}.$$

so we can add vectors one at a time in components. That is,

$$\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}.$$

We can add vectors in any order.

What is the geometric significance of $\mathbf{a} + \mathbf{b}^2$? Suppose that \mathbf{a} and \mathbf{b} was represented by lines on a sheet of paper, what would it look like? This is shown in Fig.

Fig. 11-4. We see that we can add the components of \mathbf{b} to those of \mathbf{a} and conveniently place \mathbf{c} in such a way as to represent the components of \mathbf{b} as $-\mathbf{b}$, thus representing the components of \mathbf{a} in the system formed by \mathbf{b} . Since \mathbf{b} is just "this" force, really, it has a true meaning; this is done merely by putting the "tail" of \mathbf{b} at the "head" of \mathbf{a} , the arrow from the "tail" of \mathbf{a} to the "head" of \mathbf{b} being the vector \mathbf{c} . Of course, if we add \mathbf{a} to \mathbf{b} the other way around, we would get the same result as in the "tail" of \mathbf{b} by the parallelogram process of Fig. 11-2, yet the same result does. Note that vectors can be added in this way without changing their magnitudes.

Suppose we multiply a vector by a number n , what does this mean? We define n to mean a new vector whose components are $n a_x$, $n a_y$, and $n a_z$. We have no problem for horizontal vectors, but in 3D?

Now let us consider vector subtraction. We may define subtraction in the same way as addition, but instead of adding, we subtract the a portions. So we might define subvector by defining a new vector $\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b})$ and then we would add the components. It comes to the same thing. The result is shown in Fig. 11-5. This figure shows $\mathbf{d} = \mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b})$; we can see that the difference $\mathbf{a} - \mathbf{b}$ is to be found very easily from \mathbf{a} and \mathbf{b} by using the commutative relation $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$. Thus the difference is even easier to find than the sum $\mathbf{a} + \mathbf{b}$, just like the vector sum $\mathbf{a} + \mathbf{b}$, in Fig. 11-2!

Now we discuss velocity. "Velocity is a vector." Position is given by the three coordinates (x, y, z) , what is the velocity? The velocity is given by dx/dt , dy/dt , and dz/dt is it a vector or not? We can find out by differentiating the expression of Eq. (11-1) to see if v is a vector in the right way. We see that the components dx/dt and dy/dt do transform according to the same laws as x and y , and therefore the "velocities" dx/dt and dy/dt are velocity vectors. We can write the velocity in an interesting way as

$$\mathbf{v} = \dot{x} \mathbf{i} + \dot{y} \mathbf{j} + \dot{z} \mathbf{k}.$$

What the velocities \dot{x} , \dot{y} , and \dot{z} are is a vector, can be understood more plausibly. How far does a particle move in a short time Δt ? Answer: Δr , so if a particle is "there" at one instant and "there" at another instant, then the vector difference in the position $\Delta \mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1$, which is in the direction of motion. Given in Fig. 11-6, suppose by the time interval $\Delta t = t_2 - t_1$ is the "average velocity" vector

In other words, by vector velocity we mean the limit, as Δt goes to 0, of the displacement over the time interval Δt , i.e., $\Delta \mathbf{r}/\Delta t$. In terms of \mathbf{r} defined by (11-1)

$$\mathbf{v} = \lim_{\Delta t \rightarrow 0} \frac{\Delta \mathbf{r}}{\Delta t} = \frac{d\mathbf{r}}{dt} = \dot{x} \mathbf{i} + \dot{y} \mathbf{j} + \dot{z} \mathbf{k}. \quad (11-10)$$

This velocity is a vector because it is the difference of two vectors. It is also the definition of velocity because its components \dot{x} , \dot{y} , and \dot{z} are, in fact, dx/dt , dy/dt , and dz/dt . To sum this argument: that if we differentiate any vector with respect to time we produce a new vector. So we have several ways of obtaining new vectors: (1) multiply by a constant, (2) difference of two vectors with respect to time, (3) add two vectors.

11-6 Newton's Laws in vector form

In order to write Newton's laws in vector form, we have to project our x -axis, y -axis, and z -axis on the center of mass. This is the "inertial frame" for velocity vector, and it is easy to demonstrate that its components are the second derivatives of x , y , and z with respect to time:

$$\mathbf{a} = \frac{d^2 \mathbf{r}}{dt^2} = \left(\frac{d^2}{dt^2} (x) \right) \mathbf{i} + \left(\frac{d^2}{dt^2} (y) \right) \mathbf{j} + \left(\frac{d^2}{dt^2} (z) \right) \mathbf{k}, \quad (11-11)$$

$$a_x = \frac{d^2 x}{dt^2} = \ddot{x}, \quad a_y = \frac{d^2 y}{dt^2} = \ddot{y}, \quad a_z = \frac{d^2 z}{dt^2} = \ddot{z}, \quad \mathbf{a} = \ddot{x} \mathbf{i} + \ddot{y} \mathbf{j} + \ddot{z} \mathbf{k}. \quad (11-12)$$

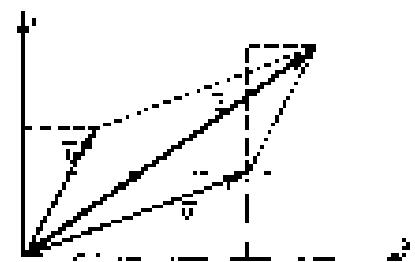


Fig. 11-4. The addition of vectors.

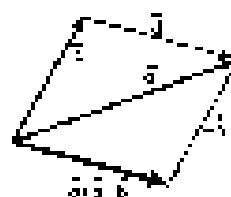


Fig. 11-5. The subtraction of vectors.

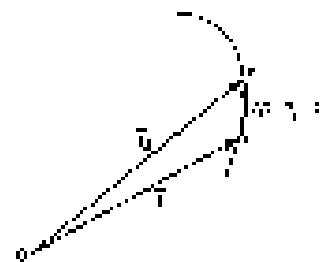


Fig. 11-6. The displacement of a particle in a short time interval $\Delta t = t_2 - t_1$.

With this definition, then, Newton's law can be written in this way:

$$m \ddot{\mathbf{r}} = \mathbf{F} \quad (11.1)$$

$$\text{or} \quad m\ddot{\mathbf{r}}(t)/t^2 = \mathbf{T}. \quad (11.1a)$$

Now the problem of deriving the equations of Newton's laws under rotation, or curvilinear motion, is reduced to taking vector differentiation just there. Since then \mathbf{F} is a vector, $m\ddot{\mathbf{r}}$ is also a vector; so if \mathbf{r} is a vector, then, since we know acceleration is a vector, Eq. (11.1a) will hold in the same as in any coordinate system. Writing it in a form which does not explicitly mention \mathbf{r}_x , \mathbf{r}_y , and \mathbf{r}_z has the advantage that from now on we need not write three axes every time we write Newton's equations or other laws of physics. We have scalar tools like one law, but really, of course, it is the free law for any particular set of axes, because any vector equation involves the statement that each of the components is equal.

The first law, the acceleration is the rate of change of the vector velocity, helps us to calculate the acceleration in our curvilinear coordinate circumstances. Suppose, for instance, that a particle is moving on some complicated curve (Fig. 11-7) and that at a given instant, it had a certain velocity \mathbf{v}_1 , but that when we go to another instant t_2 a little later, it has a different velocity \mathbf{v}_2 . What is the answer? Answer: Acceleration is the difference in the velocity caused by the same time interval, so we need the difference of the two velocities. How do we get the difference of the velocities? To do that two ways, let Δt be the time interval between t_1 and t_2 ; then $\Delta \mathbf{v}$ is the difference of the two vectors, right? Not! That only works when the ratio of the vectors is 1 : the set is closed! It has no meaning if we move the vector around when they are then in fact not closed, since about. We have to draw a new diagram to subtract the vectors. In Fig. 11-8, \mathbf{v}_1 and \mathbf{v}_2 and $\Delta \mathbf{v}$ drawn parallel to their components in Fig. 11-7, and now we can discuss the subtraction. If you try the subtraction is simply $\Delta \mathbf{v}/\Delta t$. It is interesting to note that we can compute the velocity difference out of the $\Delta \mathbf{v}$'s, without much of a calculation, by using the angle α between \mathbf{v}_1 and \mathbf{v}_2 to the path and α_1 of right angles to the path, as indicated in Fig. 11-8. The acceleration tangent to the path is, of course, just the change in the length of the vector, i.e., the change in magnitude:

$$a_r = d|\mathbf{v}|/dt. \quad (11.1b)$$

The other component of acceleration, at right angles to the curve, is easy to calculate, using Figs. 11-7 and 11-8. In the short time Δt let the change in angle between \mathbf{v}_1 and \mathbf{v}_2 be the small angle $\Delta\alpha$. If the magnitude of the velocity is called v , then of course

$$\Delta\alpha = 1.48$$

and the acceleration, a , will be

$$a_\theta = v^2/\Delta\alpha$$

Now we need to know $v/\Delta\alpha$, which can be found this way. If, at the beginning, the curve is supposed to be a circle of a certain radius R , then in Δt the distance s is, of course, $v\Delta t$, where v is the speed,

$$\Delta s = v\sin(\Delta\alpha/R), \quad \text{or} \quad \Delta s/\Delta\alpha = v/R.$$

Therefore, we find

$$a_\theta = v^2/R, \quad (11.1c)$$

as we have seen before.

11-7 Scalar product of vectors

Now let us end this talk further the properties of vectors. It is very important that the length of a vector in space would be the same in any coordinate system, and so, if a particular step $\Delta \mathbf{r}$ is represented by $\Delta_x, \Delta_y, \dots$ in one coordinate system,

and by x', y', z' the other components of system, since the orientation of world to the same in both form

$$r = \sqrt{x^2 + y^2 + z^2}$$

we also

$$r' = \sqrt{x'^2 + y'^2 + z'^2}.$$

So what we wish to verify is that these two quantities are equal. It is much more convenient now to take the square root, so that we'll want the square of the distance from the origin:

$$x^2 + y^2 + z^2 = x'^2 + y'^2 + z'^2. \quad (11.18)$$

If had better see if we substitute Eq. (11.8) we do indeed find that it is. So we see that there are no hidden difficulties with summing the squares of the components that is, let us find out whether

Something new is involved. We can introduce a new quantity, a function of a , b , and c , to be a scalar, *spherical angle*, in which two or three vectors which is the same as been systems. But at a later we can make a scalar. We have to find a general rule to that, i.e., a scalar value. The rule is to the case we consider add the squares of the components. Let me now define a new thing, often you will call it a dot. This is not a vector, but a scalar; it is a number that is the same in all coordinate systems, and it is defined to be the sum of the squares of the three components of the vector:

$$a \cdot b = a_x^2 + a_y^2 + a_z^2. \quad (11.19)$$

Now you say "But with what does?" It does not depend on because the vector is the same in every set of axes. So we have a new kind of quantity, a new kind of scalar, called "a scalar product." If we now define the following quantity for any two vectors a and b :

$$a \cdot b = a_x b_x + a_y b_y + a_z b_z. \quad (11.20)$$

we find that this quantity, calculated in the primed and unprimed systems, also have the same. To prove it we note that it is true of a_x , b_x and c_x since $a_x = a + b_x$. Therefore the sum of the squares $(a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2$ will be invisible to

$$(a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2 = (a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2 + (b_x - b_x)^2 + (b_y - b_y)^2 + (b_z - b_z)^2. \quad (11.21)$$

Both sides of this equation are expanded, there will be cross products of just the type appearing in Eq. (11.20), as well as the sums of squares of the components of a and b . The expansion of terms of the form of Eq. (11.18) then shows there are no power terms (a , b) involved also.

The quantity $a \cdot b$ is called the scalar product of two vectors a and b , and it has many interesting and useful properties. For instance, it is easily proved that

$$a \cdot (b + c) = a \cdot b + a \cdot c. \quad (11.22)$$

Also, there is a simple geometrical way to calculate $a \cdot b$, without having to calculate the components a_x and b_x . $a \cdot b$ is the cosine of the angle θ between a and the length of b times the cosine of the angle between a and b . Why? Suppose that we choose a special coordinate system in which the x -axis lies along a ; in these coordinates, the only component of a that will be there is a_x , which is of course the whole length of a . Thus Eq. (11.18) reduces to $a \cdot b = a_x b$; for this reason, this is the length of a times the component of b in the direction of a , that is, b_x :

$$a \cdot b = ab \cos \theta.$$

Therefore, in this special coordinate system, we have proved that $a \cdot b$ is the

length it takes the longer path times (c/s) . But (c/s) is unique to one coordinate system, so it must be off, because ω is independent of the coordinate system. That is our argument.

What good is the dot product? Are there any cases in physics where we need it? Yes, we need it all the time. For instance, in Chapter 4 the kinetic energy was called $\frac{1}{2}mv^2$, but if the object is moving in space it should be the velocity squared in the direction the y -axis has, and the x -direction, and w . In formula for kinetic energy accounting for both the x and y is

$$K.E. = \frac{1}{2}m(v_x^2 + v_y^2) = \frac{1}{2}m(v_x^2 + v_y^2 + v_z^2) \quad (11.22)$$

Energy does not have dimension. Momentum has dimension kg , is a vector, and it is the mass times the velocity vector.

Another example of a dot product is the work done by a force when some thing is pushed from one place to the other. We have no yet defined work, but it is equivalent to the energy change, the weights used, when a box is pushed along a distance s :

$$\text{Work} = F \cdot s \quad (11.23)$$

It is sometimes very convenient to talk about the component of a vector in a certain direction (say the vertical direction, because that is the direction of gravity). For such purposes, it is useful to invent what we call a unit vector in the direction that we want to study. By a unit vector we mean one whose dot product with itself is equal to unity. If we call this unit vector even if $|F| = 1$ new, or we want the component of some vector \mathbf{v} in the direction of \mathbf{i} , we see that the dot product $\mathbf{v} \cdot \mathbf{i}$ will be a nice value, the component v_i of \mathbf{v} in the direction of \mathbf{i} . There is a nice way to get the component to zero; it permits us to get all the components and to write a rather amazingly simple. Suppose that in a given system of coordinates, \mathbf{i} , \mathbf{j} , and \mathbf{k} are defined. If we take \mathbf{i} as a vector in the direction of \mathbf{i} , \mathbf{j} and \mathbf{k} perpendicular to the direction of \mathbf{i} , and \mathbf{k} a unit vector in the direction of \mathbf{k} . Note that that $i \cdot i = 1$. What is $j \cdot j$? When two vectors are at right angles, their dot product is zero. Thus,

$$\begin{aligned} \mathbf{i} \cdot \mathbf{i} &= 1 \\ \mathbf{i} \cdot \mathbf{j} &= 0 \quad \mathbf{j} \cdot \mathbf{i} = 1 \\ \mathbf{i} \cdot \mathbf{k} &= 0 \quad \mathbf{j} \cdot \mathbf{k} = 0 \quad \mathbf{k} \cdot \mathbf{i} = 1 \end{aligned} \quad (11.24)$$

Now with these definitions, any vector whatever can be written this way:

$$\mathbf{v} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k} \quad (11.25)$$

By this, we go from the components of a vector to the vector itself.

The properties of vectors is by no means complete. However, rather than try to go more deeply into the subject now, we shall first learn to use it physically since it comes up in the next few discussions. Then, when we have properly mastered the basic material, we shall find it easier to perceive more clearly if we do not go without getting too confused. We shall later find that it is useful to define another kind of product of two vectors, \mathbf{a} and the vector \mathbf{b} , $a \times b$, and written as $\mathbf{a} \times \mathbf{b}$. However, we shall postpone a discussion of such matters as a little while.

Characteristics of Force

12-1 What is a force?

Although it is interesting and worth while to study the physical laws, simply because they help us to understand and use the future more and more successfully, one in a while we might ask, "What does *law* really mean?" The meaning of *law* is a concept in a subject that has interested and troubled philosophers from time immemorial, and the meaning of classical laws is even more interesting, because it is generally believed that these laws represent some kind of real knowledge. The meaning of knowledge is a deep problem in epistemology, and it is always important to ask, "What does it mean?"

Let us ask, "What is the meaning of the physical laws of Newton, which you will learn?" or "What is the meaning of force, mass, and acceleration?" Well, we can say it very easily. The meaning of mass, and we can always assume it, is if we know the meaning of position and time. We shall not discuss those meanings, but shall concentrate on the new concept of force. The answer is quite simple. "If a body is accelerating, then there is a force on it." That is what Newton's law says; so the most precise and beautiful definition of force imaginable might simply be to say that force is the cause of the mass times the acceleration. Suppose we have a law which says that the conservation of momentum is valid if the sum of all the external forces is zero; then the question arises, "What does it mean, that the sum of all the external forces is zero?" A reasonable way to define this situation would be, "When the total momentum is constant, then the sum of the external forces is zero." There must be something wrong with that, because it is just not saying anything new. If we have defined a function of time, which shows that the force is equal to the mass times the acceleration, and then define the force to be the mass times the acceleration, we have sound but nothing. We could also define force to mean the following: object with no force acting upon it can do no more with greater velocity in a straight line. If we can move an object at constant velocity in a straight line, if we can move an object at constant velocity in a straight line, then we say that there is no force on it. Now such things hardly come up in the subject of physics, because many other things are going on at once. The Newtonian programme, however, seems to be a most precise definition of force, and one that appeals to the mathematician; nevertheless, it is completely useless, because no practical experimenter can measure force by definition. One might sit in somewhere all day long and define words at will, but to find out what happens when two bodies push against each other, or when a weight is hung on a spring, is another matter altogether, because the way the bodies behave is something completely outside any device of definition.

For example, if we want to choose to say that an object left to itself keeps its position, one does not move, even when we see something drizzling, we can't say that, just like it is a "force" a grain's chance of change of position. Now we have to add in new law, everything changes if except when a grain is moving. Yes, see, that would be analogous to the above definition of force, and it would contain no information. The real content of Newton's law is this, that the law is supposed to have some predictive properties, or utility, to the best of our part, but the specific inessential properties that the force has were not completely described by Newton or by anybody else, and therefore the physical law, $F = ma$, is an incomplete law. It means that if we study the mass times the acceleration and call the product the force, i.e., if we study the derivative force as a program

12-2 What is a force?

12-2 Attractors

12-3 Multicenter forces

12-4 Fundamental forces, I

12-5 Pseudo forces

12-6 Nuclear forces

of interest. Then we shall find that forces have some simplicity, the law is a great simplification, and, in fact, it is a very simple kind of the forces we experience.

Now the next example of a set forces may be the component of gravitation which was given by Newton, and in stating the law he announced the relation, "Who is the Lord?" If there were nothing but gravity, then the consequences of this law and the forces law (second law of motion) would be a complete theory. But there is much more than gravitation, and we want to use Newton's law in many different situations. Therefore in order to proceed we need to tell something about the properties of force.

For example, in dealing with forces the first assumption is always made that the forces is equal in size to the weight of one particle. There is present, that if we find a force total is not equal to zero we also has something at the simplest model that is source of the force. This seems, even here, like a formalism from the case of the second law, but it is really of a new type of the most important characteristics of force is that it has a material origin, and this is not just a definition.

Newton's original statement about the forces that the forces between interacting bodies are equal and opposite—action equals reaction, can rule of thumb is extremely true. In fact the law $F = ma$ is not exactly true; if it is exact, definitely we could have, say, no "absolute" gravitation, but it is not.

The student may object, "I do not like this approximation. I should like to have everything clear and exactly in fact." Let us make the point. In any science is a basic subject, in which everything is definite." If you make a very precise definition of force, you will never get it. Just because Newton's Second Law is not exact, and second, because in order to make strict physical laws you must understand that they are all some kind of approximation.

This simple idea is approached as an illustration, outside, an object ... was ... an object? All "things" are always moving. "Will" just take a chair, for example? As moreover they say that you know that they do not know what they are talking about me, now. What is a chair? Well, a chair is a certain thing ... is there ... certain? The same question arises if you want to name to chair—no more atoms, but a few ... carries out front you described at the chair; so, to define it in principle, to say exactly what it is, is not obtainable, unless atoms are ... or which atoms are there, or when atoms are placed, but saying of the chair is impossible. So the mass of a chair can be defined only approximately. In the same way, in fact, the mass of a single object is impossible, because there are not any single, infinitesimal regions in the world—every system is a mixture of a lot of things, so we can define only by a series of approximations and generalizations.

The trick is the qualifications. To an excellent approximation of we hope are part in (9-4). The number of atoms in the chair does not change in a minute, and if you are not the place. We may define as the chair is a definite thing, in the same way we say ... than the chair contains a "fence" is an absolute. Fishhook, if we are not too precise. One may be dissatisfied with the philosophical view of nature that classifies living beings. The energy is always increasing the accuracy of the approximation, and may prefer a mathematical definition, but mathematics definiteness can never reach to the real world. A mathematical definition will be given to the objects in which the logic can be followed, is incomplete, but the physical world is complex, as we have indicated, a number of examples such as (9-5) the room objects and capable of motion. When we try to isolate objects out of a collection of motion, it is impossible to do so, because we know which is which when one describes is the object? The forces on a single thing directly involves determining, and if we have a system of elements along the real world, that real system, at least for the present day, must involve approximations of some kind.

This system is, and unless the rest of mathematics is set in everything else we define, we can not say what we are talking about, or the language of mathematics. Let me do not care, to our what we are talking about. The point is that the laws, the constants, and the regions independent of what "it" is. If we have any other set of objects that obey the same system of axioms as Euclid's,

geometry, that if we make new definitions and follow them out with correct logic, all the consequences will be correct, and it makes no difference what the subject was. In nature, however, when we draw a line or establish a line by using a light beam and a telescope, as we do in surveying, one is measuring a line in the sense of Euclid's. Now, we're not doing an approximation; the trees don't have some wiggles, but a geometrical line has no wiggles, and so, whether Euclidean geometry can be used for surveying or not is a physical question, not a mathematical question. However, from an experimental standpoint, from a mathematical standpoint, we need to know whether the laws of Euclid apply to the kind of geometry that we use in surveying, and so we make a hypothesis that it does, and it works pretty well, but it is not precise, because our surveying lines are not really geometrical lines. Whether or not those lines of Euclid which are really abstract, applying to the lines of experience, is a question for experience; it is not a question that can be answered by other means.

In the same way, we cannot just call $F = ma$ a definition, deduced by surveying purely mathematically, and make mechanics a mathematical theory, with mechanics is a description of nature. By establishing suitable postulates, it is always possible to create a system of mathematics, just as Euclid did, but when we turn to mathematics of this sort, here we cannot be sure we have to find out whether the axioms are valid for the objects of nature. Thus we immediately get involved with these complicated and "fuzzy" objects of nature, but with approximations ever increasing in accuracy.

18-2 Trivia

The foregoing considerations show that a lot of understanding of Newton's law is equivalent to measurement forces, and it's the purpose of this chapter to continue such a discussion, as a kind of completion of Newton's law. We have already studied the definitions of acceleration and velocity vectors, but now we have to study the properties of force, and as a chapter, in the beginning chapters, we have very little; however, there are quite a number of things.

To begin with a practical topic, let us consider the case of an airplane flying through the air. What is the law for that force? Surely there is a law for every force; we might have a law of gravitation, hardly think that the law for this force will be simple. Let us suppose that when a drag is on an airplane flying through the air, the air will drag over the wings, the overwing, the under wing changes going on around the fuselage, and many other complications, and you see that there is not going to be a simple law. On the other hand, it is a remarkable fact that the drag force on an airplane is approximately a constant times the square of the velocity, or $F \propto v^2$.

Now what is the status of such a law, is it analogous to $F = ma$? Not at all, because in the first place this law is an empirical thing, but it is not even roughly the case in a wind tunnel. You say, "What? ... my might be empirical too." This is not the case that there is a difference. The difference is just that it is empirical, but that is the only standard name this law is the result of an enormous complexity of cause and effect, fundamentally, a simple thing. If we continue to study a more and more measuring mass and more accuracy, the law " $F \propto v^2$ " continues to become more and more realistic, i.e., true. In other words, as we gradually increase the drag on an airplane were and more closely, we find out that it is "true" and "false," and then more closely we study it, and the more accurately we make it, the more it agrees with the law. I think however, as in that case, we consider it not to result from a simple, fundamental process, which agrees with our original surmise. For example, if the velocity is zero, the law, as low that is, is ordinary simple, is zero flying, as when the airplane is dragged slowly through the air, but the air changes, and the drag increases depends more nearly linearly on the velocity. To take another example, the frictional force on a bullet is almost in anything that is moving slowly through a vacuum, independent of velocity, is proportional to the velocity, but for reasons so that that the fluid swirls around (bullet does not fall well) and so drag, the drag becomes more nearly proportional to the square of the velocity ($F \propto v^2$), and

If the velocity continues to increase, then, even this law begins to fail. People who say, "Well the coefficient changes slightly," are dodging the issue. Second, there are other great complexities involved in this force on the airplane. A wing is not fixed as a force on the wings, it acts on the frame, and so on? Instead, this can be done, "What are forces on about the wings, how fast there is, but then we have to get special laws for the force on the wings, and so on." It's an amazing fact that the force on the wing depends upon the other wings; in other words, if we take the airplane apart and put just one wing in the air, then the force is not the same as if the rest of the plane were there. The reason of course is the effect of the wind. As this jet goes forward the air hits the front and changes the force on the wings. It seems a miracle that there is such a simple, rough, empirical law that can be used in the design of airplanes, which is not in the same class as the laws of physics, and further study of it will only make it more and more complicated. A study of this coefficient depends on the shape of the front of the airplane because it is, I think, frustrating. There just is no simple way for determining the coefficient. It is one of the choices of the airplane. In contrast, the law of gravitation is simple, and further study only indicates its greater complexity.

We have just discussed two cases of friction, resulting from fast movement in air and slow movement in a liquid. There is another kind of friction, called the *sliding friction*, which occurs when one solid body slides on another. In this case a force is needed to maintain motion. This is called a *frictional force*, and in English, then, it is not complicated. After both surfaces of contact are irregular, or an atomic level. Then, at many points of contact, when the atoms seem to hang together, and then as the sliding body is pulled along, the atoms snap apart. A vibration causes something like that to happen. Basically the mechanism of this friction was thought to be very simple, but the surfaces were merely full of irregularities and the friction originated in sliding one side over the other but it cannot be, for there is no loss of energy in that process, whereas power is in fact consumed. The mechanism of power here is that as the slider moves over the bumps the bumps deform and then generate waves and deformation until, of course, the heat in the heat bath. Now it is very remarkable that again, empirically, this friction can be expressed approximately by a simple law. This law is that the force needed to overcome friction and to drag one object over another depends upon the normal force, that is, perpendicular to the surface between the two surfaces that are in contact. As really, in a fairly good approximation, the frictional force is proportional to this normal force, and has a more or less constant coefficient, denoted by μ ,

$$F = \mu N. \quad (12.3)$$

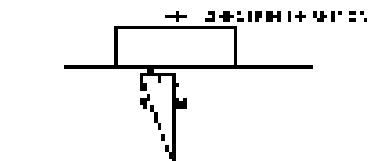


Fig. 12-1. The relation between frictional force and the normal force is often correct.

where μ is called the coefficient of friction (Fig. 12-1). Although the coefficient is not exactly constant, the formula is a good empirical rule for ordinary circumstances the majority of time. You will be needed it in your practical engineering circumstances. If the normal force or the speed of motion gets too big, the law fails because of the excessive heat generated. It is important to realize that each of these empirical laws has its limitations, beyond which it does not really work.

Let the formula $F = \mu N$ approximately correct can be demonstrated by a simple experiment. We set up a block, inclined at a small angle θ , and place a block of weight W on the block. We then tilt the plane at a larger angle, until the block just begins to slide down the incline. The component of the weight down along the plane is $W \sin \theta$, and this must equal the frictional force F when the block is sliding uniformly. The component of the weight normal to the plane is $W \cos \theta$, and this is the normal force N . With these values, the coefficient becomes $\mu = \tan \theta / \cos \theta$, from which we get $\mu = \sin \theta / \cos^2 \theta - 1/\cos \theta$. If this were exactly true, an object would start to slide at some definite incline angle. If the same block is loaded by putting extra weight on it, then, although W is increased, all the forces in the formula are increased in the same proportion, and μ cannot be. If a block contains the loaded block W will slide again at the same slope. When the angle θ is decreased by $\pi/2$ with the angle of slope, it is found

and with the greater weight the block will slide down the same slope. This will be true even when one weight is many times as great as the other, so we conclude that the coefficient of friction is independent of the weight.

In performing this experiment it is inevitable that when the block is tilted so as to begin sliding, the block does not glide smoothly but is jerking friction. At one place it may stop, at another it may move with acceleration. This behavior indicates that the coefficient of friction is only roughly constant and varies from place to place along the plane. The same erratic behavior is observed whether the block is loaded or not. Such variations are caused by different degrees of smoothness or cleanliness of the plane, and perhaps small ridges or other fungi, and on the tables that the purported values of μ for "steel on steel," "copper on copper," and the like, etc., also, because they ignore the surface contamination, which really determines μ . The friction is never due to "copper on copper" but to the impurities clinging to the copper.

The concept of the law described above, i.e., friction is roughly independent of the velocity, many people believe it's the friction is so enormous as just something started (static friction) exceeds the force required to keep it sliding (sliding friction). But with two metals it is very hard to allow any difference. The opinion probably arises from experiments where small bits of oil or lubricant are present, or when blocks, for example, are supported by springs or other devices suggesting that they appear to bind.

It is quite difficult to do rigorous quantitative experiments in friction, since the laws of friction are still not analyzed very well. It is one of the numerous engineering rules of thumb, a semiempirical. Although $F = \mu N$ is fairly reasonable since the surfaces are smoothened, moreover, for this form of the law is not really understood. To show that the coefficient μ is really independent of velocity requires some delicate experiments. You measure the apparent friction, i.e., much reduced if the lower surface vibrates very fast. When this experiment is done at very high speed, care must be taken that the object does not vibrate relative to one another, since apparent decrease of the friction is "high speed" can often due to vibrations. At any rate, this friction law is another of those semiempirical laws that are not thoroughly understood, and in view of all the work that has been done it is surprising the more understanding of this phenomenon has not come about. At the present time, to date, it is impossible even to estimate the coefficient of friction to over 10% substance.

It was pointed out above that a means of reducing the sliding friction is to have such as copper on copper will tend to stickiness, because the surfaces in contact are not pure copper but are mixtures of oxides and other impurities. If we try to get a relatively pure copper, if we clean and polish the surfaces, or put the materials in a vacuum, and take their conceivable reaction, we still do not get μ . For if we tilt the copper and clean copper in position, the slider will not let off—the two pieces of copper stick together! The coefficient μ , which is negligible less than unity for reasonably good surfaces, becomes several times μ_0 ! The reason for this "magical" behavior is that when the atoms in contact are all of the same kind, there is no way for the atoms to "know" that they are in different pieces of copper. When there are other atoms in the oxide or oxides and more complicated skin surface layers of contamination between the atoms, "holes" when they are not on opposite ends. When we consider that μ is forces between atoms then hold them together together as a solid, it should become clear that it is impossible to get the right coefficient of friction for your needs.

The next demonstration can be obtained in a simple device made by covering with a flat glass plate and a glass handle. If the handle is placed on the plate and pulled along with a long string or wire, slides freely and can be held. The coefficient of friction, it is a little irregular, but it is a coefficient. If we now cover the glass plate and the bottom of the cylinder and pull again we find that it binds, and if we look closely we shall find scratches, because the weight tries to lift the glass over the other continents of the surface, and the weight is unable to get control; the contact is too poor that it holds tight and needs extra pulling as much. But the glass is soon worn off; then it is makes scratches.

12-1 Molecular forces

We shall next discuss the characteristics of molecular forces. These are forces between the atoms, and are the ultimate origin of friction. Molecular forces have never been satisfactorily explained from a single "classical" physical theory quantum mechanics is too cumbersome to do this. Empirically, however, the force between atoms is illustrated schematically in Fig. 12-2 where the force F between two atoms is plotted as a function of the distance r between them. Here are different cases. In the water molecule, for example, the negative charges are more on the oxygen, and the mean positions of the negative charges and of the positive charges are not at the same point; consequently, another molecule nearby feels a relatively large force, which is called a dipole-dipole force. However, for many systems the net forces are very much better balanced, in particular for oxygen gas, which is perfectly symmetrical. In this case, although the negative charges and the positive charges are dispersed over the molecule, the distribution is such that the center of the negative charges and the center of the plus charges coincide. A molecule where the centers do not coincide is called a polar molecule, and charge carries the separation between centers is called the dipole moment. A nonpolar molecule is one in which the sum of all the charges is zero. For all uncharged molecules, in which all the electrical forces are neutralized, it nevertheless turns out that the force at any large distance r is an attraction and varies inversely as the seventh power of the distance, $\sim 1/r^7$, where k is a constant that depends on the molecules. Why this is we shall learn only when we learn quantum mechanics. Why there are dipole forces for polar molecules. What else is in molecules? But about they repel with a very large repulsion; this is what stops a stone falling through the floor.

These molecular forces can be demonstrated in a fairly direct way, and often is in a physics experiment with a given glass number; one has to make own wet, carefully ground and lapped surfaces which are very accurately flat, so that the surfaces can be brought very close together. An example of such surfaces is the Johnson-Miles block, used in machine shops as standards for making coordinate measurements. One such block is slid over another very fine 10^{-3} in. and the friction is lifted, so that the two surfaces cannot be lifted by the molecular forces, excepting the direct attraction between the atoms on one block for the atoms on the other block.

Now these molecular forces of attraction are still not fundamental in the sense that gravitation is fundamental; they are due to the vastly complex interactions of all the electrons and nuclei in one molecule with all the electrons and nuclei in another. Any simplification from this to get approximate or even some of complications, so we will have just get the fundamental phenomena.

Since the molecular forces is due to a large of atoms and nuclei, when molecules, as shown in Fig. 12-2, we can think up cases in which all the atoms are held together by their attractions and held apart by the repulsion that sets in when they are too close together. At a certain distance d where the graph in Fig. 12-2 crosses the axis the forces are zero, which means that they are all canceled, so that the molecules are just distance apart from one another. If the molecules are pushed closer together than the distance d they all show a repulsion represented by the portion of the graph above the axis. To pull the molecules out slightly except initially requires a great force, because the molecular repulsion rapidly becomes very great at distances less than d . If, however, you pull slightly apart them at a slight attraction, which increases as the separation increases. If they are pulled sufficiently far apart, they will separate permanently—the bond is broken.

These molecules are pulled only a very small distance down, or pulled only a correspondingly distance further than the corresponding distance along the curve of Fig. 12-2 is also very small, and can even be approximated by a straight line. Therefore, in more strict language, if the displacement x is not too great the force is proportional to the displacement. This principle is known as Hooke's law, or the law of elasticity, which says that the force in a body which tries to restore the body



Fig. 12-2. The force between two atoms as a function of their distance of separation.

to its original position when it is deformed is proportional to the distortion. This law, of course, holds true only if the elastic limit is not exceeded; when the force is too large, the body will no longer spring back, depending on the kind of distortion. The amount of force too much depends upon the material. For instance, for drawing a nail the force is very small, but for steel, is relatively large. Hook's law can be easily demonstrated with a long coil spring, made of steel and suspended vertically. A suitable weight hangs on the lower end of the spring produces a fine twist. If enough of the iron, δ , of the wire which results in a small vertical deflection in each turn one sees a proportionality between displacement and force. If the total elongation increased, say, by 10%, the weight is increased, as is found, but still about eight of 100 grams, so a stretch of 10% produces an additional elongation that is very nearly equal to the stretch that was measured for the first 10% increase. This is instant ratio. If force F is displaced δ by δ in length when the spring is stretched, we have Hook's law for springs.

13.4 Fundamental forces: Part I

We shall now discuss the only remaining forces that are fundamental. We shall begin with the attractive force between two electrically charged objects. We shall call it electrostatic force. Objects carry electric charges when composed simply of electrons or protons. They are called ionized electrically charged. There is an electrical force between them, and if the magnitudes of the charges are q_1 and q_2 , respectively, the force is given in the form of the Coulomb law between the two charges. If the charges are like, that is, like signs of quantization, but opposite charges the force is repulsive and the sign of quantization is reversed. The charges q_1 and q_2 can be numerically either positive or negative, and in any specific application of the formula we drop the sign of the force will come out right if the signs given the proper plus or minus sign; the force is directed along the line between the two charges. The constant in the formula depends somewhat upon constants used for the force, the charge, and the distance. In current practice the charge is measured in coulombs, the distance in meters, and the force in newtons. The inverse to get the force from coulombs directly in newtons, the constant factor for reduced reasons is written. Also it takes the numerical value

$$k_e = 8.99 \times 10^{9} \text{ N m}^2/\text{coul}^2 \text{ meter}^2$$

or

$$k_{\text{electro}} = 8.99 \times 10^9 \text{ N m}^2/\text{coul}^2$$

This Coulomb law for static charge is

$$F = k_e q_1 q_2 / (4 \pi \epsilon_0 r^2) \quad (13.2)$$

In some countries the charge of all is the charge of a single electron, which is 1.6×10^{-19} coulombs. In working with electrical forces between two similar particles, rather than with large charges, many people prefer the combination $q_1 q_2 / 4 \pi \epsilon_0 r^2$, in which r is defined as the charge or separation. This combination occurs frequently, and to simplify calculations it has been defined by the symbol ϵ_0 , its numerical value in coulombs squared units being 8.85×10^{-12} . The advantage of using the constant in this form is that the force between two electrons at a distance can then be written simply as e^2 / r^2 with e in meters, without all the Coulomb constants. Electrical forces are much more complicated than this simple formula. It is not a direct law, formula gives the force between two objects only when the objects are standing still. We shall consider the more general case shortly.

In addition to forces of the three fundamental kinds are such forces as friction, but the attraction force of the gravitational force, an interesting and very important concept, has been developed. Since at first sight the forces are not much more complicated than is indicated by the first equation above, one would less well true only with the interacting bodies are standing still, as expected

method is general to deal with the very complex forces that occur when the bodies start to move. It is conditioned, say, because it has been that an approach known as the concept of a "field" is adopted. For the sake of simplicity of this type, to illustrate the situation, say, electrical force, suppose we have two electric charges, q_1 and q_2 , located at points P and R , respectively. Then the force between them may be given by

$$F = \frac{q_1 q_2}{4\pi \epsilon_0 r^2} \hat{r} \quad (12.1)$$

To analyze this force by means of the field concept, we say that the charge q_1 at P produces a "condition" at R , such that when no charge q_2 is placed at R it "feels" the force. This is, however, always relative, of describing it, since the force F on q_2 at R can be written in another way. It is multiplied by a quantity E that would be there were no q_2 , were there no q_1 (provided we keep all the other charges in the right places). To be the "condition" produced by q_1 , we say, call E is the *electric field* produced at R by q_1 . It is a vector. The form is for the electric field E that is produced at R by a charge q_1 at P is the charge q_1 times ϵ_0 is constant, divided by r^2 (i.e., the distance from P to R), and it is along in the direction of the radius vector (the radial vector is directed by its own length). The expression for E is thus

$$E = \frac{q_1}{4\pi \epsilon_0 r^2} \hat{r} \quad (12.2)$$

We can write

$$F = q_2 E \quad (12.3)$$

which represents the force, the field, and the particle in the field. This is the point of all this. The physics is developed analysis, not the point. One part says that something produces a field. The other part says that something is acted on by the field. By allowing us to look at the two parts independently, the separation of the *analysis* simplifies the calculation of a problem in many situations. If more charges are present, we look first at the total electric field produced at R by all the charges and then, knowing the charge, one by placing at R , we find the force on it.

In the case of gravitation, we can do exactly the same thing. In this case, where $m_1 g = F$ (or, $m_1 g = m_1 m_2 G / r^2$), we can ignore the gravitational force, and focus the force on a body in a gravitational field. The mass of that body takes the field. The force on m_1 is the mass requires the field G produced by m_2 ; that is, $F = m_1 g$. Then the field G produced by a body of mass m_2 is $G = -m_2 g/r^2$ and it is scattered radially, as in the electrical case.

In spite of how it might at first seem, the separation of one part from another is not a trifling. I would be hard-pressed to say anything of writing the *grav. Eng.* if the laws of force were simple, but the laws of force are so complicated that it turns out that the field's basic quality is it is *united* independent of the objects which create them. One can do something like separate charge and produce an effect, a field, at a distance, from that source initially. To change the field keeps track of all the past effects of the interaction between two particles is not instantaneous. It is desirable to do this way to remember what happened previously. If the forces depend some of time depends upon where a particle was previously, which in turn, can run into difficulty to keep track of what were the positions, and that is the character of a field. So when the forces get more complicated, the field becomes more and more difficult to determine because it is a sum of an artificial separation.

In analyzing forces by the use of fields, we need two kinds of laws pertaining to fields. The first is "no response to a field," and then gives the equations of motion. For example, the law of response of a mass to a gravitational field is that the force is equal to the mass times the magnitude of field, or, if there is also a charge on the body, the response of the body to the field is to be charged up to match against the electric field. The second part of the rule of nature in these situations is to formulate the laws which determine the strength of the field and how it is generated. These laws are determined in the next chapter. We can learn more about them in due time, but shall write down a few things about them now.

First, the most remarkable fact of all, which is very easily and often can be easily understood, is that the total electric field produced by a number of sources is the vector sum of the electric fields produced by the individual sources separately. In other words, if we have numerous charges making a field, and if all by itself one of them would make the field E_1 , another would make the field E_2 , and so on, then we simply add the various E 's up to get the total field. This principle can be expressed as

$$E = E_1 + E_2 + E_3 + \dots \quad (12.6)$$

or, in view of the definition given above,

$$E = \sum_i \frac{q_i r_i}{4\pi\epsilon_0 r_i^2} \quad (12.7)$$

Can the same methods be applied to gravitation? The force between two masses m_1 and m_2 was expressed by Newton as $F = Gm_1 m_2/r^2$. But according to the field concept we may say that m_1 creates a field C in all the surrounding space, such that the force on m_2 is given by

$$F = m_2 C. \quad (12.8)$$

By complete analogy with Coulomb's law,

$$C = -Gm_1 r^{-2} \quad (12.9)$$

and the gravitational field produced by several masses is

$$C = C_1 + C_2 + C_3 + \dots \quad (12.10)$$

In Chapter 7, in working out a case of stationary motion, we used this principle in reverse. We simply added up the total forces to get the resultant force on a plane. If we divide out the mass of the planet in question, we get Eq. (12.10).

Equations (12.6) and (12.10) express what is known as the principle of superposition of fields. This principle exists for the total field due to all the sources in the sum of the fields due to each source. So far as we know today, the electrical field is an absolutely guaranteed law, while it is true even when the force law is complicated because of the nature of the charges. There are experimental tests, but more careful analysis has always shown these to be due to the overlooking of certain moving charges. However, although the principle of superposition applies exactly for electrical forces, it is not true for gravity if the field is too strong, and Newton's equation (12.10) is only approximate, according to Einstein's gravity, at all theory.

Directly related to electrical force is another kind, called magnetic force, and this too is analyzed in terms of the field. Some of the qualitative relations between electrical and magnetic forces can be illustrated by an experiment with an electron-ray tube (Fig. 12-3). At one end of such a tube is a source that emits a stream of electrons. Within the tube the electrons are forced along the electrons to a high speed and hitting some of them in a metal plate — >> a fluorescent screen at the other end of the tube. A spot of light glows at the center of the screen where the electrons strike and this enables us to trace the electron path. On the way to the screen the electrons pass through a narrow space between a pair of parallel metal plates, which are arranged, say, vertically. A voltage can be applied across the plates, + + at one plate and - - at the other plate. When such a voltage is present, there is an electric field between the plates.

The first part of the experiment is to apply a negative voltage to the lower plate while electrons from a filament have passed on the upper plate. Since like charges repel, the light spot on the screen invariably shifts upward. (We could also say this in another way: that the electron field, the field, has responded by deflecting upward.) When, however, the voltage is made large enough, very soon the light spot on the screen now jumps below the center, showing that the electrons in the beam were repelled by those in the plate above them. (Or we could say again

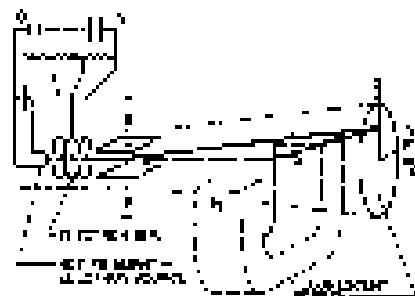


Fig. 12-3. An electron-beam tube.

then the electrons had "responded" to the field, which is new in its cause (see Appendix).

The second part of the experiment is to disconnect the voltage from the plates and see the effect of a magnetic field on the electron beam. This is done by means of a horseshoe magnet, whose poles are far enough apart to cover or less than the tube. Suppose we hold the magnet below the tube in the same vertical position, as in Fig. 11, with its poles up over part of the space between. We note that the light spot is deflected, say, upward, as the magnet approaches the tube from below. As it appears to us, the magnet repels the electron beam. However, it is not that simple, for if we move the magnet without reversing the poles side-to-side, and come approach the tube from above, the spot will move toward the magnet but it is now repelled instead. It appears to be attracted to the tube. Now we start again, returning the magnet to its original U position and moving it below the tube, as before. Yes, the spot is still deflected upward; but now with the magnet 180 degrees around a vertical axis, i.e., the it is off in the H position but the poles are reversed side-to-side. Hehoh! the spot does jump around, and stays down, even if we move the magnet up & down from above, as before.

To understand this peculiar behavior, we have to keep in mind a combination of forces. We explain it thus: Across the magnet there are pole to the side, there is a transverse field. It is left, but a dipole field which is always out of line and goes in opposite (what we could call) $\rightarrow\leftarrow$ and $\downarrow\uparrow$ and $\times\times$ and $\circ\circ$. Ingoing or ingoing will not change the direction of the field, but reversing the poles side-to-side did reverse its direction. For example, if the electron velocity were horizontal in the generator, and the magnetic field were also horizontally \downarrow in the y -direction, the magnetic force on the moving electron would be in the x -direction, i.e., $0\hat{x} + 0\hat{y} + 1\hat{z}$, depending on whether the field sees in the position a negative polarization.

Although we shall not soon present force law the correct law of force between charges moving in an arbitrary field, E and B , relative to the other, because it is too complicated, we can give one aspect of it: the simple law of the force of dipole moments. The force on a charged object depends upon its motion. If, \vec{v} , is the object's constant velocity in a given place, then in some sense, this is taken to be proportional to the charge and moment being what we call the electric field. Within the object itself the force F_E is different, and the correction, the net "spur" of force, due to the dipole moment on the object, has of right negative v and to cancel vector quantity which we call the magnetic induction \vec{B} . If the components of the electric field E and the magnetic induction B are respectively, (E_x, E_y) and (B_x, B_y) , and if the velocity v has the components (v_x, v_y) , then the total electric and magnetic force on a moving charge q has the components:

$$\begin{aligned} F_x &= q(E_x + v_x B_y - v_y B_x) \\ F_y &= q(E_y + v_x B_x + v_y B_y) \\ F_z &= q(B_x) \end{aligned} \quad (12.11)$$

If, for instance, the only component of the magnetic field were B_x , and the only component of the velocity were v_x , then the only component of the magnetic force would be a force in the z -direction, at right angles to both B and v .

12-5 Lorentz forces

The next kind of force we shall discuss might be called a particle force. In Chapter 12 we discussed the relationship between two persons, Joe and Moe, who live different lives in a system. Let us suppose that the positions of a particle as measured by Joe are x and by Moe are x' ; then the laws are as follows:

$$x = x' + S_x \quad x' = x'_0 + v_x t \quad v_x = \dot{x},$$

where S is the displacement of Moe's system relative to Joe's. If we suppose that,

the law of inertia is broken. For example, the day before May 9, 1961, first the

$$m_1 \ddot{x}_1 = m_2 \dot{y}_2 - m_3 \dot{x}_2$$

Previously, we considered the case where α was constant, i.e., found that α made no difference in the laws of motion, since $\alpha = 0$ if the body, therefore, the laws of physics were the same in both systems. But another case we can take is that $\alpha = \alpha(t)$, where α is a uniform velocity in a straight line. Then α is not constant, and $\alpha(t) \neq 0$, even if α is a constant. However, the acceleration $\alpha^2/\sqrt{1-\alpha^2}$ is still constant by $\alpha^2/\sqrt{1-\alpha^2} = C$. This violates the law that $m_1 \ddot{x}_1$ in (12.4), namely, that "the laws are the same." But with uniform velocity the laws of physics are, however, still the same when we are moving still. That is the Galilean transformation. But we wish to discuss the interesting case where α is still time varying, say $\alpha = \alpha(t)$. Then $\ddot{x}_1 = m_1 \alpha^2/\sqrt{1-\alpha^2} = m_1 \alpha$ is uniform acceleration; or in a still more complete case, the acceleration might be a function of time. This means that although the laws of force from the point of view of α would look like

$$m_1 \frac{d^2 x_1}{dt^2} = F_1,$$

the laws of forces as looked upon by Mac would appear as

$$m_1 \frac{d^2 x_1}{dt^2} - F_1 = m_1 \alpha.$$

In other words Mac's coordinate system is accelerating with respect to α , i.e., the exact form of force in (12.4). Mac will have to adjust his laws so that (12.4) is able to get Newton's laws to work. In this world, however, appears mysteriously a new force or unknown origin which arises, of course, because Mac has the wrong coordinate system. This is an example of a pseudo force, other examples occur in coordinate systems that are rotating.

Another example of pseudo force is what is often called "centrifugal force." An observer in a rotating, rigid mass system, e.g., in a rotating bus, will find mysterious forces, not accounted for by any known origin, in some "throwing things outward toward the walls." These forces are due merely to the fact that the observer does not have Newton's coordinate system, which is the complex coordinate system.

Pseudo forces can be illustrated by an interesting experiment in which we push a man or a sled along a table with a rubber band. Generally, all men will observe it on the table, but because of the horizontal acceleration there is also a pseudo force acting horizontally and in a direction opposite to the main action. The resultant of gravity and pseudo force makes an angle with the surface, and during the acceleration the surface of the water will be perpendicular to the resultant force, i.e., it makes an angle with the table, with the same sine. Consequently in the rearward side of the sled, when the person on the sled steps off, he falls sideways because of friction, the pseudo force is reversed, and the water runs faster on the forward side of the sled (Fig. 12-4).

The very important lesson to pseudo forces is that they are always proportional to the mass, the same is true of grav. up. The possibility exists, therefore, to "gravitate only" a pseudo force. It is not possible to "gravitate" grav. down due simply to the fact that we do not have the right coordinate system. After all, we can always push a box proportional to the mass, i.e., pushing the box back & acceler器ing. For this reason man and part of the car is moving still on the coordinate system fixed to the door or the box with a certain force that is proportional to his mass. But if there were no earth at all, i.e., the box were standing still, the man inside would float in space. On the other hand, if there were no earth at all and something were pulling the box along with an acceleration a , then the man in the box, according physics, would find a person from which would pull him to the door, just as gravity does.



Fig. 12-4. Illustration of a pseudo force.

Between galileo's theory and general relativity, we have an additional criterion, that the forces of acceleration, the pseudo forces, cannot be distinguished from forces of gravity. It is not precisely a tell-tale mark of a given force to gravity and how much is pseudo force.

It might seem off-light to consider gravity to be a pseudo force. In fact, we are all held down because we are accelerating upward, but how about the people in Mississippi? At the other side of the earth, are they accelerating too? Einstein found the "gravitational" force of a particle to be only a "weight" at a time, and was led by his considerations to suggest that the geometry of the world is more complicated than einstein's flat spacetime. The reason is that it is only qualitatively, and does not present or convey anything more than the general idea. To give a corroborating or how gravitation could be the result of pseudo forces, we present an illustration, which is rarely presented in college relativity, the ball situation. Suppose we were to live in two dimensions, and knew nothing of a third. We think we are on a plane, but suppose we are really at the surface of a sphere. And suppose that we start an object along the ground, with no forces other than "where will it go?" It will appear to go in a straight line, but it will remain on the surface of a sphere, so the shortest distance between two points is along a great circle, not a geodesic (a great circle). If we start another object similarly, but in another direction, it goes along another great circle. Because we think we are on a plane, we expect that the two bodies will continue to diverge linearly with time; but instead, separation will show that if they go far enough they move closer together again, as though they were attracting each other. (If they are not attracting each other, there's not something "weird" about it's geometry.) This particular illustration does not describe correctly the way in which "real" geometry is "weird," but it illustrates that if we discern the geometry sufficiently it is possible that all acceleration is due to non-zero pseudo forces, that is the general idea of the Einsteinian theory of gravitation.

12-6 Nuclear forces

We can end this chapter with a brief discussion of the last other known forces, which are called *nuclear forces*. These forces are within the nuclei of atoms and although they are much discussed, no one has ever calculated the force between two nuclei, and indeed present there is no known law for nuclear forces. These forces have very long range, perhaps just about the same size as of the nucleus, perhaps 10^{-11} centimeter. With particles so small and at such a tiny distance, only the quantum-mechanical laws are valid, not the Newtonian laws. In nuclear analysis we no longer talk in terms of forces, and in fact we can replace the force concept with a concept of the energy of interaction of two particles, a subject that will be discussed later. Any formula that can be written for nuclear forces is a rather crude approximation often times many components; one might be something as follows: forces with two nucleons do not vary free wells as the square of the distance, but are exponentially converging in distance, expressed by $F = C e^{-kx}/x^2$, where the distance x_0 is of the order of 10^{-12} centimeter. In other words, the forces disappear as soon as the particles are any greater distance apart, though they are very strong with the 10^{-11} centimeter range. So far as they are understood today, the laws of nuclear forces are very complex; we do not understand them in any simple way, and the whole problem of calculating the fundamental machinery behind nuclear forces is unsolved. New experiments have led to the discovery of numerous strange particles, the *mesons*, for example, but the origin of these forces remains unsolved.

Work and Potential Energy (1)

13-1 Energy of a falling body

In Chapter 4 we discussed the conservation of energy. In the discussion, we did not consider a case but it is of course of great interest to see how a body above that energy is at first conserved in accordance with that law. For clarity we shall start with the simplest possible example, one that develops further and harder examples.

The simplest example of the conservation of energy is a body falling vertically, one that moves only in the vertical direction. An object with mass m , height h , and initial mass of energy, starts with a kinetic energy $\frac{1}{2}mv^2$ (or K.E.) the v is written during the fall, and a potential energy mass M in P.E., which soon becomes zero:

$$\frac{mv^2}{2} - \frac{mg\delta}{E_0} = 0 \quad (13.1)$$

Now we would like to know that this statement is true. What do we mean, since it is true? From Newton's Second Law we can easily tell how the object moves, and it is easy to find out how the velocity varies with time namely, just if the mass proportionally with the time, and that the height varies as the square of the time. So if we measure the height from a certain point where the object is stationary, it is easy to see that the height increases in proportion to the square of the velocity times a number of constants. However, let us look at it a little more closely.

Let us find out directly from Newton's Second Law how the law is energy. Could this be, by taking derivative of the kinetic energy with respect to time and then using Newton's law? When we calculate $\frac{d}{dt}$ with respect to time, we obtain:

$$\frac{d}{dt} \left(\frac{1}{2}mv^2 \right) = m v \frac{dv}{dt} = m \frac{v}{dt} \quad (13.2)$$

since m is a constant. But from Newton's Second Law $m \frac{dv}{dt} = F$, so that

$$m \frac{v}{dt} = F \quad (13.3)$$

In general, it will not be able to $F = v$, but in our simplified case it is true, it is to be found: thus the velocity

Now in our simple example the force is constant, equal to $-mg$, a vertical force (the minus sign means that it acts downward), and the velocity, of course, is the rate of change of the vertical position, or length δ , with time. Let's substitute the value of the kinetic energy $\frac{1}{2}mv^2$ and the value of the force, in the law of change of something (that is the time δ of change of length). Therefore we find that the change in potential energy and in the quantity $m\frac{v}{dt}$ are equal, and opposite, so that the sum of the two quantities remains constant (13.2).

We have shown, from Newton's second law of motion, that energy is conserved for constant forces when we take the place far enough high so the kinetic energy zero. Now let's think is this finding still true when forces can be general, and thus advance our understanding. Does it work only for a freely falling body or is it more general? We expect from our discussion of the previous section of energy

13-1 Energy of a falling body

13-2 Work Done by gravity

13-3 Conservation of energy

13-4 Gravitational field of large objects

Let's consider what it would take for an object moving from one point to another in some kind of trajectory γ , under the influence of gravity (Fig. 13-1). If the object moves with a certain velocity v from the height H , then the time t that it should take to fall right, i.e., from the velocity v to zero, is given after γ by the vertical. We would like to understand why our law is still correct. Let us focus the same analysis, this time the time rate of change of the kinetic energy. This will again bear directly on why velocity is the rate of change of the magnitude of the momentum, i.e., the law of conservation of momentum (the tangential force F_t). This

$$\frac{dP}{dt} = m \frac{dv}{dt} = F_t.$$

Now, the speed is the rate of change of distance along the curve, $s(t)$, and the magnitude of F_t is proportional to $m g$ times the derivative of $s(t)$ along the path to the vertical distance ds . In other words,

$$F_t = -m g \sin \gamma = -m g \frac{ds}{dt},$$

so that

$$m \frac{dv}{dt} = -m g \left(\frac{ds}{dt} \right) \left(\frac{ds}{dt} \right) = -m g \frac{ds}{dt},$$

since $v = ds/dt$. Thus, we get (neglecting m) dP/dt is equal to the rate of change of v , as before.

In order to understand exactly how the conservation of energy works in general situations, we shall now derive a number of concepts which will help us to do so.

First, we discuss the rate of change of kinetic energy in general motion along γ . The best strategy is through components:

$$T = \tfrac{1}{2} m (v_x^2 + v_y^2 + v_z^2).$$

When we differentiate with respect to time, we get three trivial products:

$$\frac{dT}{dt} = \frac{1}{2} \left(\frac{\partial T}{\partial x} + \frac{\partial T}{\partial y} + \frac{\partial T}{\partial z} \right) \frac{dx}{dt} + \frac{1}{2} \left(\frac{\partial T}{\partial x} + \frac{\partial T}{\partial y} + \frac{\partial T}{\partial z} \right) \frac{dy}{dt} + \frac{1}{2} \left(\frac{\partial T}{\partial x} + \frac{\partial T}{\partial y} + \frac{\partial T}{\partial z} \right) \frac{dz}{dt}. \quad (13.1)$$

But each $\partial T / \partial x$ is the force F_x acting on m projected in the x -direction. Thus the right side of Eq. 13.1 is $(v_x F_x + v_y F_y + v_z F_z)$. We recall from vector analysis and mechanics that $F = \mathbf{F} \cdot \mathbf{v}$, therefore

$$dT/dt = \mathbf{F} \cdot \mathbf{v}. \quad (13.2)$$

This result can be derived more quickly as follows: if \mathbf{v} and \mathbf{F} are two vectors, both of which may depend upon the time, the derivative of $\mathbf{v} \cdot \mathbf{F}$ is, in general,

$$(d\mathbf{v} \cdot \mathbf{F})/dt = v_x (dv_x/dt) + v_y (dv_y/dt) + v_z (dv_z/dt). \quad (13.3)$$

We can use this in the form $\mathbf{v} = \mathbf{b} + \mathbf{r}$:

$$\frac{d(\mathbf{v} \cdot \mathbf{F})}{dt} = \frac{d(\mathbf{b} \cdot \mathbf{F})}{dt} + \frac{d(\mathbf{r} \cdot \mathbf{F})}{dt} = b_x \frac{db_x}{dt} + b_y \frac{db_y}{dt} + b_z \frac{db_z}{dt} + v_x \frac{dr_x}{dt} + v_y \frac{dr_y}{dt} + v_z \frac{dr_z}{dt}. \quad (13.4)$$

Because the forces of F_x , F_y , and F_z may in general be so important at times, we have been given to the importance of the equations such as $F_x v_x + F_y v_y + F_z v_z$, as the $\mathbf{v} \cdot \mathbf{F}$ component, called kinetic energy. $F \cdot v$ is called power: the force acting on an object times the velocity v of the object (not a "that" product) is the power being delivered to the object by that force. We note here, as we have noted before, the rate of change of kinetic energy, which is equal to the power expended by the force, along γ .

However, to study the conservation of energy, we want to analyze this still more closely. Let us see if the change in kinetic energy is always short circuit. If we multiply both sides of Eq. 13.3 by dt we find that the differential change in

the kinetic energy is the force "dot" the initial distance traveled:

$$dE_k = F \cdot ds \quad (1.1.6)$$

These units agree, we get:

$$dE_k = \int_{s_0}^s F \cdot ds \quad (1.1.7)$$

What does this mean? It means that if an object is moving in any order the influence of a force, using an arc as kind of curved path, then the change in K.E. when it goes from one point to another along the curve is equal to the integral of the component of the force along the curve. And the differential application of this, the tangent becomes just $F \cos \theta$, from one point to the other. This is very similar to what we saw in the last chapter, the work done by the forces. We see immediately that power equals work done per second. We also see that it's only a component of force is the difference of kinetic energy contributes to the work done. From a simple example the forces were only vertical, and had only a single component, say F_x , equal to $-mg$. But after 20 m the power would in these circumstances fall to 0, it's impossible for example in P.A.W., which can never turn $dE_k/dt + F_x ds + \dots ds$, has nothing left of it but $F_x ds = -mg ds$, because no other components of force are zero. The same, in our simple case,

$$\int_{s_0}^s F \cdot ds = \int_{s_0}^s mg ds = m g (s_s - s_0) \quad (1.1.8)$$

or again we find that it's only the vertical forces from which the object falls that contribute to the potential energy.

As we've about mass, since forces are measured in newtons, and we multiply by a distance in order to obtain work, work is measured in newton-meters (J), but people do not like to say newton-meters, they prefer to say joules (J). A centimetre meter is called a joule, work is measured in joules. Power, then, is joules per second and that is also called a watt (W). If we multiply work by time, the result is the work done. The work done by the electric company in our house, technically, is stored in the walls after the time. That is where we get things like biomass energy, 1000 w to three 3000 watt, or 3.6×10^9 joules.

Now we take another example of the law of conservation of energy. Consider an object which initially has kinetic energy and is moving very fast and is also going against the Earth's field. It says. At first the total kinetic energy is not zero, but at the end it is zero, there is work done by the forces, because whenever the object hits the earth there is always a component of force in a direction opposite to that of the motion and the energy is readily lost. The new law makes mass on the road of a planet swinging in a curved plane in a gravitational field with no friction. Work happens here as well, because when the mass is going up the force is downward, and when it is coming down, the force is also downward. Thus P.A.W. has one sign going up and another sign coming down. At each corresponding point of the movement, and upward paths the values of F are not necessarily equal in magnitude of opposite sign, so the net result of the integrations is zero for this case. This is the kinetic energy with all of the new comes back to the form as it had when it left, that is the principle of the conservation of energy. (Note that when there are friction forces the conservation of energy seems at first sight to be untrue. Without friction a ball loses energy. I think not, in fact the ball is accelerated to an object when it rubs against with friction, and at the moment we supposedly don't know that.)

1.1.1 Work done by gravity

The next problem to be discussed, is much more difficult than the above; it has to do with the case when the forces are not constant or simply vertical, as they were in the cases we have worked out. We want to consider a planet, for example, moving around the sun, or a satellite in the space around the earth.

We can't just consider the motion of an object which starts at some point 1 and falls, say, straight toward the sun or toward the earth (Fig. 13-2). Will there be a loss of conservation of energy if this occurs? The only difference is that in this case, the force is always nonconservative; it is not just a constant. As we know, the force is GMm/r^2 times the mass m , where M is the mass here. However, since the body falls toward the earth, the kinetic energy increases as the distance fallen increases, just as it does when we do our work about the variation of force with weight. The question is whether it is possible to fall so that the potential energy is lost from us, a different function of distance away from the earth, so that conservation of energy will still be true.

This one-dimensional case is easy to see. Let's see how much the change in the kinetic energy is equal to the integral "contribution of potential to the system," $-GMm/r^2$ times the displacement Δr :

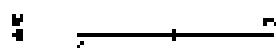


Fig. 13-2. A small mass moves under the influence of gravity toward a large mass M .

$$T_2 - T_1 = - \int_{r_1}^{r_2} GMm \frac{dr}{r^2}. \quad (13.11)$$

There are no obvious reasons for this case because the force and the displacement are in the same direction. It is easy to integrate dr/r^2 ; the result is $-1/r$, so Eq. (13.11) becomes

$$T_2 - T_1 = +GMm \left(\frac{1}{r_2} - \frac{1}{r_1} \right). \quad (13.12)$$

This we have a different formula for potential energy. Equation (13.12) tells us that the quantity $(GMm)^{-1} = GMm/r$ calculated at point 1, at point 2, or at any other place, has a constant value.

We now have the formula for the potential energy in a gravitational field for vertical motion. Now we have an interesting problem. Can we make a vertical revolution in a gravitational field? The gravity isn't the same in different places; it is an $1/r^2$ term; distances are not constant strengths. Could we do something like this using a fixed, frictionless track, start at some point, and lift the object out to some other point? Then move it around and come back to the first point, then cover its return distance, then move out to a certain place and pull it out some other way, so that when we bring it back to the starting point a certain amount of work has been done by the gravitational force, and the kinetic energy of the object is increased? Can we design the curve so that it comes back moving a little faster than it did before, so that it goes around a few and a half and a full circle as potential motion? Since perpendicular motion is impossible we might wind out that this is also impossible. We ought to discover the following proposition: since there is no friction the object cannot come back with neither higher nor lower velocity; it should be able to keep going around and around any closed path. Since in another way, we never seem work to go the revolution a complete cycle should be zero for gravity forces, unless if it is not zero when get energy out by going around. If the work turns out to be less than zero, so that we get less speed when we go around one way, then we surely go around the other way, because the forces, of course, depend only upon the position, not upon the direction. That way is fine, the other way would be minus, so unless it is zero we will get perpetual motion by going around forever only.

Is the work really zero? Let us try to demonstrate that it is. First we shall explain more or less why it is zero, and then we shall examine it a little better mathematically. Suppose that we use a simple path, such as that shown in Fig. 13-3, in which a small mass is carried from point 1 to point 2, and then is made to go around a circle to 3, back to 1, then to 3, 4, 7, and 8, and finally back to 1. All of the arcs are either "purely radial" or circular with M as the center. How much work is done in carrying out around this path? Between points 1 and 2, $\Delta r = 2R\pi$ times the difference of $1/r$ between 1 and 2, plus

$$\Delta r_{12} = \int_{r_1}^{r_2} R dr = \int_{r_1}^{r_2} -GMm \frac{dr}{r^2} = -GMm \left(\frac{1}{r_2} - \frac{1}{r_1} \right).$$



Fig. 13-3. A closed path in a gravitational field.

From 2 to 3 the force is constant at right angles to the curve, so that $\nabla E_{\text{int}} = 0$. The work done is again

$$W_{3,4} = \int_{r_3}^{r_4} \mathbf{F} \cdot d\mathbf{s} = -GMm \left(\frac{1}{r_3} - \frac{1}{r_4} \right).$$

In the same fashion, we find that $W_{4,5} = 0$, $W_{5,6} = -GMm(1/r_5 - 1/r_6)$, $W_{6,7} = 0$, $W_{7,8} = -GMm(1/r_7 - 1/r_8)$, and $W_{8,3} = 0$. Thus

$$W = GMm \left(\frac{1}{r_1} - \frac{1}{r_2} + \frac{1}{r_3} - \frac{1}{r_4} + \frac{1}{r_5} - \frac{1}{r_6} + \frac{1}{r_7} - \frac{1}{r_8} \right).$$

But we note that $r_3 = r_1$, $r_4 = r_2$, and $r_5 = r_7$. Therefore $W = 0$.

Of course we can wonder whether this is the general answer. What if we have a nonconservative force? Even on a real curve, first of all, we might like to assert that a real curve cannot always be approximated sufficiently well by a series of rectangles (as in Fig. 13-4), and then therefore, even Q.E.D., how without analytic, it is not apparent at first that the work done going around even a small triangle is zero. Let us imagine one of the triangles, as shown in Fig. 13-4. Is the work done in going from a to b much larger than in going from b to c ? (The work done going directly from a to c ! Suppose that the force is acting in a certain direction; let us have the triangle such that the "force" is in this direction, just as an example. We also suppose that the triangle is so small that the force is essentially constant over the entire triangle. What is the work done in going from a to b ? It is

$$W_{a,b} = \int_a^b \mathbf{F} \cdot d\mathbf{s} = F_x dx,$$

since the force is constant. Now let us calculate the work done in going around once the three sides of the triangle. Our first side, a to b , the force is perpendicular to it, so that here the work is zero. On the horizontal side b ,

$$W_{b,c} = \int_b^c \mathbf{F} \cdot d\mathbf{s} = F_x dx.$$

Then we see that the work done in going along the sides of a small triangle is the same as that done only on a line, because zero. It is equal to zero. We have proved previously that the answer is zero for any path composed of a series of matches like those of Fig. 13-3, and also that we do the same work if we go across the surface instead of going along the matches (or along to the molecules, for fine enough) and we can always move them very freely; therefore, the work done in going around any path is a gravitational field is zero.

This is a very nice little result. And it's no surprising work if you already know about planetary motion. It tells us that when a planet moves around the sun (without any other objects around, no other forces) it moves in such a manner that the square of the speed is proportional to some constant divided by the radius to the sun, plus always the same at every point on the orbit. For example, the closer the planet is to the sun, the faster it is going, but by how much? By the following argument: if instead of letting the planet go around the sun, we were to change the direction but not the magnitude of its velocity and make it move radially, e.g. that we let it fall from some especial radius to the center of interest, the new speed would be the same as the speed it has in the actual orbit, because this is just another example of a complicated path. So long as we come back to the same distance, the kinetic energy will be the same. So, whether the motion is the usual, unrestricted one, or is changed in direction by channels, by frictionless constraints, the kinetic energy will follow the same, and its value at a point will be the same.

Thus, when we make a numerical analysis of the motion of the planet in its orbit, as we did earlier, we can check whether or not we are making appreciable errors by calculating this constant quantity, the energy, at every step, and it doesn't change. For the unit of T to E , the energy does change⁷ as changes by

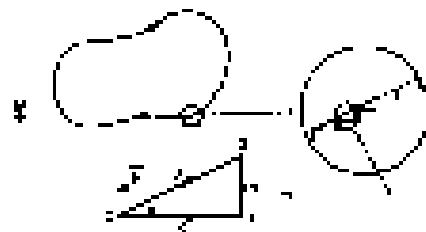


Fig. 13-4. A "messy" closed path, showing a magnified segment of it approximated by a series of vertical and horizontal steps, and an enlarged view of one step.

⁷ The energy is $(1/2)mv^2 + U$ in the general form.

some 1.5 percent. From the beginning to the end. Why? Either because the spring is not linear, or there are other frictional effects, or else because we made a slight mistake somewhere in our work.

Is it reasonable that energy is constant over the oscillation of a mass on a spring? When we consider the mass from its balanced position, the restoring force is proportional to the displacement. In those circumstances, can we think out a law for conservation of energy? Yes, because the work done by such a force is

$$W = \int_{x_1}^{x_2} F(x) dx = \int_{x_1}^{x_2} -kx dx = -\frac{1}{2} kx^2. \quad (13.1)$$

Therefore, for a mass on a spring we have that the kinetic energy of the oscillating mass plus $\frac{1}{2} kx^2$ is a constant. Let us see how this works. We pull the mass down; it is extending still and so it's speed is zero. But at equilibrium, $x=0$, its potential energy is zero. Energy, the sum of kinetic energy, decreases. Now we release the mass and things begin to happen (the mass will be horizontal). But as any instant the kinetic plus potential energy must be a constant. For example, when the mass is on its way up, the initial equilibrium point, the position is again zero but the k when it's at its highest x^2 , and it's going more, v^2 it gets less k^2 , and so on. So the balance $1/2 v^2 + kx^2$ is maintained as the mass goes up and down. Thus we have another rule now. i.e., the potential energy for a spring is $\frac{1}{2} kx^2$. If the force is $F = -kx$.

13-4 Harmonics of Energy

Now we go on to the topic of a consideration of what happens when there are several numbers of objects. Suppose we have the complicated problem of many objects, which we label $i = 1, 2, 3, \dots$, all exerting gravitational pulls on each other. What happens then? We recall once that if we add the total energies of all the particles, we can always keep away all forms of particle motion gravitational potential energy. Gravity, the total is a constant.

$$\sum_i \text{kinetic}_i = \sum_i \text{pot}_i = \frac{\text{constant}}{r_{ij}} \quad \text{since} \quad (13.1)$$

How do we prove it? We differentiate both sides with respect to time and get $\dot{m}_i \ddot{r}_{ij}$. We're not differentiating pot_i , we're taking the derivative of the velocity. So it's the forces, just as in Eq. (13.1). We replace those forces by the law of force that we know from Newton's law of gravity and then we notice that what is left is the same as the time derivative of

$$\sum_i \frac{G m_i m_j}{r_{ij}},$$

The time derivative of the kinetic energy is

$$\begin{aligned} \frac{d}{dt} \sum_i \text{kinetic}_i &= \sum_i m_i v_i \cdot \frac{dv_i}{dt} \\ &= \sum_i \dot{m}_i v_i \\ &= \sum_i \left(\sum_j \frac{-G m_i m_j}{r_{ij}^2} \right) v_i. \end{aligned} \quad (13.2)$$

The time derivative of the potential energy is

$$\frac{d}{dt} \sum_i \text{pot}_i = \sum_i \left(\frac{-G m_i m_j}{r_{ij}^2} \right) \left(\frac{dr_{ij}}{dt} \right).$$

But

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2},$$

so that

$$\begin{aligned}\frac{dr_{ij}}{dt} &= \frac{1}{2r_{ij}}[2(r_i - r_j)\left(\frac{dv_i}{dt} - \frac{dv_j}{dt}\right) \\ &\quad - 2v_i \cdot v_j\left(\frac{dv_i}{dt} - \frac{dv_j}{dt}\right) \\ &\quad + 2(r_i - r_j)\left(\frac{dz_i}{dt} - \frac{dz_j}{dt}\right)] \\ &= r_{ij} \cdot \frac{v_i - v_j}{r_{ij}} \\ &= r_{ij} \cdot \frac{v_i}{r_{ij}} + r_{ij} \cdot \frac{v_j}{r_{ij}},\end{aligned}$$

since $v_{iz} = -v_{jz}$ while $r_{iz} = r_{jz}$. Thus

$$\frac{d}{dt} \sum_{1 \leq i < j \leq n} \frac{Gm_i m_j}{r_{ij}} = \sum_{i,j} \left[\frac{(mv_i m_j v_{iz})_i v_j}{r_{ij}^2} + \frac{(Gm_i m_j v_{iz})_i v_j}{r_{ij}^2} \right]. \quad (13.16)$$

Now we must note carefully what \sum , \sum_i , and $\sum_{i,j}$ mean. In Eq. (13.15), $\sum_i \sum_j$ means that i takes on all values $i = 1, 2, 3, \dots$ in turn, and for each value of i , the index j takes on all values except i . Thus if $i = 3$, j takes on the values $1, 2, 4, \dots$.

In Eq. (13.16), on the other hand, \sum means that given values of i and j occur only once. Thus the particle pair 1 and 2 contributes only one term to the sum. To keep track of this, we might agree to let i range over all values $1, 2, 3, \dots$, and for each i let j range only over values greater than i . Thus if $i = 1$, j could only have value 2, 3, 4, ... But we notice that for each i, j value there are two contributions to the sum, one involving v_i and the other v_j , and that these terms have the same superposition as that of Eq. (13.14), where all values of i and j (except $i = j$) are taken in the sum. Therefore, by matching the terms one by one, we see that Eqs. (13.16) and (13.15) are precisely the same, but of opposite sign, so that the time derivative of the kinetic plus potential energy is indeed zero. Thus we see that, for many objects, the kinetic energy is the sum of the contributions from each individual object, and that the potential energy is also simple, it being also just a sum of contributions, the energies between all the pairs. We can understand why it should be the energy of every pair this way: suppose that we want to find the total amount of work that must be done to bring the objects to certain distances from each other. We may do this in several steps, bringing them in from infinity where there is no force, one by one. First we bring in number one, which requires no work, since no other objects are yet present to exert force on it. Next we bring in number two, which does take some work, namely $W_{12} = -Gm_1 m_2 / r_{12}$. Now, and this is an important point, suppose we bring in the next object to position three. At any instant the force on number 3 can be written as the sum of two forces: the force exerted by number 1 and that exerted by number 2. Therefore the work done is the sum of the work done by both, because if F_3 can be resolved into the sum of two forces,

$$\mathbf{F}_3 = \mathbf{F}_{31} + \mathbf{F}_{32},$$

then the work is

$$\int \mathbf{F}_3 \cdot d\mathbf{s} = \int \mathbf{F}_{31} \cdot d\mathbf{s} + \int \mathbf{F}_{32} \cdot d\mathbf{s} = W_{13} + W_{23}.$$

That is, the work done is the sum of the work done against the first force and the second force, as if each acted independently. Proceeding in this way, we see that the total work required to assemble the given configuration of objects is precisely the value given in Eq. (13.14) as the potential energy. It is because gravity obeys the principle of superposition of forces that we can write the potential energy as a sum over each pair of particles.

13-4 Gravitational field of large objects

Now we wish to calculate the fields which are met in a few physical circumstances involving distributions of mass. We have not so far considered contributions of mass, only of force, so it is interesting to consider the forces which they can produce by more than just one particle. First we shall find the gravitational force due to a mass that is produced by a plane sheet of constant density in extent. The force on a unit mass at the given point P , produced by this sheet of mass (Fig. 13-5), will of course be directed toward the sheet. Let the distance of the point from the sheet be a , and let the constant mass per unit area of the sheet sheet be σ . We shall suppose it to be constant. This is a uniform sheet of mass σa^2 . Now, when small dudm is produced by the mass dm lying between p and $p + dp$ from the point C of the sheet nearest point P . Answer: $dC = Gdm/r^2$. But this force is directed along r , and we know this, only the component of dC we want when we add all the little vector dC 's to produce C . The y -component of dC is

$$dC_y = G \frac{dm}{r^2} \cos \theta = G \frac{dm}{r^2},$$



Fig. 13-5. The gravitational force F on a mass m produced by an infinite plane sheet of matter.

Now all contributions which are at the same distance r from P will yield the same dC_y , so we may at once write for dm the total mass in the ring between p and $p + dp$, namely $dm = \sigma 2\pi p dp$ ($2\pi p dp$ is the area of a ring of radius p and width dp , if $dp \ll p$). Thus

$$dC_y = G\sigma 2\pi p \frac{dm}{p^2}.$$

But, since $r^2 = p^2 + a^2$, $dm = r dr$. Therefore,

$$C_y = 2\pi G\sigma \int_p^\infty \frac{dr}{r^2} = 2\pi G\sigma \left(\frac{1}{p} - \frac{1}{r} \right) = 2\pi G\sigma. \quad (13.7)$$

Thus the $G\sigma$ is independent of distance a . Why? Have we made a mistake? One might think that the farther away we go, the weaker the force would be. But no! If we are close, most of the matter is pulling at an unfavorable angle; if we are far away, most of the matter is situated quite favorably to exert a pull toward the plane. A simple diagram the matter which is most effective lies in a certain range. When we are farther away the force is smaller by the inverse square, but in the same range, in the same angle, there is much more matter, larger by just the square of the distance. This analysis can be made rigorous by just noticing that the differential contribution to any given force is in fact independent of the distance because of the reciprocal variation of the strength of the force from a given mass, and the amount of mass included in the force, with changing distance. The force is not really constant, of course, because when we go on the other side of the sheet it becomes $-$ in sign.

We have also, in effect, solved an electrical problem: if we have an electrically charged sheet, with an infinitesimal of charge per unit area, then the electric field at a point a is, outside the sheet, $E = \sigma/2\epsilon_0$, and is in the outward direction of the sheet if positively charged, and inward if the sheet is negatively charged. To prove this, we merely note that the gravity, $(2\pi G\sigma)/a$, is due to the conductivity.

Now suppose that we have two plates, with a positive charge $+q$ on one and a negative charge $-q$ on another at a distance a from the first. What is the field? Outside the two plates it is zero. Why? Because one side is at the other repels the force being withdrawn of charge, so that it has to move out. Also, the force between the two plates is clearly taken as q/q as two sheets one side, namely $E = q/4\pi\epsilon_0 a^2$, and is directed from the positive plate to the negative one.

Now we come to a most interesting and important problem, whose solution we have been assuming all the time, but now, that the force produced by the earth is a pull, and the surface of the earth it is the same as if all the mass of the earth were located in its center. The validity of this assumption is not obvious, because when we are close, some of the mass is very close to us and some is further away.

and so on. When we add the effects all together, it seems a miracle that the net force is exactly the same as we would get if we put all the mass in the middle!

We now decide to be convinced of this miracle. In order to do so, however, we shall consider a thin spherical shell instead of the whole earth. Let the total mass of the shell be m , and let us calculate the potential energy of a particle of mass m' at a distance R away from the sphere (Fig. 13-6) and show that the potential energy is the same as it would be if the mass m were a point at the center. (The potential energy is easier to work with than is the field because we do not have to worry about angles; we merely add the potential energies of all the pieces of mass.) If we call x the distance of a certain plane section from the center, then all the mass that is in a slice dx is at the same distance x from P , and the potential energy due to this ring is $-Gm'm\mu dx/x$. How much mass is in the small slice dx ? An amount

$$dm = 2\pi\mu x dx = \frac{2\pi\mu x dx}{\sin\theta} = 2\pi\mu dx,$$

where $\mu = m/(4\pi R^2)$ is the surface density of mass on the spherical shell. (It is a general rule that the area of a zone of a sphere is proportional to its axial width.) Therefore the potential energy due to dm is

$$\delta W = -\frac{Gm'm\mu}{x} = -\frac{Gm'm\mu dx}{x}.$$

But we see that

$$\begin{aligned} r^2 &= x^2 + (R-x)^2 = x^2 + x^2 - 2Rx = R^2 - 2Rx \\ &= x^2 + R^2 - 2Rx. \end{aligned}$$

Thus

$$2r dx = -2R dx$$

so

$$\frac{dx}{r} = \frac{dx}{R}.$$

Therefore,

$$\delta W = -\frac{Gm'm\mu dx}{R},$$

and so

$$\begin{aligned} W &= -\frac{Gm'm\mu}{R} \int_{R-x}^{R+x} dr \\ &= -\frac{Gm'm\mu}{R} dx = -\frac{Gm'(4\pi x^2\mu)}{R} \\ &= \frac{Gm'm}{R}. \end{aligned} \tag{13.18}$$

Thus, for a thin spherical shell, the potential energy of a mass m' external to the shell, is the same as though the mass m of the shell were concentrated at its center. The earth can be imagined as a series of spherical shells, each one of which contributes an energy which depends only on its mass and the distance from the center; adding them all together we get the total mass, and therefore the earth acts as though all the material were at the center.

But notice what happens if our point is on the inside of the shell. Making the same calculation, but with P on the inside, we still get the difference of the two r 's, but now in the form $x + R$ ($x = R$) = $2R$, or twice the distance from the center. In other words, W comes out to be $W = -Gm'm/a$, which is independent of R and independent of position, i.e., the same energy no matter where we are inside. Therefore no force; no work is done when we move about inside. If the potential energy is the same no matter where an object is placed inside the sphere, there can be no force on it. So there is no force inside, there is only a force outside, and the force outside is the same as though the mass were all at the center.

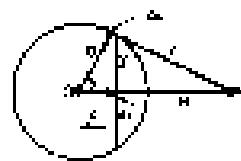


Fig. 13-6. A thin spherical shell of mass or charge.

Work and Potential Energy (continued)

14.1 Work

In the preceding chapter we have presented a great many new ideas and results that play a central role in physics. These ideas are so important that it seems worthwhile to devote a whole chapter to a closer examination of them. In the present chapter we shall not repeat the "proofs" or the specific tricks by which the results were obtained, but shall concentrate instead upon a discussion of the ideas themselves.

In learning any subject of a technical nature where mathematics plays a role, one is confronted with the task of understanding and solving, usually, in the extremely large body of facts and laws. Let us imagine by way of example something which can be "proved" or "shown" to rest between them. It is easy to confuse the proof itself with the result which it establishes. Clearly, the important thing is to know and to remember is the relationship, not the proof. In any particular case, however, we can safely say that it can be shown that "such and such is true," if we can show it. In almost all cases, the particular "such" that is most interesting, that is, the interest lies in its being done quickly and easily on the computer or on paper, and so that it will be as simple looking as possible. Consequently, the proof may look deceptively simple. This is fact. In fact, you might have worked for hours trying different ways of calculating the same thing until you found the easiest way, so as to be able to show that it can be done in the shortest amount of time. The thing to be remembered when seeing a proof, is that the proof is off, but for the it can be shown that such and such is true. Of course, if the proof involves some mathematical procedure or "tricks" that one has not seen before, then you should try to get back to the basic theory, turn to the mathematics, and re-investigate.

It is certain that at all the examinations that I have taken up to now, as well as this, nothing has been remembered from me from where the author studied freshman physics. Quite the contrary, he merely remembers what was done, since such is true, and to explain how it can be shown he uses a demonstration of the human kind. Anyone who has really learned a subject ought to know how to follow a similar procedure, but it is not necessarily the point. That is why, in this section, we shall avoid the words of the various statements made previously and simply summarize the results.

The first idea that has to be discussed is work when a force. The classical word "work" is not the word in the ordinary sense of "the labor of the workman," but by a different idea. Physical work is expressed as $\int F \cdot ds$, called "the line integral of F dot ds ," which means that if the force F , for instance, is in one direction and the object on which the force is working is displaced in a certain direction, then only the component of force in the direction of the displacement counts as work. If the direction of the force were constant and the displacement were a straight line, then the work done in moving the object distance through that distance is only the component of force along ds in ds . That is, "Work times Distance" here, as really means only the component of force in the direction of the displacement times the displacement. As you, naturally, the component of displacement in the direction of force times F . It is evident that a work whatever is done by a force which is at right angles to the displacement.

Now if the vector displacement is resolved into components. In other words, if the actual displacement is ds and we write to consider it "vertically" as a component of displacement dy in the y -direction, dz in the z -direction, and dx

14.1 Work

14.2 Conservation motion

14.3 Conservative forces

14.4 Nonconservative forces

14.5 Potentials and fields

In an equilibrium, when the work done in carrying an object from one place to another can be calculated in three parts, i.e. calculating the work done along x , along y , and along z . The work done in going along x involves only the component of force, namely F_x , and so on, so the work is $F_x \Delta x - F_y \Delta y - F_z \Delta z$. When the force is not constant, and we have to consider curved motion, then we must follow the path in a series of little steps, and the work done in carrying the object along x , y , z would take us until as it goes to zero. This is the meaning of the "line integral."

Everything we have just said is contained in the formula $W = \int F \cdot ds$. It is all very well to say that it is a mathematical formula, but it is important to understand what it means, so what some of the consequences are.

The word "work" in physics has a meaning quite different from that of the word as it is used in ordinary circumstances that it may be close, yet still. But there are some peculiar circumstances in which it appears not to be the same. For example, according to the physical definition of work, it does not have a hundred pound weight off the ground for a while. Is it doing no work? None of us even one knows. Let us begin to see why this, and because harder, as if he were running up a flight of stairs. Yet running upstairs is considered as doing work for running downstairs, we do not see all the work according to physics. In simply holding an object in a fixed position, no work is done. I think the physical definition of work differs more in psychological definition, for reasons we shall shortly examine.

It is a fact that when one holds a weight he does no "physiological" work. Why should he sweat? Why would he need to concentrate hard to hold the weight up? Why is the machinery inside him operating at full force, giving his heart, his lungs up? Actually, he might as well be holding up with an effort by suspending it on a table, then the table, quietly and calmly, without the supply of energy, is able to hold it. This is the way of the static weight. The physiological situation is something like the following. There are two kinds of muscles in the human body and, in other animals, too, called smooth muscle and the type of muscle we have in our arms, for example, which is under voluntary control. The other kind, called smooth muscle, is like the muscle of the heart muscle, or the stomach, the greater omentum muscle that closes the fist. The smooth muscles are not very strong, but they can make their job. If we try to hold, for example, a six pound fish to hang, to catch it, it will hold a position under load for hours and hours without getting tired, because it is very much easier to not having to a weight of "pounds" into a certain position, and the skeleton just does this, keeping it with no work being done, more work being performed by the brain. The fact then we have to generate effort to hold up a weight is simply due to the design of smooth muscle. What happens is that when we use voluntary muscles then, the brain gives a little twitch and then relaxes, so that when we hold something the enormous volleys of nerve impulses are coming to the muscle. Large numbers of impulses are maintaining the weight while the brain relaxes. We can see this, of course, when we hold a heavy weight and get tired, we begin to shake. The reason is that the volleys are coming continually, and the muscle is tired and not reacting, restlessly... Why such a inefficient system? Which we know exactly why, but evolution has not been able to evolve just smooth muscle. Smooth muscle would be much more effective for holding up weights because you could just stand there and it would kick in, there would be no work, the less work energy would be required. However, it has the disadvantage that it is very slow-operating.

Returning now to physics, we may ask why we want to calculate the work done. The answer is that it is interesting and useful to do so, since the work done on a particle is the resultant of all the forces acting on it. It's easily related to the change in kinetic energy of that particle. That is, if an object is being pushed, it picks up speed, and

$$K.E. = \frac{1}{2} m v^2$$

14-1 Constrained motion

Another interesting feature of forces and work is this. Suppose that we have a string in a curved path, and a particle that must move along the track, for some reason. If we now have a particle, with a string and a weight, the string connects the weight to some fixed point about the pivot point. The pivot point may be changed by pulling the string "tilt a yard," so that the part of the weight is along two circles of different radii. These are examples of what we call *geometric constraint*.

In motion with a fixed track, constrained motion is done by the constraint because the forces of constraint are always at right angles to the motion. By the "Principle of Least Action," the same thing is true which is applied to the object directly by the one pulling itself—the tension force with the tends or the tension in the string.

The forces involved in the motion of a particle on a slope moving under the influence of gravity are quite complicated since there is a normal force, a gravitational force, and so on. However, if we take into account the law of conservation of energy and the law involving work alone, we get the right result. This seems rather strange because it is not strictly the right way to do it; we should use the resulting force. However, the work done by the gravity force, force \mathbf{F}_g , will turn out to be the change in the kinetic energy, because the work done by the conservative force of the forces is zero (Fig. 14-1).

The important feature here is that the forces can be analyzed as the sum of the various "pieces." Then the work done by the resultant force in going along a certain curve is the sum of the works done by the various "independent" forces, into which the forces are resolved. Thus if we analyze the forces as being the vector sum of several effects of gravitational, normal, frictional, etc., then, as in the assumption of a frictional and the "parametrization" of all forces in any other way than we wish to split them, then the work done by the net force is equal to the sum of the works done by all the parts in it, which we have divided the force in making the analysis.

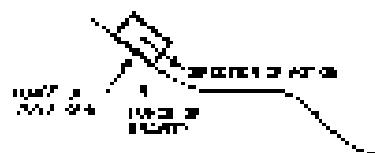


Fig. 14-1. Forces acting on a sliding body (no friction).

14-2 Conservative forces

To introduce these we start with a lot of gravity, for example, which have a very remarkable property which we call "conservative" (we will find this involved, it is again one of those crazy words!). If we calculate how much work is done by a force in moving an object from one point to another along some curved path, in general the work depends upon the curve, but in gravity's case it does not. If it does not depend upon the curve, we say that the force is a conservative force. In other words, if the integral of the force times the distance in going from position 1 to position 2 in Fig. 14-2 is evaluated along curve 4 and then along 5, we get the same number of joules, and this is true for the pair of points in every case, and if the same procedure is taken with pairs of points we have them say the force is conservative. In such circumstances, the work integral along from 1 to 2 can be evaluated in a simple manner, and we can give a formula for the result. Ordinarily it is not this easy, because we also have to specify the curve, and why we have a case where the work does not depend upon the curve, then, of course, the work depends only upon the positions of 1 and 2.

To demonstrate this, let's consider the following. We take a "fixed" point P , at an arbitrary location (Fig. 14-3). Then, the work U_{12} (along curve 1) from 1 to 2, which we want to calculate, can be evaluated as the work done in going from 1 to P plus the work done in going from P to 2, because the forces are conservative and the work does not depend upon the curve. Now, the work done in going from position P to a particular position of 2 is a function of that position of 2 itself. Of course it really depends on P also, but we have an arbitrary point P does not妨碍 our analysis. If that is done, then the work done in going from point P to point 2 is some function of the final position of 2. It depends upon where 2 is, if we go somewhere else, we get a different answer.

We shall call this function of position = $U(x_2, y_2, z_2)$, and when we want to refer to some particular point 2 whose coordinates are (x_2, y_2, z_2) , we shall write $U(2)$.



Fig. 14-2. Possible paths between two points in a field of force.

is an abbreviation for $\mathbf{U}(x_1, y_1, z_1)$. The work done in going from point 1 to point P can be written also by using the other form along the integral, referring to the $d\mathbf{r}_1$. That is, the same curve in going from 1 to P is known. The work done in going from the point P to 1:

$$\int_1^P \mathbf{F} \cdot d\mathbf{r} = \int_{\mathbf{r}_P}^{\mathbf{r}_1} (-\mathbf{F}) \cdot d\mathbf{r} = - \int_P^1 \mathbf{F} \cdot d\mathbf{r}$$

The total work done in going from P to 1 is $-\mathcal{U}(P)$, and from P to 1 the work is $\mathcal{U}(P)$. Therefore the integral from 1 to P is equal to $-\mathcal{U}(P)$ plus $(-\mathcal{U}(P))$ backwards, i.e. $-\mathcal{U}(P) + \mathcal{U}(P)$.

$$\begin{aligned} \mathcal{U}(P) &= - \int_P^1 \mathbf{F} \cdot d\mathbf{r} = \mathcal{U}(1) = - \int_1^P \mathbf{F} \cdot d\mathbf{r} \\ &= \int_1^P \mathbf{F} \cdot d\mathbf{r} = \mathcal{U}(1) - \mathcal{U}(P) \end{aligned} \quad (14.1)$$

The quantity $\mathcal{U}(P) - \mathcal{U}(1)$ is called the change in the potential energy, and we call \mathcal{U} the potential energy. We say, "when the object is located at position 2, it has potential energy $\mathcal{U}(2)$ and at position 1 it has potential energy $\mathcal{U}(1)$. If it is located at position P, it has zero potential energy." If we had used another potential, say \mathcal{U}' , instead of \mathcal{U} , we would get $\mathcal{U}'(P)$ and we shall leave it to you to convince yourself that the potential energy is changed only by the addition of a constant. Since the numerical value of energy depends only upon changes, it does not matter if we add a constant to the potential energy. Thus the point 1 is arbitrary.

Now, we have the following two propositions: (1) that the work done by a conservative plus the non-conservative force of the problem, i.e., (2) and non-conservative forces, is zero; (2) that the work done is minus the change in a function \mathcal{U} which we call the potential energy. As a consequence of these two, we arrive at the proposition that if any conservative force has no non-conservative part, its potential energy is conserved invariant.

$$\mathbf{F} + \mathbf{V} = \text{constant}. \quad (14.2)$$

Let us now discuss the formulae for non-potential energy for a number of cases. If we have a gravitational field, it follows, if we are not going to neglect non-conservative forces, the work done is minus the change in a function \mathcal{U} which we call the potential energy. As a consequence of these two, we arrive at the proposition that if any conservative force has no non-conservative part, its potential energy is conserved invariant.

$$\mathbf{F}_G = m\mathbf{g}, \quad (14.3)$$

and the point P which corresponds to zero potential energy happens to be any point in the plane $z = 0$. We could also choose $z = L$ for, the potential energy is zero $= 0$ if we had chosen to—all the results would, of course, be the same in our analysis except that the value of the potential energy at $z = 0$ would be $-mgL$. It may be an infinitesimal, because only difference in potential energy is small.

The energy needed to complete a linear spring a distance x from an equilibrium point is

$$\mathcal{U}(x) = \frac{1}{2}kx^2 \quad (14.4)$$

and the value of potential energy is at the position $x = 0$, the equilibrium position of the spring. A given weight will stay constant as we wish.

The potential energy of a capacitor at point x is, as we know, a capacitor (see Eq. 1.1)

$$\mathcal{U}(x) = Q^2/2C \quad (14.5)$$

The constant has been chosen here so that the potential is zero at infinity. Of course, this choice certainly applies to electric charges, because it is the same law.

$$Q_{\infty} = q_1 q_2 / 4\pi\epsilon_0 r_{\infty} \quad (14.6)$$

Now let us actually use all of these formulas, to see whether we understand what "means". Please! This has always been to show a rocket away from the

can't go out far if it leaves? Because if the kinetic plus potential energy must be constant; when it "leaves," it will be no longer of finite energy, and since it is just barely going, it can't be radius of the curve, size of its mass. The kinetic plus potential energy is then initially given by $\frac{1}{2}mv^2 + GMm/r$. On the end of the motion the two energies must be equal. The kinetic energy is taken to be zero at the end of the motion, because it is supposed to be just barely falling away at essentially zero speed, and so potential energy is exactly divided by infinity, which is zero. So everything is now in our control that $v^2 = rGM$; the square of the velocity must be $2GM/r$. Our GGM is what we call the acceleration of gravity g . Thus

$$v^2 = 2gr$$

At what speed must a satellite travel, in order to keep going around the earth? We worked this out long ago and found that $v^2 = GM/r$. Therefore, if you want to leave the earth, you must $\sqrt{2GM/r}$ be velocity you need to just go around the earth near the surface. We know, in other words, *not to touch things*, because initial speed is the square of the velocity to leave the earth or the distance around it. Therefore, the first thing that we have to do is to find v in m/s we have to go and to go around the earth, which requires a speed of five miles per second. This next time, we have to send a satellite away from the earth perpendicularly. It's required twice the energy, or about seven miles per second.

Now, continuing our discussion of the characteristics of potential energy, let us consider the interaction of two molecules, or two atoms, like oxygen atoms for instance. When they are very far apart, the force is one of attraction, which scales up like inverse seventh power of the distance, and when they are close together there is a very large repulsion. If we integrate the inverse seventh power to find the work done, we find that the potential energy V , which is a function of the center distance between the two oxygen atoms, varies as the inverse sixth power of the distance r in general.

If we sketch the curve of the potential energy $V(r)$ as in Fig. 14-2, we thus start at the point $r = \infty$ with an infinite negative potential, but if we come in sufficiently far, we reach a point where there is a minimum in potential energy. The minimum in potential energy is $-V_0$. If r comes closer to r_0 we start at V_0 and move a small distance, a very small distance, the next time, which is the change in potential energy when we move this distance, is nearly zero, because there is very little change in potential energy at the bottom of the curve. Thus the r_0 is the equilibrium point, and V_0 is the equilibrium potential. Another way to see this is the equilibrium point is that it takes work to move away from it in either direction. When the two oxygen atoms are seated down, so that no more energy can be extracted from the force between them, they are in the lowest energy state, and they will be. The expansion of this gas is why an oxygen cylinder looks when it is cold. When we heat it up, the atoms shake and move farther apart, and we can extract less and less work, but it still requires a certain amount of work to move, which is the potential energy difference between $r = r_0$ and $r = \infty$. When we try to push two masses very close together the energy goes up very rapidly, because the potential V of r .

The reason we bring this up is to make the idea of potential energy particularly suitable for mechanics, where the idea of energy is most natural. We find that although forces and velocities "divided" and "supplied" when we consider the many independent forces between particles, they do not represent any of the energy except potentials. Therefore we have curves of potential energy of one form, velocities another, but we can't do the same sort of curve for the forces between two masses, even if we try. Very people who are doing mechanics are thinking in terms of energy rather than of forces.

Next we note that if you're moving along a curve in an object of finite mass, then the maximum energy of the object is the sum of the potential energies from each of the separate forces. This is the same proposition that we mentioned before: because if the forces are proportional and additive in form, then the total force is the sum of the forces due to the partial

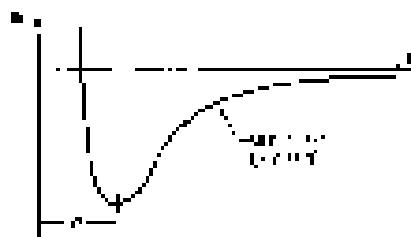


Fig. 14-2. The potential energy between two atoms as a function of the distance between them.

It does, and it can't be defined in any other terms. Changes in the potential energy is a *vector* of force *separately*. This total potential energy is the sum of all the little pieces.

We could generalize this to the case of a system of many objects, like a galaxy with many stars, like Jupiter, Saturn, Uranus, etc., oxygen, nitrogen, carbon, etc., which are acting with respect to one another in proportion to forces all of which are conservative. In these circumstances, the kinetic energy in the entire system is simply the sum of the kinetic energies of each of the particular atoms or planets or whatever, and the potential energy of the system is the sum, over the pairs of particles, of the potential energy of each pair. It's a single pair, so "length the vector" were not there. (This is not really true for Coulomb's forces, and the formula is somewhat more complicated; it certainly is not for Newtonian gravitation, and it is less accurate sometimes by molecular forces. For instance, since there's a repulsive energy, you'd have to make a more complex expression of the positions of the atoms than simply a sum of forces from particle 1, the special case of gravity, though...) The potential energy is the sum, overall, the same kind of - ∞ -infinity, as was indicated in Eq. (11.10). Equation (11.10) expressed mathematically the obvious consequence, that the total kinetic energy plus the total potential energy does not change with time. As the various planets travel around, and turn and twist and so on, as we calculate the total kinetic energy and the total potential energy we find that the total remains constant.

14-4 Nonconservative forces.

We have spent a considerable time discussing conservative forces, what about nonconservative forces? We said "like a dog: this is his domain, and outside the domain no nonconservative forces." As a matter of fact, the fundamental forces in nature appear to be conservative. It is *not* a consequence of Newton's laws. In fact, in Isaac Newton I may know, the forces would be more conservative, as Einstein apparently is. When we say "inertial reference frame," we are taking a modern view, in which it has been observed that all the deep forces, the forces between the galaxies, the major fundamental ones, are conservative.

In fact example, we analyze a system like the planet, galaxy or star cluster that we saw a picture of with the dependence of other all interacting, then if we do the for the total potential energy is simply the sum plus another term, etc., a minus sign, all pair of stars, and the result will be if the sum of the kinetic energies of all the individual stars. But the galaxies cluster as a whole is itself in space here, and the total forces of gravity from it and the other galaxies could be the sum of each single object. Then if forces were applied to it, some of those forces might add up causing it to move as a whole, and we would say "The center of the whole thing moves." On the other hand, some of the forces can be, so to speak, "wasted" in increasing the kinetic or potential energy of the "particle" alone. I.e., we suppose, for instance, the the action of these forces causes the whole cluster and makes the particles move faster. The total energy of the whole thing is still conserved, but not from the particle with a center of mass which is just the combination of previous kinetic, and just thinking of the kinetic energy of the motion, of the whole object as though it were a single particle, a weak assumption, this energy is not conserved, but this is due to a lack of appreciation of what it is that we see. And that is more or less the total energy of the system, kinetic plus potential, is a constant when we look closely enough.

When we study matter in the finest detail at the atomic level, it is not always easy to the total energy of a thing for two parts, kinetic energy and potential energy, and such separation is not always necessary. It is always always possible to do it, and it is very frequently necessary, possible and for the genuine electrostatic energy of the world is constant. Thus the total potential plus-kinetic energy, like the kinetic plus U constant, and if the "whole" is a system of isolated particles, the energy is constant if there are no external forces. But as we have seen, some of the kinetic and potential energy of a thing may be lost, for instance like internal molecular motions in the solid. So we do not make it. We know that in a glass of water everything is jiggling around, all the atoms are moving

of the atoms there is a certain kind of energy inside which we usually may not pay any attention to. We do not notice the motion of the atoms, which produces heat, and so we do not call it kinetic energy, but heat is certainly kind of one type. Internal potential energy may also be in the form of the form of electric energy; when we move positive charge is liberated because the potential energies of the atoms at the new position or arrangement are lower than in the old one. In fact, it is not entirely possible to beat "heat" being some kind of energy, for a time the potential goes on and goes over to chemical energy, so we sum the two together and say that the total kinetic and potential energy inside an object is partly heat, partly chemical energy, and so on. Anyway, all these different forms of internal energy are sometimes considered to "heat" energy in the way described above, so it will be make clear when we study thermodynamics.

As another example, when friction is present, it is not true that kinetic energy is lost, even though a sliding object stops and the kinetic energy seems to be lost. The kinetic energy is not lost because, of course, the atoms inside are juggling with a larger amount of kinetic energy than before, and although we cannot see that, we can measure it by determining the temperature. Of course if we disregard the heat energy, then the conversion of energy between will appear to be loss.

Another situation, in which energy conservation appears to be false is when we study only part of a wave. Below, the conservation of energy theorem will appear to be false, something is interacting with everything else in the outside and we neglect to take that interaction into account.

In classical mechanics potential energy involved only gravitation and electricity, but now we have nuclear energy and other energies also. Right, for example, weak forces are a new type of energy in the classical theory, but we can do, if we want, imagine that the energy of π^+ is the kinetic energy of a pion, and then our formula (14-1) would go like right:

14-5 Potentials and Fields

We shall now discuss a bit of the theory associated with potential energy and with the idea of a field. Suppose we have two large objects A and B and a third very small one which is attracted gravitationally by the two, with some resultant force \mathbf{F} . We have already noted in Chapter 13 that the gravitation of objects A and B can be written as the sum of three numbers G , each of which is dependent only upon the position of the particle:

$$\mathbf{F} = \mathbf{qC}$$

We can calculate gravitation, that is, by calculating that there is a certain vector \mathbf{C} at every position in space which "feels" up to a mass which we may place there, but which is free itself where you want to supply a mass for it to "feel" on it. \mathbf{C} has three components, and each of these components is a function of (x, y, z) , a location of position in space. Such a thing we call a field, and we say that the objects A and B generate the field, i.e., they "cause" the vector \mathbf{C} . When an object is placed below the horizon it is due to its mass times the value of the field vector at the point where the object is given.

We can also do the same with the potential energy, since the potential energy, defined as V of (14-1), can be written as $V = q\mathbf{q}\cdot\mathbf{C}$ (that is, the potential of the field). With a mass charge q at (x, y, z) the potential energy $V(q, p)$ of an object located at a point (x_1, y_1, z_1) in space can be written as a third function Φ which we may call the potential Φ . The integral $\int \mathbf{q}\cdot d\mathbf{s} = -q\int q\cdot d\mathbf{s} = -V$. There is only a plus factor between the two:

$$V = -\int q\cdot d\mathbf{s} = -q\int (x, y, z)\cdot d\mathbf{s} \quad (14-2)$$

By having the function $\Phi(x, y, z)$ at every point in space, we can immediately calculate the potential energy of a system at any point in space, namely, $V(x_1, y_1, z_1)$ and (x_2, y_2, z_2) —either a real business concern. But it is not really trivial, because it is somewhat difficult to describe the field by giving the field in

of Φ everywhere in space instead of having to give Φ a name and having to write three complicated equations of a vector function, we can also list all these for Φ in one. Furthermore, it is much easier to calculate Φ at any given component of \mathbf{E} when the field is produced by a number of charges, because the potential Φ is a scalar and is always odd, without worrying about direction. Also, the total Φ can be integrated easily from Φ , as we shall shortly see. Suppose we have point charges q_1, q_2, \dots, q_n at points $\mathbf{r}_1, \mathbf{r}_2, \dots$ and we wish to know the potential Φ at some arbitrary point \mathbf{r} . This is simply the sum of the potentials due to the individual charges taken singly and

$$\Phi(\mathbf{r}) = \sum -\frac{q_i}{r_i}, \quad i = 1, 2, \dots \quad (14.8)$$

In the last chapter we used this formula. But the potential is the sum of the potentials from all the different objects, to calculate the potential due to a spherical shell of charge by adding the contributions to the potential at a point from all parts of the shell. An example of this calculation is shown graphically in Fig. 14.4. It is negative, along the radius r to $r = \infty$ and decreasing as r increases. The radius a and then is constant inside the shell. Consider first that the potential is $-q/a$, where a is the radius of the shell, which is exactly the same as it would have been if all the q 's were located at the center. But it is not appropriate now by the same for inside the shell. Its potential is known to be $-q/a$, and is a constant. Then the potential is constant, there is no force, or rather the potential energy is constant; there is no force, because if we move an object from one place to another anywhere inside the sphere the work done by the force is exactly zero. Why? Because the work done in moving the object from one place to the other is equal to minus the change in the potential energy, so the work required to change the energy of the potential is zero. But the potential energy of the shell at any two points inside, so there is zero change in potential energy, and therefore the work is zero in going between any two points inside the shell. The only way the work can be zero for all directions of displacement is that there is no force at all.

This gives us a clue as to how we can obtain the force at the field given the potential of charge. We are supposed that the potential energy of an object is known at the position (x, y, z) and we want to know what the force on the object is. It will not do to know the potential at only this one point, as we shall see; it requires knowledge of the potential at neighboring points as well. Why? They can be calculated by summation of the forces. If we can do this, of course, we can also find the x and y components, and we will then know the whole force. Now, if we want to move the object a small distance Δr in the x direction, the force on the object would be the change ΔU in the force times Δr . If Δr is sufficiently small, this should equal the change in potential energy in going from one point to the other:

$$\Delta U = -\Delta V = F_x \Delta x. \quad (14.9)$$

We have nearly used the formula $\Delta U = -\Delta V$, but for a very short while how we divide by Δx and Δr and that the force is

$$F_x = -\partial V / \partial x. \quad (14.10)$$

Of course this is not exact. What we really want is the limit of (14.10) as the distances Δx and Δr get smaller, because it is only *approximately* right in the limit of infinitesimals Δx . This we recognize as the derivative of V with respect to x , and we will be defining the force in some "mathematical" way. F depends on x, y, z , and θ , and the mathematicians have invented a different symbol to remind us to be very careful when we use the unitless θ as a function, so as to distinguish that we are considering the field \mathbf{E} around some point x, y, z every "angle" of θ is being simply made a "backwards 0," or $\bar{\theta}$. ($\bar{\theta}$ should have been used to the beginning of calculus because we always want to cancel that off but we never seem to notice it with θ .) They write $\partial V / \partial x$ for "force," in "mathematical" terms, if they want to be very careful, they put a bar under θ and a tilde over V in the notation $\partial \tilde{V} / \partial \bar{\theta}$, etc., etc.

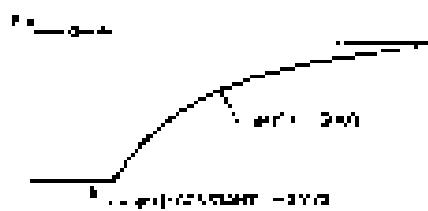


Fig. 14.4. Potential due to a spherical shell of radius a .

which we've "Take the derivative of U with respect to x , keeping y and z constant." Most often we leave out the remark about what is kept constant because it is usually evident from the context, so we usually do not use the $\partial/\partial x$ with the word x , instead always use $\partial/\partial x$ instead of d/dx because the latter is a derivative with some other variables kept constant. This is called a *partial derivative*; it is a derivative in which we vary only x .

Therefore, we find that the force in the x -direction is minus the partial derivative of U with respect to x :

$$F_x = -\partial U / \partial x \quad (14.12)$$

In a similar way, the force in the y -direction can be found by differentiating U with respect to y , keeping x and z constant, and the third component, of course, is the derivative with respect to z , keeping x and y constant:

$$F_y = -\partial U / \partial y, \quad F_z = -\partial U / \partial z. \quad (14.13)$$

This is the way to get from the potential energy to the force. We get the first from the potential in exactly the same way:

$$C_x = -\partial U / \partial x, \quad C_y = -\partial U / \partial y, \quad C_z = -\partial U / \partial z. \quad (14.14)$$

Incidentally, we don't mention this second notation, which we shall not actually use for quite a while; since C is a vector and has x , y , and z components, the symbolized $\partial U / \partial x$, $\partial U / \partial y$, and $\partial U / \partial z$ which produce the x , y , and z components are something like vectors. The xy -term, however, has an unusual significance, too, called "grad" or "gradient" which is a scalar quantity but an operator which makes a vector function scalar. It has the following "components": The x -component of this "grad" is $\partial/\partial x$, the y -component is $\partial/\partial y$, and the z -component is $\partial/\partial z$, and then we have the fact of writing our formulas this way:

$$\mathbf{F} = \mathbf{C} U, \quad \mathbf{C} = \text{grad } U. \quad (14.15)$$

Using Eq. 14.15 gives us a quick way of testing whether we have a true vector equation or not, but only Eq. 14.15 means precisely the same as Eqs. 14.12 and 14.13; it is just another way of writing them, and since we do not want to write three equations every time, we just write Eq. 15 instead.

One more example of fields and potentials has to do with the electrical case. In the case of electricity the force on a stationary object is the charge times the electric field: $\mathbf{F} = q\mathbf{E}$. (In general, of course, the x -component of force in an electrical problem is also a part which depends on the magnetic field.) It is easy to show from Eq. 13.10 that the force on a particle due to magnetic fields is always at right angles to its velocity, and also at right angles to the field. Since the force due to magnetism on a moving charge is at right angles to the velocity, no work is done by the magnetism on the moving charge because the motion is at right angles to the force. Therefore, in calculating the loss of kinetic energy in electric and magnetic fields we can disregard the contribution from the magnetic field, since it does not do any kinetic energy. We suppose that there is only an electric field. Then we can calculate the energy, or work done, in the same way as for gravity, and calculate a quantity ϕ which is minus the integral of $\mathbf{F} \cdot d\mathbf{s}$, first in the initial specified point in the path, where we make the calculation, and then the potential energy in an electric field is just charge times this quantity ϕ .

$$\phi(r) = \int_{r_0}^r \mathbf{E} \cdot d\mathbf{s}$$

Let us take, as an example, the case of two parallel metal plates, each with a surface charge of σ per unit area. This is called a parallel-plate capacitor. We learned previously that there is zero force outside the plates and that there is a uniform electric field between them, E , equal from 1 to 2 and of magnitude σ/ϵ_0 (Fig. 14-1). We would like to know how much work would be done in

carrying a charge from one plate to the other. The work would be the (force) · (the integral, which can be written as charge times the potential) value of units 1 coulombs times 1 volt = 1 joule.

$$W = \int_1^2 F \cdot ds = q(\phi_1 - \phi_2)$$

We can actually work out this integral because the force is constant, and if we call the separation of the plates d , then the integral is easy:

$$\int_1^2 F \cdot ds = W \int_{\phi_1}^{\phi_2} dx = Wd.$$

The difference in potential, $\phi_2 - \phi_1/d$, is called the voltage difference, and V is measured in volts. When we say a pair of plates is charged to a certain voltage, what we mean is that the difference in electrical potential of the two plates is equal to many volts. (For a capacitor made of two parallel plates carrying a surface charge (i.e., the voltage, or difference in potential, of the pair of plates is $V = Q/C_0$).

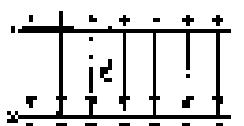


Fig. 14-5. Field between parallel plates.

The Special Theory of Relativity

15-1 The principle of relativity

For over 300 years no experiments of motion originated by Newton were performed in space; he could conceive, and the first time that either or in these laws was disagreed, one way to correct it was to disprove. Both the error and its correction were discovered by Galileo in 1602.

Newton's Second Law, which we have expressed by the equation

$$F = m a = m \frac{dv}{dt},$$

was tested with the first assumption that it is consistent, but we now know that this is not true, and can measure how it truly increases with velocity. In this it increased to much more than the value

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}} \quad (15.1)$$

where the "rest mass" m_0 represents the mass of a body that is not moving and c is the speed of light, which is about 3×10^8 cm/sec 2 or about 186,000 miles/sec.¹

For those who want to learn just one problem that does not cause problems, that is all there is to the theory of relativity—it just changes Newton's law by introducing a correction factor to the mass. From the formula, we find it easy to see that this mass increases very small in ordinary circumstances. If the velocity is even as great as that of a satellite, which goes around the earth at 5 mi/sec, then $m/m_0 = 5/(5c/300,000)$, putting this value into the formula shows that the correction to the mass is only one part in ten billion billion, which is easily impossible to observe. Actually, the correctness of the formula has been simply confirmed by the observation of many kinds of particles, moving at speeds close to exactly the speed of light. However, because the effect is really so small, it took a hundred years for it to be discovered theoretically before it was discovered experimentally. Empirically, at a sufficiently high velocity, the effect is very large, but it was not discovered that way. Therefore it is interesting to see how the Einsteins involved to develop a modification (at the time when it was first discovered) was brought up by a combination of experiments and physical reasoning. Contributions to the discovery were made by a number of people, the final result of whose work was Einstein's discovery.

The second major Einstein theory of relativity, this chapter is connected with the Special Theory of Relativity, which dates from 1905. In 1915 Einstein published an additional theory, called the General Theory of Relativity. This latter theory deals with the extension of the Special Theory to the case of the law of gravitation; we shall not discuss the General Theory here.

The principle of relativity was first stated by Newton in one of his axiomatics to the laws of motion: "The motions of bodies included in a given space are the same among themselves, whether that space is at rest or moves uniformly forward in a straight line." This means, for example, that if a space ship is moving along at a uniform speed, all experiments performed in the cockpit and within ρ distance in the space ship will appear the same as if the ship were not moving, provided, of course, that one does not look outside. That is the meaning of the stated principle of relativity. This is a simple enough idea, and the only question is whether it is true for all experiments performed inside a moving system the laws of physics

15-2 The principle of relativity

15-3 The Lorentz transformation

15-4 The Michelson-Morley experiment

15-5 Transformation of time

15-6 Length contraction

15-7 Time-dilation

15-8 Relativistic dynamics

15-9 Equivalence of mass and energy

will appear the same to Jay as if the system were standing still. Let us first investigate whether Newton's laws appear the same in the moving system.

Suppose that Moe is moving in the x direction with a uniform velocity v , and he measures the position of Jones's x -axis at point, shown in Fig. 15-1. This defines the " x -distance" of the point in his coordinate system as x' . Jones is at rest, and measures the position of the same point, designating its x -distance in "his system" as x . The relationship of the coordinates in the two systems is clear from the diagram. After $v t$ Moe's origin has moved a distance $v t$, and if the two systems originally coincide,

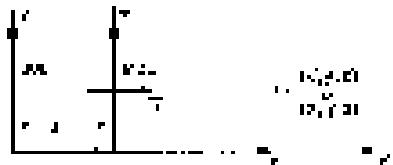


Fig. 15-1. Two coordinate systems in uniform relative motion along their x -axes.

If we substitute the coordinate law of conversion into Newton's laws we find that these laws transform to the static laws in the primed system, that is, the laws of Newton are of course true in a moving system as in a stationary system, and therefore it is impossible to tell, by carrying out such experiments, whether the system is moving or not.

The principle of relativity has been used in mechanics for a long time. It was employed by various people, in particular Huygen, to obtain the rule for the collision of billiard balls in which the same rule as we use it in Chapter 10 to discuss the conservation of momentum. In the past century, indeed, it was heightened as the result of investigations into the phenomena of electricity, magnetism, and light. A long series of such studies of these phenomena by many great scientists as in Maxwell's equations of the electromagnetic field, which describe electricity, magnetism, and light as one unified system. However, the Maxwell equations did not seem to obey the principle of relativity. That is, if we transform Maxwell's equations by the substitution of equations (15-1), both forces that we remain the same; therefore, in a moving space ship the electrical and optical phenomena should be different from those in a stationary ship. Thus one could use these optical phenomena to determine the speed of the ship; in particular, one could determine the absolute speed of the ship by looking outside. It applies to electrical means, however. One of the consequences of Maxwell's equations is that if there is a disturbance in the field such that light is generated, then electromagnetic waves go out in all directions equally and at the same speed c , or 186,000 miles/sec. Another consequence of the equations is that if the source of the disturbance is moving the light emitted goes through space at the same speed c . This is analogous to the case of sound. The speed of sound waves being obviously independent of the motion of the source.

Can the speed of the source of the waves, in the case of light, affect upon its traveling speed?

Suppose we are riding in a car that is going at a speed v , and light from the sun is going past the car with speed c . On the initial path the first equation of (15-1) gives

$$\frac{dx'}{dt} = \frac{dx}{dt} - v,$$

which means that according to the Galilean transformation the apparent speed of the passing light, as we measure it in the car, should not be c but $c - v$. For instance, if the car is going 100,000 miles/sec, and the light is going 186,000 miles/sec, then apparently the light going past the car should go 86,000 miles/sec. In any case, by measuring the speed of the light going past the car in the Galilean transformation, we can test if "light" can travel faster than c . A number of experiments were made in this group; they were performed to determine the velocity of the earth, but they all failed; they gave no velocity at all! We shall discuss one of these experiments in detail, to show exactly what was done and why was the law concerning light the matter, whereas something was wrong with the equations of physics. What could it be?

15-2 The Lorentz transformation

When the failure of the equations of physics in Newton's case came to light, the first thought that occurred was that the motion must be to the new Maxwell equations of electrodynamics, which were by 20 years older than him. I accepted almost without doubt that equations must be wrong or the thing to do was to change them in such a way that under the Galilean transformation the principle of relativity would be satisfied. When this was done, the new terms had to be put into the equations so the predictions of new electrodynamic laws did not exist at all when tested experimentally, so this attempt had to be abandoned. That is, probably because it appeared that Maxwell's laws of electrodynamics were correct, and the trouble must be sought elsewhere.

In the meantime H.A. Lorentz added a modification to his theory when he made the following substitution in the Maxwell's equations:

$$\begin{aligned} x' &= \frac{x - ct}{\sqrt{1 - v^2/c^2}} \\ y' &= y \\ z' &= z \\ t' &= \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}} \end{aligned} \quad (15.2)$$

namely, Maxwell's equations remain in the same form when this transformation is applied to them! Equations (15.2) are Lorentz's *contracting hypothesis*. Einstein, following a suggestion originally made by Poincaré, then proposed that all the physical laws should be of such a kind that they remain unchanged under a Lorentz transformation. In other words, we should change, not the laws of electrodynamics, but the laws of mechanics. Now that we change Newton's laws, then they will remain unchanged by the Lorentz transformation.⁴ If this goal is set, we can have to rewrite Newton's equations in such a way that the contraction becomes implemented. As it turned out, the only requirement is that the mass m in Newton's equations must be replaced by the form shown in Eq. (15.3). When this change is made, Newton's laws are the laws of electrodynamics now. Furthermore, if we use the Lorentz transformation, ... comparing Moe's measurements with ours, we should never be able to detect whether there is motion, because the form of the experiments will be the same in both coordinate systems!

It is interesting to observe what it means then we neglect the transformation between the coordinates and then with a new one, because the old one (Galilean) seems to be self-evident, and the new one (Lorentz) looks peculiar. We wish to know whether it is logically and experimentally possible that the new and not only the transformation is correct. A fact that alone is not enough to justify the laws of mechanics but, as Lorentz says, we also must compare the laws of gravity and force in order to make this confirmation. We shall focus on this in these laws and their implications for mechanics at some length, so we say in advance that the effort will be justified, since the results agree with experiment.

15-3 The Michelson-Morley experiment

As mentioned above, attempts were made to determine the speed of velocity of the earth through the hypothetical "ether" that was supposed to pervade all space. The most famous of these experiments is the performed by Michelson and Morley in 1887. It was 14 years later before the negative result of the experiment was finally explained by Lorentz.

The Michelson-Morley experiment was performed with an apparatus like that shown schematically in Fig. 15.2. This apparatus is essentially composed of a light source A , a partially silvered glass plate B , and two mirrors C and D , all mounted on a rigid base. The mirrors are placed at equal distances L from B . The plates E and F are used for mounting the A light and the beam splitter beam com-

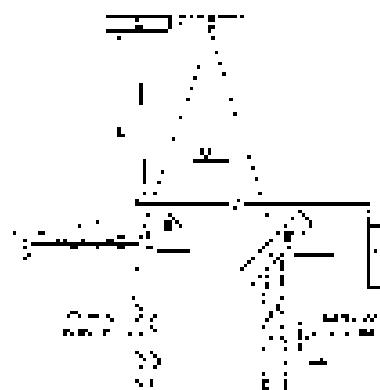


Fig. 15.2. Schematic diagram of the Michelson-Morley experiment.

travels in mutually perpendicular directions in the mirrors, where they are reflected back to B . On arriving back at B , the two beams are recombined as two separate parallel beams, B and C . If the time taken for the light to go from B to C and back is the same as the time from A to B and back, the emerging beams B and C will be in phase and will reinforce each other, but if the two times differ slightly, the beams will be slightly out of phase and interference will result. If the apparatus is "at rest" in the sense, the times should be precisely equal, but if it is moving toward the right with a velocity v , there should be a difference in the times. Let us calculate.

First, let us calculate the time required for the light to go from B to C and back. Let us say that the time for light to go from point B to point C is t_1 , and the time for the return trip is t_2 . Now, while the light is on its way from B to C , moreover, the apparatus moves a distance L , so the light must travel a distance $L + v t_1$ at the speed c . We can also express this distance as $c t_1$, so we have

$$ct_1 = L + vt_1 \quad \text{or} \quad t_1 = L/c - v/c$$

(This result is also obtained from the point of view that the velocity of light relative to the apparatus is $c - v$, so that one is the length L divided by $c - v$.) In a like manner, the time t_2 can be calculated. During this time the plane B travels a distance $v t_2$, so the total distance of the light is $L - v t_2$. Then we have

$$ct_2 = L - vt_2 \quad \text{or} \quad t_2 = L/c + v/c$$

Then the total time is

$$t_1 + t_2 = 2L/c + 2v/c = 2L/c(1 + v/c)$$

For convenience in later comparison of times we write this as

$$t_1 + t_2 = \frac{2L/c}{1 - v^2/c^2}. \quad (15.4)$$

Our second calculation will be of the time t_3 for the light to go from B to the mirror C' . As before, during time t_3 the mirror C' moves as the light accelerates to the position C' ; in the same time the light travels a distance $c t_3$ along the hypotenuse of a triangle, which is BC' . For this right triangle we have

$$(c t_3)^2 = L^2 + (v t_3)^2$$

or

$$L^2 = c^2 t_3^2 - v^2 t_3^2 = (c^2 - v^2) t_3^2,$$

from which we get

$$t_3 = L/\sqrt{c^2 - v^2},$$

For the return trip from C' the situation is the same, so we have. From the symmetry of the figure, therefore the return time is also the same, and the total time is $2t_3$. With a little rearrangement of the form we can write

$$2t_3 = \frac{2L}{\sqrt{c^2 - v^2}} = \frac{2L/c}{\sqrt{1 - v^2/c^2}}. \quad (15.5)$$

We are now able to compare the times taken by the two beams of light. In expressions (15.4) and (15.5) the same terms are identical, and represent the time that would be taken if the apparatus were at rest. In each denominator, the term v^2/c^2 will be small, unless this is comparable in size to v . The denominators represent the modifications in the times caused by the motion of the apparatus. And finally, since modification terms are not very great, the time to go to C' is less than the time to go back to C and back, even though the mirrors are equidistant from B , and C , we have v/c to increase. Total difference $\approx v/c$ percent.

There is another mechanical point arises—suppose the two lengths L are not exactly equal? In that we would expect some time effect. In that case we simply turn the apparatus 90 degrees, so that between the lines of motion and BC is perpendicular to the motion. Any small difference in lengths then becomes

diminution, and with we look for a shift in the interference fringes when we rotate the apparatus.

In carrying out the experiment, Michelson and Morley oriented the apparatus so that the line BC was nearly parallel to the Earth's motion, i.e., orbit, at certain times of the day and night. This is little speed is about 18 miles per second, and any "transversal" speed would have had that much in components of the day or night and no南北 speed during the year. The apparatus was finely sensitive to observe such an effect, but no time of the day was found. The velocity of the earth through the ether could not be detected. The result of the experiment was null.

The result of the Michelson-Morley experiment was very puzzling and somewhat disconcerting. This is first for failing a very basic of the important relativity from Lorentz. He suggested that material bodies contract when they are moving, and that this length-shrinking is only in the direction of the motion, and also that if the length is L_0 when a body is at rest, then when it moves with speed v parallel to its length, the new length, which we call L , is given by

$$L = L_0 \sqrt{1 - v^2/c^2} \quad (15.6)$$

When this modification is applied to the Michelson-Morley interferometer setup, the distance from B to C does not change, but the distance from B to E is shortened to $L_0 \sqrt{1 - v^2/c^2}$. Therefore $L_0 / (L_0 \sqrt{1 - v^2/c^2})$ is not changed, and the λ in Eq. (15.4) must be changed in accordance with Eq. (15.6). When this is done we obtain

$$\Delta t = \frac{(2L_0 \sqrt{1 - v^2/c^2})}{c} = \frac{2L_0}{c \sqrt{1 - v^2/c^2}} \quad (15.7)$$

Comparing Eq. (15.7) with Eq. (15.5), we see that $\Delta t = \tau_{\text{obs}} - \tau_{\text{exp}}$. If the apparatus rotates at the manner just described, we have a way of understanding why the Michelson-Morley experiment gives no result at all. Although the contraction hypothesis successfully accounted for the negative result of the experiment, it was open to the objective that it was invented for the express purpose of explaining away the difficulty, and was not utilized. However, it must other experiments to a similar set either with, or the different, however, it appears that nature was at the "conspiracy" to cover man by introducing some new phenomenon to make the experiment fail, he thought, would give it a measurement of c .

However, it is distinguished, as Poincaré pointed out, that, *provided conspiracy is still a law of nature*, Poincaré then proposed that there is still a law of nature, that it is not possible to determine in either way by any experiment, that is, there is no way to determine an absolute velocity.

15-4 Transformation of time

Is accepting our wavelet contraction idea as an analogy with the facts in other experiments is it reasonable. But everything is except provided that the wavelets are also modified in the manner proposed in the fourth, or the fifth section. So it turns out because the time t_0 , calculated for the trip from A to C and back, is not the same when calculated by a man performing the experiment in a moving spaceship as when calculated by a stationary observer who is watching the spaceship. To the man in the ship, the time is simply $2L_0/c$, but to the other observer it is $2L_0/c(1 - v^2/c^2)$ (Eq. 15.5). In other words, when the man looks the man in the spaceship lighting a cigarette, the screens appear to be slower (than normal), while to the other man, everything moves at a normal rate. So between them, the lighter the atom, but also the interatomic forces (which is called γ) must apparently slow down. That is, when the clock in the spaceship moves it slows down, as seen by the man in the ship; it takes $1/\gamma$ more "normal" seconds to the man in the ship.

This slowing of the clock in a moving system is a very peculiar phenomenon, and is worth an explanation. To begin to understand this, we have to watch the mechanics of the clock and see why it appears when it is moving. Since the β

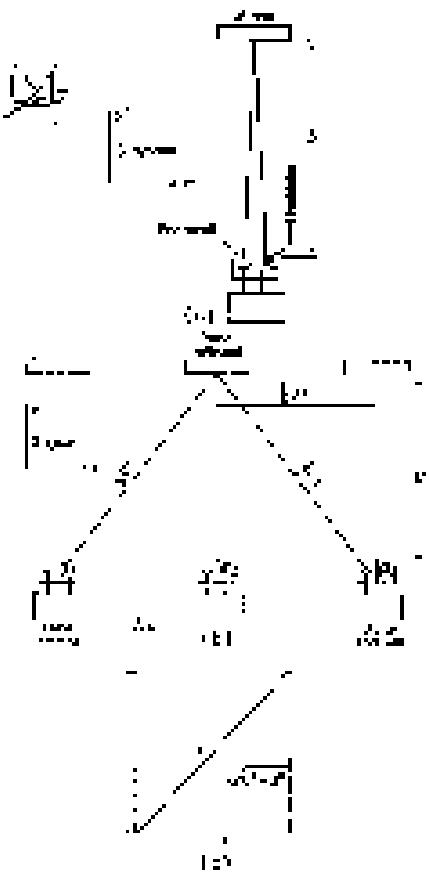


Fig. 15-2. (a) A "light clock" of two mirrors in the S system. (b) The same clock moving through the S' system. (c) Illustration of the diagonal path taken by the light beam in a moving "Twin" clock.

rather difficult, we shall take a very simple kind of clock. The one we choose is not a really kind of clock, but it will work if principles it is based upon are valid when a mirror is present, and, when we send a light signal between the mirrors, the light beam goes up and down, making a click every time it comes down, like a clock and ticking clock. We "bank" our twin clocks with exactly the same length, and synchronize them thereby starting them together, then they agree to agree thereafter, because they are the same in length, and light always travels with speed c . We give one of them, the S clock, no motion to take along in its space ship, and let it make the run perpendicular to the direction of motion of the ship; then the lengths of the rods will not change. How do we know that perpendicular lengths do not change? Two men can agree to make marks on each other's arms or necks. They pass each other. By symmetry, the two marks must come up the same way and change distance, since otherwise when they get together, to compare each other's neck with his own or below his other arm, he would really move it.

Now let us see what happens to the moving clock. Before the man does it, should he agree that it was a true constant clock, and when he goes along in the space ship, he will not see anything peculiar. If he did, he would know he was moving, if anything at all changed, because of the motion, he could tell he was moving. But the principle of relativity says this is impossible in a uniformly moving system, so nothing has changed. On the other hand, when the external observer looks at the clock with him, he sees that the light is going from center to mirror, & "nearly" taking a zig-zag path, since the rod is running sideways all the while. We have already analyzed such a zig-zag motion in connection with the Michelson-Morley experiment. I'm given to think the rod moves forward a distance proportional to $v t$, in Fig. 15-1, the distance the light travels in the same time is proportional to c , and the vector distance is therefore proportional to $\sqrt{c^2 - v^2} = c$.

That is, it takes a longer time for light to go round and back in the moving clock than in the stationary clock. Therefore the apparent time between clicks is longer for the moving clock, in this case, proportional to shown at the hypotenuse of the triangle (that is, the sum of the equivalent expressions in our equation). From the figure it is also apparent that the greater v is, the more slowly the moving clock appears to run. Not only does this particular kind of clock run more slowly, but if our theory of relativity is correct, any clock does, occurring on any principle whatsoever, would also appear to run slower, and in the same proportion—no matter who makes it. Why? This is why?

To answer the above question, suppose we had two identical clocks made only alike with tubes and gears, or perhaps based on radioactive decay, or something else. Then we adjust these clocks so they both run in precise synchronism with our fire-starts. When light goes up and back in the fire-starts and synchronizes them with a clock, the new model can complete some sort of cycle, which then simultaneously synchronizes by some similarly independent flash, a long, narrow signal. One of these clocks is taken into the space ship, along with the first kind. Perhaps this clock will not run slow, but will continue to keep the same time as its stationary twin part, and thus disagree with the other moving clock. At least, if that should happen, the man in the ship could use this mismatch between his two clocks to determine the speed of his ship, which we have been supposing is impossible. We now can draw a fitting picture of how every bit of the new clock that might cause the effect we simply knew and whatever the reason, it will appear to run slow, just like the first one.

Now if all moving clocks run slow, if for any reason changing time gives anything but a slower rate, we still just have to sit in a constant state. One thing may appear to be faster in a space ship. All the phenomena there— he needs time to eat, his thought processes, the time to raise to light a cigar, how long it takes to open his mouth, all these things must be slowed down in a space ship even, however, because all he is moving. The big experts and learned men sometimes say it is not quite certain that the time it takes for a cigar to develop will be the same in a space ship, but from the viewpoint of a modern physicist it is nearly impossible to imagine anything but the rate of time development to determine the speed of the ship!

A very interesting example of the working of time dilation is furnished by muons from cosmic-ray particles that disintegrate spontaneously after an average lifetime of 2.2×10^{-6} sec. They come to the earth from space, and can also be produced artificially in the laboratory. Some of them are created in mid-air, but the continuous disintegrations only after they encounter a mass to remain in c . It's clear that in its short lifetime a muon cannot travel, even at the speed of light, much more than 100 meters. But although the muons are released at the top of the atmosphere, some 10 kilometers up, they are usually found in a laboratory down here, in cosmic rays. How can this be? The answer is that different muons move at various speeds, some of which is a very close to the speed of light. While from their own point of view they only travel 2 meters, from our point of view they live considerably longer, i.e., much longer than they may reasonably expect. The factor by which the time is increased has already been given as $\gamma = \sqrt{1 - u^2/c^2}$. The γ -invariance factor is measured quite accurately for clusters of different velocities, and the values agree exactly with the formula.

We do not know *why* the theory, despite the γ 's, is valid, but we do know *why* or *where* the principle of relativity. That is to say, many of the predictions of relativity attempt to make predictions, even about things that otherwise we do not know much about. For example, before we knew anything about what matter is made of, it was not predicted that if it is moving at the speed of light, the speed it is all of time that it goes to is $c^2 = 1/c^2 = 0.73$ sec, and no prediction works—that is the kind thing about it.

15-5 The Lorentz contraction

Now it is return to the two-dimensional motion (15.2) and try to make better understanding of the relationship between the x -position and the (x', y', z') coordinates system, which we shall call the S' and S systems, or Joe and Mrs systems, respectively. We have already noted that the first equation is based on the Lorentz suggestion of synchronizing the two systems; how can we prove that a synchronization takes place? In the Michelson-Morley experiment we note apparent times were measured around 1900 m.s., but length by the principle of relativity, the null result of the experiment demands that the times must be equal, so, in order for the experiment to give a null result, the length of the $S'E$ must happen, by the requirement $\gamma(1 - u^2/c^2)$. What does this synchronization mean, a series of measurements made by Joe and Mrs? Suppose that Mrs, driving at the S' system in the x -direction, is measuring the x -coordinate of some point with a meter stick. She says the stick is a m. long, so he prints the distance is a' meters. From the viewpoint of Joe in the S system, however, Mrs is moving a longitudinal distance the "real" distance measured is $a\sqrt{1 - u^2/c^2}$ meters. Then if the S' system has traveled a distance a away from the S system, the S observer would say that the same point measured in his coordinates is at a distance $x = a\sqrt{1 - u^2/c^2} + a$ m.

$$x' = \frac{x - u t}{\sqrt{1 - u^2/c^2}},$$

which is the first equation of the Lorentz transformation.

15-6 simultaneity

In a non-inertial system, there are c different times. To make the simultaneous expression is introduced into the fourth equation of the Lorentz transforms. The most interesting term in this equation is the $u x'/c$ in the numerator, because that is quite new and unexpected. Now what does this mean? If we look at the situation carefully we see that events that occur at two separate places at the same time as seen by Mrs in S' , do not happen at the same time as viewed by Joe in S . If one event occurs at position x_1 at time t_1 , and the other occurs at position x_2 at time

Thus we find that the two corresponding times t_1 and t_2 differ by an amount

$$t_2 - t_1 = \frac{(v_{x_1} - v_{x_2})^2}{c^2}.$$

This time interval is called "Effect of simultaneity of a Galileo;" and to make the idea a little clearer let us consider the following experiment.

Suppose that a man moving in a space ship system S' has placed a clock at each end of the ship and is interested in making sure that the two start at the same time. How can the clocks be synchronized? There are many ways. One way involving very little calculation would be just to have exactly the same pulse between the clocks. Then from this a man would send a light signal which will go both ways at the same speed and will return at both clocks, clearly, at the same time. This signal arrives at all of the ship's clocks. He need to synchronize the clocks at the same time so he can synchronize his clock by the pulses of motion. Let us see whether an observer in system S would agree that the two are in synchronism. The man in S' sees light as follows: he sees because he does not know that he is moving. But the man in S reasons that since the ship is moving forward, the clock in the front end has gained time. The light signal, hence the light bulb to perform more to "keep up." In other words the clock, however, was advancing to meet the light signal so the distance was shorter. Therefore, the man in S needed the rear clock. But, although the man in S' thought that the signals traveled him separately. When the signals when a man in a space ship thinks the times at two locations are simultaneous, equal values of c in the coordinate system must correspond to different values of c in the other coordinate system.

13-4 Transformation

Let us see what she we can discover in the Lorentz transformation. It is interesting to note that the transformation between the x 's and x 's is analogous to that of the transformation between x 's and y 's that we studied in Chapter 11 for a relative of coordinates. We have

$$\begin{aligned} x' &= x \cos \theta - y \sin \theta, \\ y' &= y \cos \theta + x \sin \theta, \end{aligned} \quad (13.5)$$

in which the new x' means the old x and y' are the new y also in which the old x and y ; similarly, in the Lorentz transformation we find a new x which is a mixture of x and y , and a new y which is a mixture of x ' and y . So the Lorentz transformation is analogous to a rotation, only it is a "rotation" in space and time, which appears to be a curious concept. A check of the analogy is provided can be made by calculating the quantity

$$x'^2 - y'^2 + z'^2 = x^2 - y^2 + z^2 = c^2 t^2. \quad (13.6)$$

In this equation the first three terms on each side represent, in three-dimensional geometry, the square of the distance between a point and the origin (surface to a sphere) which remains unchanged (invariant) regardless of choice of the coordinate axes. Similarly, Eq. (13.6) shows that there is a certain combination which includes x , y , z that is invariant in a Lorentz transformation. Thus the analogy to a rotation is complete, and x 's and y 's are like unit vectors, i.e., quantities involving "components" which transform the same way as the coordinates and time, are transformed in connection with relativity.

Thus we continue our treatment of the idea of vectors which we have so far considered to have only space components, to include a time component. That is, we expect the them will be one \vec{v} with four components, three of which are the components of an ordinary vector, one with time well represented in the transformation, which is the analog of the time part.

This concept will be analyzed further in the next chapters, where we shall find that, if the ideas of the preceding paragraphs are applied to momentum, the transformation gives three space parts, that are: the ordinary momentum components, and a fourth component, the time part, which is the energy.

L5.9 Rekurrenzbedingungen

We are now ready to investigate, more generally, what form the laws of mechanics take under the Lorentz transformation. [We have thus far explained how length and time change, but not how we get the modified formulae for m (Eq. 291). We shall do this in the next chapter.] To see the varying stages of Einstein's modification of the Newtonian mechanics, we start with the Newtonian law that $\ddot{x}_i = \frac{d}{dt}(\dot{x}_i)$ is the rate of change of velocity x_i , or

$$F = \sigma_{\text{max}}^2$$

When I am in a C program, but after we use the `DEF` of this becomes

$$B = \sin \frac{\pi \omega t}{\sqrt{C^2 - I^2 \omega^2}}. \quad (15.10)$$

This is Einstein's modification of Newton's laws. Under this modification, if action and reaction are still equal (which they may not be in detail, but are in the long run), there will be conservation of momentum in the same way as before, but the quantity that is being conserved is not the old $m v$ with its constant mass, but instead \rightarrow the quantity given in (15.10), which has the modified mass. When this change is made in the form of a law of momentum, conservation of momentum still works.

Now let us see how momentum varies with speed. In Newtonian mechanics, it is proportional to the speed v , according to (15.10), over a considerable range of speed, but small compared to c . It is nearly the same in relativistic mechanics, because the significant modification differs only slightly from 1. But when v is almost equal to c , the square-root expression approaches zero, and the momentum therefore goes toward infinity.

Who happens? To consider, how does a body for a long time? In Newtonian mechanics the body increases picking up speed until it goes faster than light. But this is impossible in relativistic mechanics. In relativity, the body keeps picking up, not speed, but momentum, which can continuously increase because the mass is increasing. After a while there is practically no acceleration in the sense of a change of velocity, but the momentum continues to increase. Of course, whenever a force produces any small change in the velocity of a body, we say that the body has a great deal of inertia, and that is exactly what our formula for relativistic mass says (see Eq. 13.10). It says that the inertia is very great when v is nearly as great as c . As an example of this effect, to collect the high-speed electrons in the synchrotron that is used here at Caltech, we need a magnetic field that is 2000 times stronger than would be expected on the basis of Newton's laws. In other words, the mass of the electrons in the synchrotron is 2000 times as great as their normal mass, and it is greater than \sqrt{e} times of a proton. That is, it should be 2000 times m_p , so that $L = 1.6 \times 10^3$ must be $1.6 \times 10^3 m_p$, and that means that mc^2 differs from E by one part in 4,000,000, or that c differs from v by one part in 5,000,000, so the electrons are getting pretty close to the speed of light. If the electrons and light were born to gear from the synchronous (estimated at 100 feet away) and run out in Bridge Hall, who would arrive first? The light, of course, because light always travels faster.² How much earlier? That is not hard to tell; instead, we tell by what distance the light is ahead: it is about $1/2000$ of an inch, $\sim 3 \times 10^{-5}$ inch—of a piece of glass! When the electrons are going that lose their masses are electrons, but their speed cannot exceed the speed of light.

* The electrons would normally scatter from atoms under light because of the index of refraction of air. A permanent would make no losses.

Now let us look at some further consequences of relativistic change of mass. Consider the mass of all the molecules in a small unit of gas. When the gas is heated, the speed of the molecules is increased, and therefore the mass is also increased and the gas is heavier. An expression for this increase in the mass of mass, for example, when the velocity is small, can be found by expanding $m_0/\sqrt{1 - v^2/c^2} \approx m_0(1 + v^2/c^2)^{-1/2}$ as a power series. Using the form of expansion, we get

$$m_0(1 + v^2/c^2)^{-1/2} = m_0(1 + \frac{v^2/c^2}{2} + \frac{v^4/c^4}{8} + \dots).$$

We see clearly from the formula that the series converges rapidly when v is small, and the terms after the first two or three are negligible. So we can write

$$m \approx m_0 + M_{\text{kin}} \left(\frac{v^2}{c^2} \right) \quad (15.11)$$

in which the second term on the right expresses the increase of mass due to molecular velocity. When temperature increases the molecular propagation, so we can say that the increase in mass is proportional to the increase of temperature. But since $m_0 v^2$ is the kinetic energy in the non-relativistic Newtonian sense, we can also say that the increase in mass of all the molecules of gas is equal to the increase in kinetic energy divided by c^2 , or $M_{\text{kin}} = m_0 K_{\text{kin}}/c^2$.

15.4 Equivalence of mass and energy

The above observation led Einstein to the suggestion that the mass of a body can be expressed more simply than by the formula (15.1), if we say that the mass is equal to the total energy of content divided by c^2 . If Eq. (15.11) is multiplied by c^2 the result is

$$mc^2 = m_0 c^2 + M_{\text{kin}} c^2 + \dots \quad (15.12)$$

Here, the term on the left is called the total energy of the body, and we recognize the last term as the ordinary kinetic energy. Einstein interprets the long constant term, $m_0 c^2$, as part of the total energy of the body, as "rest energy," known as the "rest-mass."

Let us follow our own interpretation of meaning with respect to that the motion of a body without energy $m_0 c^2$. As an interesting result, we shall find the formula (15.12) for the conversion of mass to energy, which we have already assumed up to now. We start with the body at rest, when its energy is $m_0 c^2$. Then we apply a force to the body, which starts it moving and gives it kinetic energy; therefore, the total energy has increased. The new law of mechanics is implicit in the original assumption: as long as the force continues, the energy and the mass both continue to increase. We have already seen (Chapter 13) that the law of a type of energy with time varying force varies the velocity, v ,

$$\frac{dE}{dt} = F \cdot v. \quad (15.13)$$

We also have (Chapter 9, Eq. 9.1) that $F = \rho(x)v(x)$. When these relations are put together with the definition of E , Eq. (15.12) becomes

$$\frac{d(m_0 c^2)}{dt} = \rho(x) v(x) \frac{d(m_0 v)}{dt}. \quad (15.14)$$

We wish to solve this equation for $v(x)$. To do this we first use the technique just described of multiplying both sides by dm_0 , which changes the equation to

$$c^2 \left(\frac{dm_0}{dt} \right) \frac{dv}{dx} = \rho(x) v \frac{d(m_0 v)}{dt}. \quad (15.15)$$

We now let go one of the derivatives, which can be accomplished by integrating

both sides. The quantity $(\partial^{\mu} A_{\mu})/dt$ can be recognized as the time derivative of A^{μ} , and $(\partial^{\mu} A_{\mu}) - \partial A^{\mu}/\partial t$ is the time derivative of A^{μ} . But Eq. (15.17) is then simply

$$\partial^{\mu} (\frac{\partial A_{\mu}}{\partial t}) = \frac{\partial m^2 c^2}{\partial t}. \quad (15.18)$$

If the derivatives of two quantities are equal, the constants themselves differ at most by a constant, ϵ , i.e., C . This point is as before:

$$m^2 c^2 = m^2 c^2 + C. \quad (15.19)$$

We must again make the constant C massless initially. Since Eq. (15.19) must be true for all velocities, we can choose a special case where $v = 0$, and set $C = 0$; in this case the mass is m_0 . Substituting these values into Eq. (15.17) gives

$$m^2 c^2 = 0 + C.$$

We can now substitute values of C in Eq. (15.19), which becomes

$$m^2 c^2 = m^2 c^2 + m^2 c^2. \quad (15.20)$$

Dividing by c^2 and rearranging terms gives

$$m^2(1 - c^2/c^2) = m_0^2$$

from which we get

$$m = M_0 / \sqrt{1 - 1/c^2}. \quad (15.21)$$

This is the form of Eq. (15.1), and is exactly what is necessary for the agreement between mass and energy in Eq. (15.19).

Ordinarily these energy changes represent extremely slight changes in mass, because most of the time we cannot disentangle much energy from a given amount of mass-energy; but in an atomic bomb of explosive energy equivalent to 20 kilograms of TNT, for example, it can be shown that the difference between the initial and final masses is $\sim 10^{-13}$ g. The initial mass of the resulting neutron is m_1 because of the energy that was released, i.e., the released energy was a mass of 1 gram, according to the relationship $\Delta E = \Delta m c^2$. This famous "conservation of mass" may have been fully verified by experiments in which mass is annihilated—converted entirely to energy: an electron and a positron come together at rest, each with a rest mass m_1 . When they come together they annihilate and lose gamma-ray energy, each with the measured energy of $m_1 c^2$. This experiment provides a direct demonstration of the energy associated with the existence of the rest mass of a particle.

Relativistic Energy and Momentum

16-1 Reference and the philosopher

In this chapter we shall continue to discuss the principle of relativity of Einstein and Poincaré, as it affects our areas of physics and other because of human thought.

Twentieth-century philosophers seem of one mind in their view of the principle of relativity. According to the principle of relativity, the laws of physical phenomena must be the same for a fixed observer, as for an observer who has a uniform motion of translation relative to him, or "that we have no reason to precisely know, any means of discerning whether or not we are carried along at such a motion."

When this idea descended upon the world, it is said a great number of philosophers, particularly by "cocktail-party philosophers," who say, "Oh, it is very simple. Einstein's theory says all is relative!" In fact a surprising number of philosophers, not only those found at cocktail parties, but others, too, than philosophers, seem to think all is relative. A consequence of Einstein's theory is just now illustrated in an ad. In addition, they say, "It has been demonstrated in physics that phenomena depend upon your frame of reference." We can take a great deal but it is difficult to find out what it *means*. Probably the source of this idea lies were originally referred to were the coordinate systems which are in the very basis of the theory of relativity. So the fact that "things depend upon your frame of reference" is supposed to have had a profound effect on public thought. One might well wonder why *any* person, after all, that things depend upon one's point of view is an simple idea and certainly should have been necessary to go to all the trouble of the physical relativity theory in order to discover it. The whole matter depends upon his frame of reference is actually known to anybody who walks around because he sees an airplane and suddenly lost from the front and then from the back; this is nothing deeper than most of the philosophy which is said in the course from the theory of relativity than to comment that "A person looks different from the front than from the back." The old story about the elephant that several blind men describe in different ways is another example, perhaps, of the theory of relativity from the philosopher's point of view.

But actually there can be deeper things in the theory of relativity than just this single remark that "you can look different from the front than from the back." Of course relativity is deeper than this, because we can make definite predictions with it. I would say it is rather remarkable if we could predict the behavior of atoms from just a simple theory like this.

There is another school of philosophers who see very uncomfortable about the theory of relativity, yet it seems that one cannot live with our civilization seriously without ruling in something outside, and who would say "It is obvious that one cannot measure his velocity without looking outside. It is self-evident that it is impossible to talk about the velocity of a thing without looking outside. The physicists are quite simple for having thought otherwise, but it was just known from them that this is the case. If only the philosophers had realized what the problem were! The physicists had, so could have decided immediately by themselves that it is impossible to tell how fast one is moving without looking outside, and we could have had an immediate confirmation to ourselves." These philosophers are always with us, struggling in the periphery to try to tell us something but, they never really understood the subtleties and causes of the problem.

16-2 Relativity and the philosopher

16-2-1 The twin paradox

16-2-2 Transformation of velocities

16-2-3 Velocity-addition rules

16-2-4 Relativistic energy

Our inability to detect absolute motion is a result of experiment and not a result of plain thought, as we can easily illustrate. In the first place, Newton believed that it was true that one could not tell how fast one is moving with uniform velocity in a straight line. In fact, Newton himself was the author of this belief and one of earliest writers on the last topic was also an element of Newton's. Why then did the philosophers not make all this fuss about "all is relative?" or whatever is Newton's line? Because it was not until Maxwell's theory of electromagnetism was developed that a physicist knew that one could measure his velocity without looking outside; soon it was found that one could do this and could not.

Now if it were my definition, philosophy is necessary to one whom I do not be able to tell how fast he is moving almost looking outside. One of the consequences of relativity was the development of a philosophy which said "You can only define what you can measure." Since it is self-evident that one cannot measure a velocity without telling what he is measuring is relative to, therefore it is clear that there is no moving in absolute motion. The physicist should realize that one can only say about a thing "you can measure" by which the philosopher would mean that one can only define absolute velocity in the sense of what he can measure without looking outside; whether he is moving. Likewise, whether something is measurable is not something to be decided a priori. I might think he is moving that can be decided only by experiment. Given the fact that the velocity of light is 186,000 miles per second, one will find the few philosophers who will easily note that it is still evident that if light goes 186,000 miles per second and the car is going 100,000 miles, that the light can go 86,000 miles per second on the ground. This is a shocking fact to them, but very does with each, it is obvious that, when you have them a specific fact, that it is no problem.

Finally, there is even a philosopher who says that one cannot detect any motion except by looking outside. This is simply not true in physics. Thus, one can not provide a uniform motion in a straight line, but if he places a man who moves he would certainly know it, for everybody would be aware of the pull. There would be all kinds of "non-uniform" others. That, however, is a matter for the philosopher to determine without looking outside. Until a man moves, he does not know he. When this is said to a philosopher, he is very apter that he did not really understand it, because to him it seems impossible that one could be able to detect an motion about an axis without looking outside. If the philosopher is given enough of an excuse, he may come back and say, "I understand. We really do not have such a thing as absolute motion; we are really rotating relative to the stars, you see. And a some influence exerted by the stars on the object must cause the centrifugal force."

Now, let us see how we know that a man who has a car, at the present time, of telling whether there would have been centrifugal force if there were no stars and no other objects. We have not been able to do the experiment of removing all the people and cars, leaving the situation, so we simply do not know. We might, also, let the philosopher say "right." He comes back, therefore, to the philosopher and says, "It is necessary necessary that the world uniformly turn out to be the way uniform rotation means, nothing, it is only relative to the world." Then we say to him, "Now my friend, is not the uniform motion uniformity in a straight line, relative to the world, should produce no effects outside a car?" Now that the man is a little also interested in the relative, it becomes a question of question, and a question that can be answered only by experiment.

What then are the philosophical influences of the theory of relativity? If we continue his influences in the sense of a kind of one who had suggested one type to the physicist by the principle of relativity, we could describe some of them as follows. The first discovery is essentially that one does know which have been held for a very long time and which turn from very necessary to the "ignor be wrong." It was a shocking discovery, at once, and Newton's laws are it."

working, after all the years in which one seemed to be correct. Of course it is also true that the experiments were wrong, but that they were done was only a limited range of velocities, so small ones; the relativistic effects might not have been evident. But nevertheless, we now have a much more humble point of view of our physics; however, only thing one is wrong.

Secondly, if we come a bit of strange times such as the time goes slower when one moves, and so forth, whether we like them or do not like them is a nonscientific question. The only unscientific question is whether the Einstein's relativity was right or wrong experimentally. In other words, the "strange ideas" need only agree with experiment, and the only reason that we have to discuss the behavior of objects and so forth is to disprove it. But although the notion of the time dilation is strange, it is consistent with the way we measure time.

Finally, there is a third suggestion which is a little more technical but which seems good to be of importance: if you are only at other points, how are that to be read at the "beginning" of the time so, more specifically, to look for the ways in which the laws can be unchanged under a shift from the origin. When we develop the theory of motion, we never say the fundamental laws of motion are not changed when we rotate the coordinate system, and now we learn that they are not the "right" when we change the space and time variables in a suitable way, given by the Lorentz transformation. So this idea of finding the particular conditions under which the fundamental laws are not changed has proved to be a very useful one.

14.2 The relativity paradox

To continue our discussion of the theory of relativity and relativistic effects, we consider a famous story that "paradox" of Peter and his son who are supposed to be twins, born at the same time. When they are old enough to drive a car, say, Paul flies away at very high speed. Because Peter, who is left on the ground, sees Paul going so fast, all of Paul's clocks appear to go slower. His heart beats go slower, his thoughts go slower, everything goes slower, from Peter's point of view. Of course, Paul notices nothing unusual, but "he . . . accelerated and about ten minutes and then comes back, he is . . . older than Peter, the man on the ground! That is actually right, it is one of the consequences of the theory of relativity when two have every driven to that—less as the man who last longer when they are moving, so also will Paul last longer when he is moving. This is just this "paradox," only the people who believe that the principle of relativity means that all motion is relative; they say, "Well, look, you from the point of view of Paul, can't we say that Peter was moving and should therefore appear to age more slowly?" By symmetry, the only possible . . . is that Paul should be the same age when they meet." For in order for them to come back together and make the comparison, Paul must take, say, at the end of the trip and make a comparison of clocks or something, simply he has to come back, and the one whom comes back must be the man who was moving, and so know this, because he had to turn around. What he turned around? A kind of unusual things happened to his space ship—the rocks went off, things jammed against one wall and so on—while Peter left, etc.

At the very bottom, the only to know that the man who has felt the acceleration, who has seen things fall against the walls, and so on, is the one who would be the younger. Just > the difference between that in the "paradox" sense, and it is actually known. When we measured the fact that running man runs like human, we used as an example their straight-line motion in the atmosphere. The scientist also make measurement in a laboratory and cause them to go at a curve with a magnet, and even under the same field condition, they has varied as much for as they do when they are moving in a straight line. Although no one has arranged enough time especially on that we can get rid of the paradox, one could compare a magnet, which is left running with one can feel going around a magnetic field, and . . . would surely be found that the one that went around the circle last longer. Although we have not actually carried out an experiment using a magnet

circle. It is really not necessary, of course, because everything has longitude. Right. This may not satisfy those who insist that only single-fiber determinations will do, but we confidently predict the result of our experiment to watch Paul go to a complete circle.

16.2 Transformation of velocities

The main difference between the relativity of Einstein and the relativity of Newton is that the laws of transformation connecting the coordinates and times between different moving systems are different. In contrast, Isaac's motion law, the "law of motion," is:

$$\begin{aligned}x' &= \frac{x - vt}{\sqrt{1 - v^2/c^2}}, \\t' &= t, \\z' &= z, \\r' &= \frac{r - vx/c^2}{\sqrt{1 - v^2/c^2}}.\end{aligned}\quad (16.1)$$

These equations correspond to the relativity example now in which the relative velocity of the two observers is along their common x -axis. Only rectangular dimensions of motion are possible, but the time period. **Line 2** Transformation is called "coupled" with x' being split like this mixed up together. Why? **LT** continues to remain simpler form, since it contains all the essential features of the theory.

Let us now discuss more of the exact nature of this basic motion. First, it is interesting to ask: Are equations in moving—that is, here is a set of linear equations, four equations with four unknowns, and they can be solved in general, the x, y, z, t , in terms of x', y', z', t' . The result is very interesting, since it tells us how a system of coordinates "at rest" looks from the point of view of one who is "moving." Of course, since the motion can also be one of uniform velocity, the "rest" or "moving" can say if he wishes, but it is only then he follows who is moving and he himself who is at rest. And since he is moving in the opposite direction, he should see the same transformation, but with the opposite sign of velocity. That is precisely why we find by extrapolation, you don't have to think if he not come out that way, we would have real cause to worry!

$$\begin{aligned}x &= \frac{x' + vt'}{\sqrt{1 - v^2/c^2}}, \\y &= y', \\z &= z', \\t &= \frac{t' + vx/c^2}{\sqrt{1 - v^2/c^2}}.\end{aligned}\quad (16.2)$$

Next we discuss the interesting problem of the addition of velocities in relativity. We recall that one of the original postulates was that light travels c. 186,000 miles in all systems, even when they are in relative motion. This is a special case of the more general principle exemplified by the following. Suppose that an object made a space ship is flying at 100,000 miles/sec and the space ship itself is going at 100,000 miles/sec fast is the object made the space ship moving from the point of view of an observer outside? We might say no, say 200,000 miles/sec, which is more than the speed of light. This is very interesting, too, since it is required to be going faster than the speed of light. The general problem is as follows:

Let us suppose the the object made the ship from the point of view of the outside, is moving with velocity v , and that the space ship itself has a velocity u with respect to the ground. We want to know with what velocity v' the object is moving to the point of view of him in on the ground. This is, of course, still but a special case in which the motion is in the x -direction. There will also be a

transformation for velocities in the y -direction, or for any right angle can be worked out as needed. Make the same assumption, velocity is v_{xy} , which means that the displacement x is equal to the velocity times the time:

$$x' = v_{xy}t. \quad (16.1)$$

Now we have only one relation: who the position and time are from the point of view of the outside observer for the object which has the relation, if 6.2, between x' and t . So we simply substitute (16.1) into (16.2), and obtain

$$x' = \frac{v_{xy}t + vt}{\sqrt{1 - v^2/c^2}}. \quad (16.4)$$

At this we find x' expressed in terms of t . In order to get the velocity as seen by the observer, the outside, we must divide the distance by the time and by the factor c/v . So we must also take advantage to go from the outside, which is

$$z = \frac{t + v(x_{xy}/c)^2}{\sqrt{1 - v^2/c^2}}. \quad (16.5)$$

Now we must find the ratio of x to z , which is

$$x_2 = \frac{x}{z} = \frac{y + vt}{t + v(x_{xy}/c)^2}. \quad (16.6)$$

An important note having to do with. This is the law that we work the resultant velocity, i.e., "sum" of two velocities, is not just the algebraic sum of two velocities (we know that it wouldn't be so, we just in practice, but is "corrected" by $1/c^2$ term).

Now let us see what happens. Suppose that you are moving inside the space ship at the speed of light and that the space ship itself is going at the speed of light. This is impossible, but is the definition used in defining the γ factor.

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1 - 1}} = \infty.$$

So, in reality, "half" and "light" does not mean "one," it makes only " $1/\gamma$." Of course low velocities can be dealt quite easily in the familiar way, because so long as the velocities are small compared with the speed of light we can forget about the $(1 - v^2/c^2)$ factor; but things are quite interesting at high velocity.

Let us take a limiting case, just for fun, suppose that inside the space ship the man was observing light itself. In other words, $v = c$, and yet the space ship is moving. How will it look to the man on the ground? The answer will be

$$x = \frac{s + t}{1 - sv/c} = s + \frac{t}{\gamma} = s.$$

Therefore, if something is moving at the speed of light inside the ship, it will appear to be moving at the speed of light from the point of view of the man on the ground. This is good, for it is, in fact, what the Einstein theory of relativity was designed to do in the first place—so it can't work.

Of course, there are cases in which the motion is not in the direction of the uniform translation. For example, here may be an object inside the ship which is not moving "properly" with the velocity v_x with respect to the ship, and the ship is moving "horizontally." Now, we simply go through the same thing only using v_x instead of v , with the result

$$y = y' + v_x t,$$

so that $v_{xy} = 0$,

$$x' = \frac{t}{\gamma} = \frac{t}{\sqrt{1 - v^2/c^2}} = t/c^2. \quad (16.7)$$

This is a sideways velocity v_y and not v_x , but $v_y^2 = v^2 - v_x^2$. We can obtain it by substituting and simplifying the measurement equations. But we can also get it directly from the principle of relativity for the following reason: it is always good to look again at what we can and cannot do. We know only that light always travels at the speed c in the laboratory. While a moving particle appears to travel at c , it goes at the speed v in the laboratory. While a moving particle moves with the same speed in the moving system. We found that the vertical component of the velocity in the moving system is less than that of light by the factor $\sqrt{1 - v^2/c^2}$ (see Eq. 15-2). But now suppose that we let a material particle go back and forth in a some "clock" having some "frequent" f_0 at some "natural frequency" of light (Fig. 16-1). Then when the particle has gone back and forth n times the light will have gone exactly n times. That is, the "frequency" of the moving clock will coincide with each other "below" the light clock. One goes up and down, the other moves to and fro. Because the speed of the emission of photons will be approximately the same, the particle must be slower than the corresponding speed by the same small fractional. That is why the square root appears in any velocity velocity.

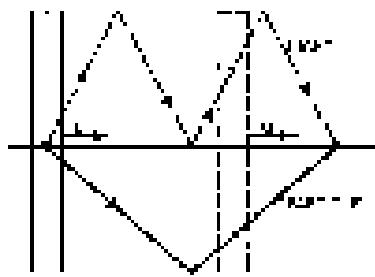


Fig. 16-1. Two diagrams showing how a light ray and particle below a moving clock.

16-4 Relativistic mass

You know or have test data that the mass of an object increases with velocity, but no demonstration of this was given, in the sense that we made no arguments to prove it. Now about the way clocks move we believe. However, we did prove that the consequence of this belief gives four very remarkable implications. The most interesting to us is this. (We have to say "a few other assumptions" because we cannot prove anything unless we have some laws which we assume to be true if we expect to make reasoning". Definitions.) Consider the case to study the transformation law of the total energy E of an object's motion, where we shall take nothing constant but the law of Newton. First we shall assume the conservation of momentum and energy. Also, we shall assume that the momentum of a particle which is moving in a straight line is always directed in the direction of the velocity. (However, we shall not assume that the momentum is necessarily along the velocity as Newton does, but only that it is scale parallel to velocity. We shall write the momentum vector as a coordinate function of the particle velocity.)

$$p = m v, \quad (16-5)$$

We put a condition on the coefficient to remind us that it is a function of velocity, and we shall agree to call the coefficient in the "rest" of expression where the velocity is zero. It is the rest mass that we would measure in the slow moving experiments with wave motion. Now we shall try to determine what the form of the mass must be. First, by arguing from the principle of causality into the laws of physics must be causal or have causal links between distant events.

Say you that we have two systems, like two particles, that are moving away and they are moving toward each other with exactly equal velocities. Their total momentum is zero. Now what happens? After the collision, their three pieces of motion must be exactly opposite to each other. Because if they are not exactly opposite, there will be a nonzero total vector momentum, and afterward it will not have been conserved. Also they must have the same speed, since after they exactly stop, all the energy is converted in the collision. So the digits of an electric charge or mass scale remain set, and the length between all the charges are the same length, all the speeds are equal. We can suppose that such causality can always be assumed. That is to say that causality, and that only speed could be used in such a collision. Next, we notice that the same particle can be visualized differently by turning the axes, and just for convenience we shift them the axes, so that the horizontal splits it evenly, as in Fig. 16-2(a). It is the same clock in motion, only with the axes "tilted".

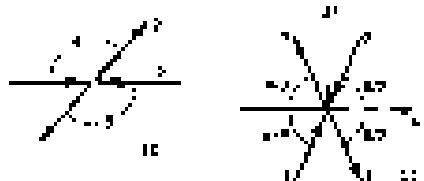


Fig. 16-2. Two stages of an elastic collision between equal objects moving at the same speed in opposite directions.

Now here is the trick: let me look at this collision from the point of view of someone riding along in a car that is moving at the speed equal to the horizontal component of the velocity of one particle. Then how does the collision look? It looks as though particle 1 is just going straight up because it has lost its horizontal component, and it comes straight down again afterwards in a shorter time than that our particle. That is, the collision appears as shown in Fig. 16.3(a). Particle 2, however, was going the other way, and so the impact appears to fly by at some terrible speed and at a smaller angle, but we see apparently that the angles before and after the collision are the same. So, the change of the horizontal component of the velocity of particle 2, and by so the vertical velocity of particle 1.

Now the question is, what is the vertical velocity v_{1y} ? If we knew that, we could get the momentum gain for the particle, assuming the law of conservation of momentum is the same. However, I think the horizontal component of the momentum is unchanged. It is the same before and after the collision for both particles and is zero for particle 1. So we must use the conservation law only for the horizontal velocity v_{1x} . Let us forget the particle theory, simply by focusing on the particle, looking the other way! If we look at the collision of big (6-kg) hammer to the 10-kg mass with speed $v_1 = v_0$, the same collision "turns over," as shown in Fig. 16.3(b). Now particle 2 is the one that goes up and down with equal speed and particle 1 has picked up the horizontal speed v_0 . Of course, now we know why the velocity v_{1y} is zero in $v_0 = v_{1y}$ (see Fig. 16.7). We know that the change in the vertical component of the velocity moving over time is

$$v_{1y} = \frac{v_0}{m_1} = \frac{v_0}{6\text{ kg}}.$$

(2), because it moves up and back down. The "slightly" moving particle has a initial velocity v_0 whose components we have found to be $v_{1x} = v_0/\sqrt{3}$, and $v_{1y} = v_0/2$. There is again recoil momentum, and this particle is therefore $\Delta p = -2m_2 v_0/3 = -v_0/3$ heavier, in accordance with our assumed law. (In 3) the horizontal component is always the mass corresponding to the magnitude of the velocity times the component of the velocity in the direction of interest. Thus in order for the total momentum to remain the vertical momenta must cancel and the ratio of the mass moving with speed v_0 and the mass moving with speed v_0 must therefore be

$$\frac{m_2}{m_1} = \sqrt{1 - v_{1x}^2/v_0^2}. \quad (16.9)$$

Let us take the limiting case that m_1 is infinitesimally small, very tiny indeed, so that the two objects are essentially equal. In this case, $m_1 = m_2$ and $m_1 = 1\text{ kg}$. Let's go ahead and do

$$\frac{m_2}{m_1} = \frac{m_2}{\sqrt{1 - v_{1x}^2/v_0^2}}. \quad (16.10)$$

From our earlier discussion, we know that if we want Eq. (16.8) to reduce to Eq. (16.9) it needs time for arbitrary values of v_0 , assuming that Eq. (16.10) is the right formula for the mass. Now that the velocity associated to Eq. (16.9) can be calculated from the right-hand side of (16.10)

$$v_0^2 = v_{1x}^2 + v_{1y}^2(1 - v_{1x}^2/v_0^2).$$

Now we find to make out automatically, although we need it only in (16.9) and not in (16.10).

Now, let us accept that momentum is conserved and that the mass depends upon the velocity according to (16.10) and prove that whatever we calculate. Let us consider what is commonly called an elastic collision. For simplicity, we can suppose that two pieces of mass m_1 and m_2 moving oppositely with equal speeds v_0 at first stick together to become some new, combined object, as shown in Fig. 16.4(a). The mass of this new object is $m_1 + m_2$, so we know $v_{1x} = -v_{2x}$. If we assume the conservation of momentum and the principle of relativity, we can demonstrate an interesting fact about the mass of the new object which has been formed. We compute the initial total velocity

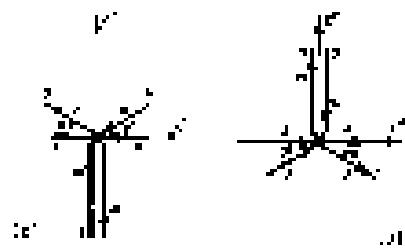


Fig. 16.3. Two views of the collision between moving cars.

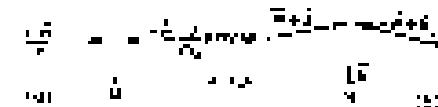


Fig. 16.4. Two views of an elastic collision between moving objects.

at right angles to it (we can do the same with other values of α , but it's easier to understand it with the horizontal velocity), then both objects move in directions parallel to an object's initial velocity, $v - p$. What we see is shown in Fig. 16-1(b). The composite object has an "run-down mass" M . Now object 1 moves with an upward component of velocity equal to the initial component which is proportionally equal to v , and so does object 2. After the collision we have the mass M moving upward with velocity v , considered as a small composite with the speed of $|v|$, and a mass m associated with it. Momentum must be conserved, so let us determine the momentum in the upward direction before and after the collision. Before the collision we have $v_1 = |v|$, and after the collision the "center" will be evidently $p_1 = Mv$, but p_2 is evidently the same as $m v$ because v is so small. These momenta must be equal because of the conservation of momentum, and therefore

$$Mv = mv \quad (16.11)$$

The mass of the whole object is shared when two objects move with equal and opposite velocities when they come together. You might say, "Yes, of course, that is the conservation of mass." But it "is not, of course" necessarily, because when two objects have exchanged mass, the masses m_1 and m_2 will be if they were exceeding still, yet they still contribute to the total M , not the mass they have taken standing still, just more. According to that may seem, in order for the conservation of mass to work when two objects come together, the mass that they form **must** be greater than the rest masses of the objects even though the objects are not after the collision!

16.5 Relativistic energy

In the last chapter we demonstrated that as a result of the dependence of the mass on velocity (see Newton's laws, the changes in the kinetic energy of an object resulting from the total work done by the forces on it), "total energy" can be

$$\Delta T = (m_1 + m_2)c^2 - \frac{m_1 c^2}{\sqrt{1 - v^2/c^2}} - m_2 c^2. \quad (16.12)$$

We can write further, and guessed that the total energy is the total mass times c^2 . Now we can take this definition.

Suggest that our two equally massive objects can still be "inside" M . For instance, a proton and a neutron are "stack together," but are still moving along, inside of M . Then, although we might at first suspect the mass M to be $2m_0$, we have found that it is not $2m_0$, but $m_1 + m_2$. Since $2m_0$ is when $v = 0$, but $m_1 + m_2$ are the rest masses of the things inside, the "new" mass of the composite object is equal to the kinetic energy brought in. This means, of course, that energy has flowed. In the last chapter we discussed the analogy of gas and showed that because the gas molecules are moving and moving atoms are heavier when we put energy into the gas its molecules move faster and the gas goes faster. In fact the argument is completely general, and our discussion of the inelastic collision shows that the mass \rightarrow time relationship is kinetic energy. In other words, "Two particles come together and sacrifice potential to any other form of energy (the pieces are closed doors to shrinking it)", being work against internal forces, or whatever, then it is true that the mass is not initially, but has been put in. We see that the conservation of mass which we have deduced above is equivalent to the conservation of energy, and therefore there is no place in the theory of relativity for strictly inelastic collisions, as there was in Newtonian mechanics. According to Newtonian mechanics ... is all right for two things to collide and so for a collision of mass $2m_0$, which is in many ways distinct from the one that would see a "summarizing them together slowly." Of course we know from the law of conservation of energy that there is more kinetic energy inside, but that does not alter the mass, according to Einstein's laws. That now we see that this is impossible: because of the kinetic energy involved in the collision, the resulting

object will be broken; therefore, it will be a different object. When we pull the object together, yet I may make something whose mass is large; when we pull them together forcefully, they may become something whose mass is small. Then the mass is different, we can not say it is different. So unless this, the conservation of energy must go along with the conservation of momentum in the theory of relativity.

This has interesting consequences. For example, suppose the we have an object whose mass M is moving and, and suppose something happens so that it flies into two equal pieces moving with speed c , so that they each have a mass m . Now suppose that these pieces contain enough material to slow them up until they stop; then they will have mass m . How much energy will they have given to the material after they have stopped? Each will give an amount $(m - m_0)^2/c^2$, by the famous theory of general relativity. This mass energy is left in the material as some form, as heat, potential energy, or whatever. Now $2m_0 = M$, so the total initial energy is $E = (M - 2m_0)c^2$. This is good for the usual calculations how much energy would be liberated under fusion in the atomic bomb. For example, although the fragments are not exactly equal, they are nearly equal. The mass of the atomic bombinium was known. It had been measured. One day -- see the atomic bomb which is split,裂變, atom, and so on, all were all known years ago. Today we do not know the masses while calculating one thing, we need the masses when the atoms裂變, or other words, heat, of and any one known. So by subtracting the two numbers one can calculate how much energy will be released. If one has a rock to split in "half". For this reason your old Einstein was to say the "father" of the atomic bomb is all the newspaper. Of course, all that means was that he could tell us ahead of time how much energy would be released if we had known what process would occur. The energy that should be released when an atom, or hydrogen undergoes fission was estimated about six months before the first direct test, and as such as the energy was in fact liberated, someone immediately wrote (and if Einstein's formula had not worked, they would have measured it anyway), and the nuclear energy measured it, they no longer needed the formula. Of course, we should credit him to him, but he also should credit the newspaper and many popular descriptions of what comes when in the history of physics and technology. The problem of how to get the thing to work is an effective and important one is a completely different matter.

The most important significant is chemistry. For instance, if we were to weight the carbon dioxide molecule and compare its mass with that of the carbon and the oxygen, we think that not much energy would be released when a hydrogen and oxygen form carbon dioxide. The only trouble here is that the substances in question are so small that it is relatively very difficult to do.

Now let me come to the question of whether we should add m_0^2 to the kinetic energy and say from now on that the total energy of the object is $E + m_0^2$. First, if we can still see the component pieces of new mass m_0 inside M , then we should say that some of the mass M of the compound object is the massless rest mass of the parts, part of it is kinetic energy of the parts, and part of it is potential energy of the parts. So we have discussed, it is quite possible of course to do such nuclear reactions just like the one we have treated above, in which when all the energy in the world, + or - goes over the mass. For instance, when E. K. Meissner distinguishes into two parts it does not necessarily do the $(E + m_0^2)$ but the idea that E is made out of m_0^2 is a useless idea, because it also distinguishes into E 's.

For this we have a situation: we cannot have to know what E and m_0 is made of inside; we cannot say that we identify individual particles, which of the energy is rest energy of the parts until which it is going to disintegrate. It is not convenient and also not possible to separate the total rest energy of an object into the energy of the individual pieces. Kinetic energy of the pieces, and potential energy of the pieces moreover, we simply ignore at the total energy of the particle. We "shut the engine" of energy by adding m_0^2 to a rest energy being, and say for the individual particle it increases in scaling times c^2 , and when the object is standing still, the energy is the mass of that object c^2 .

Typically, we find that the velocity v , momentum p , and total energy E are expressed in a rather simple way. That the mass in relation to speed c is the most reasonable one divided by $\sqrt{1 - v^2/c^2}$ (neglecting mass) is hardly useful. Instead, the following relations are easily proven, and turn out to be very useful:

$$E^2 - p^2 c^2 = m^2 c^4 \quad (16.15)$$

and

$$p_c = p_c/c \quad (16.16)$$

Space-Time

17-1 The geometry of spacetime

The theory of relativity shows us that the relationships of positions and times as measured in one coordinate system are different from what we would have expected in the Galilean intuition. This is very important, and we thoroughly understand the relations of space and time implied by the Lorentz transformation; and therefore we shall consider little matter more deeply in this chapter.

The Lorentz transform for between the positions and times (x, y, z, t) as measured by an observer "stationary at ∞ " and the corresponding coordinates and time (x', y', z') as measured inside a "moving" space ship, moving with velocity v , is:

$$\begin{aligned}x' &= \frac{x - vt}{\sqrt{1 - v^2/c^2}}, \\y' &= y, \\z' &= z, \\t' &= \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}}.\end{aligned}\quad (17.1)$$

Let us compare these equations with Eq. (10.5), which also relates measurements in two systems, one of which is in translation relative to the other:

$$\begin{aligned}x' &= x \cos \theta - y \sin \theta, \\y' &= y \cos \theta + x \sin \theta, \\z' &= z.\end{aligned}\quad (17.2)$$

In this particular case, x and y are increasing with time having an angle θ between the x' - and x -axes. In each case, we note that the "primed" quantities are "the result" of the "transformation," since the new x' is a mix of old x and y , and the new y' is also a mixture of x and y .

An analogy is useful: When we look at an object, there is no obvious thing we might call the "depth, width," and another we might call the "height." But the two being, width, and depth, are not fundamental properties of the object, however. If we step aside and look at the same thing from a different angle, we get a different width and a different depth, and we may develop some formulas for computing the new ones from the old ones and the angles involved. Equations (17.2) are those formulas. One might say that a given depth is a kind of "mixing" of all depth and all width. If it were impossible ever to move, and we always saw a given object from the same position, then this whole business would be unnecessary—we could always see the "true" width and the "true" depth, and they would agree in how quite different qualities, because one expects as a subtler optics, angle and the other involves a refraction, of the lens or even rotation, they would seem to be very different things and would never get mixed up. It is because we can walk around that we realize that depth and width, somehow or other, just two different facets of the same thing.

Can we look at the Lorentz transformation in the same way? Then also we have a mixture of positions and the time. A difference between a space measurement and a time measurement produces a new space measurement! In other words, in the space measurements of one man, there is mixed in a jolt of time, as seen by the other. One cannot seem to separate this fact. The "reality" of

17-1 The geometry of spacetime

17-2 Space-time intervals

17-3 Past, present, and future

17-4 More about four-vectors

17-5 Four-vector algebra

an object (as); we are looking at a situation where (speaking crudely and intuitively) even its "width" and its "depth" measure they depend upon how we look at it: when we move to a new position, our brain immediately reevaluates the width and the depth. This is not immediately recognizable, but it does, and this is what we mean by high speed, because we have had an effect on experience of going nearly as fast as light; to accelerate the fact that time and space are also of the same nature. It is as though we were always stuck in the position of having to think of just the width of something, not being able to move our heads independently one way or the other; if we could, we understand now, we would see some of the other man's face. We would say "horizontal," we speak a little bit.

Now we shall try to think of objects in a new kind of world, a space-time mixed together, in the same sense that the objects in our ordinary space world are real and can be looked at from different directions. We shall then consider the objects occupying space and having for a certain length of time occupy a kind of a "block" in this kind of world, and this we call at this "block" from different points of view when we are moving at different velocities. This, however, has the geometrical reality in which the "blocks" exist by occupying position and taking up a certain amount of time, is called gravitons. A given point (x_1, y_1, z_1, t_1) is called an isolated graviton. Imagine, for example, that we plot the positions horizontally, y and z in two other directions both initially at "right angles" and at "right angles" in the sense (1), and time, vertically. Now, how does a moving particle, say, look on such a diagram? If the particle is standing still, then it has a certain x , and the time goes on, it has the same x . In short, the same x space "path" is a line that runs parallel to the x -axis (Fig. 17-1(a)). On the other hand, if it drifts outward, then as the time goes on x increases (Fig. 17-1(b)). So, you know, for example, which starts to drift out and then slows down should have a motion something like that shown in Fig. 17-1(c). A particle, in other words, which is permanent and does not disintegrate is represented by a line in spacetime. A particle which disintegrates would be represented by a forked line, because it would turn into two other things, which would start from that point.

What about light? Light's speed in the space-time, and that would be represented by a line having a certain fixed slope (Fig. 17-1(d)).

Now according to our new idea, if a given event passes into a particle, say if it is suddenly dissevered at a certain speed, it points into our new axes what follows we have L and C , and this has nothing ever occurred at a certain value of x and a certain t , so if x and t then we would expect L.C., if this makes any sense, we just have to take a few pairs of ones and two there, and that will give us the new road. Our new x is our new system, as shown in Fig. 17-2(a). But this is wrong, because Eq. (17-1) is not exactly the same mathematical expression as Eq. (17-2). Note, for example, the difference in sign between the two, and the fact that one is written in terms of $\cos \theta$ and $\sin \theta$, while the other, i.e. written with algebraic quantities. Of course, it is not impossible that algebraic quantities could coincide, as cosine and sine, but actually they cannot, but still, the two expressions are very similar. As we shall see, it is not really possible to think of spacetime as a real, ordinary geometry, because of that difference in sign. In fact, although we shall not emphasize this point, it turns out that a man who is moving has to make an "axis" which are inclined equally to the light ray, using a "geometric" kind of projection parallel to the x and t axes, for his x and t as shown in Fig. 17-2(b). We shall not deal with the geometry, since it does not help much, it is easier to work with the x - y - z axes.

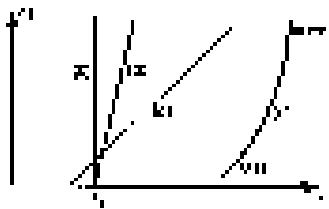


Fig. 17-1. Three particle paths in space-time: (a) a vertical line at $x = x_1$; (b) a particle with constant velocity $v = v_1$ and moves with constant velocity; (c) a particle which slows at high speed down.

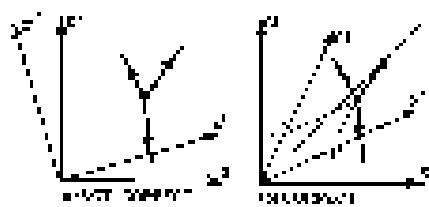


Fig. 17-2. Two views of a disintegrating particle.

17-3 Spacetime intervals

Although the geometry of spacetime is not based on the ordinary sense, there is a quantity which is very similar, but perhaps in certain respects. If this kind of geometry is right, there ought to be some functions of spacetime and time which are independent of the coordinate system. For example, under ordinary rotations, if we take two points, one at the origin, the x -axis, and the other on x -axis, the two systems would have the same x given the distance from

less to the other point is the same in both. That is one property that is independent of the particular way of measuring it. The square of the distance is $x^2 + y^2 + z^2$. Now what about spacetime? It is not hard to demonstrate that we have here, also, something which stays the same. Namely, the combination $x'^2 - x^2 - y^2 - z^2$ is the same before and after the transformation:

$$x'^2 - x^2 - y^2 - z^2 = x^2 + y^2 + z^2 - c^2. \quad (17.2)$$

This quantity is therefore something which, like the distance, is "real" in some sense. It is called the interval between the two space-time points, one of which is, in this case, at the origin. (Actually, of course, it is the interval squared, just as $x^2 + y^2 + z^2$ is the distance squared.) We give it a different name because it is in a different geometry, and the interval implying identity has some signs reversed and there is no unit.

Let us get rid of the c : this is an abuse if we are going to have a sensible system with c 's and y 's that can be unambiguous. One of the curiosities this could be caused by occurring with no experience would be to measure widths, say, by the range subtended in the eye and measure depth in a different way, like the sum of the angles needed to focus them, so that the angle would be measured in feet and the width in meters. Then one would get all enormously complicated messes of equations in making transformations such as (17.2), and would not be able to see the beauty and simplicity of the theory for a very simple technical reason, just the same thing is happening measured in two different units. Note that Eqs. (17.1) and (17.3) make c refer to that time and space unit equivalent; time becomes space; they should be measured in the same units. What distance is a "second"? It is easy to figure out from (17.1) what it is. It is 3×10^8 meters. The distance that light could cover in one second. In other words, if we were to measure all distances and times in the same units, seconds, then our unit of distance would be 3×10^8 meters, and the equations would be simpler. Or another way that we could make the units equal is to measure time in meters. What is a "meter" of time? A meter of time is the time it takes for light to go one meter, and so therefore $1/(3 \times 10^8)$ sec., or 0.3 billionths of a second.¹ We would like, in other words, to put all our equations in a system of units in which $c = 1$. If time and space are measured in the same units, as suggested, then the equations are obviously much simplified. They are

$$\begin{aligned} x' &= \frac{x - ct}{\sqrt{1 - \beta^2}}, \\ \gamma &= \frac{c}{v}, \\ \beta &= \frac{v}{c}, \\ \epsilon &= \frac{1 - \beta^2}{\sqrt{1 - \beta^2}}. \end{aligned} \quad (17.4)$$

$$t'^2 - x'^2 - y'^2 - z'^2 = t^2 - x^2 - y^2 - z^2. \quad (17.5)$$

Now we can drop the "light unit". But also we have this system with $c = 1$. We shall never be able to get our equations right again, the answer is called the impossible. It is much easier to remember them without the c 's in them, and it is always easy to put the c 's back, by looking after the dimensions. For instance, in $\sqrt{1 - \beta^2}$, we know that we cannot subtract a velocity squared, which has units, from the unity minus it, so we know that we must divide v^2 by c^2 in order to get the right answers, and this is the way to do it.

The difference between space-time and ordinary space, and the character of the interval as related to the distance, is very interesting. According to formula (17.5), if we consider a point which in a given coordinate system had zero time and only space, then the interval squared would be negative and we would have an imaginary interval, the square root of a negative number. Intervals can be either real or imaginary in the theory. The square of an interval may be still positive or negative, unlike distance, which has a positive square. When an interval is imaginary, we say that the two points have a spacelike interval between them

(region of imaginary), because the interval is more like "space than like time." On the other hand, if two objects cover the same space in a given coordinate system but it's only in time that the square of the time is positive, and the distance not zero, not the interval would be positive; this is called a pseudo-time interval. In addition to spacetime, themselves we would have a reverse "time" coordinate, such that $x^2 - t^2$ where x are two lines (position, x , time, t) and t will be "excess" called light cones) and points on these lines will be zero interval from the origin. When a light cone from a given point is always separated from it by some interval, or we can from Eq. (17-5). Incidentally, we have just proved that if light travels with speed c in a system, it travels with speed c in another. For if the interval is the same in both systems, i.e., zero at one end and zero in the other, then we see that the propagation speed of light is constant, i.e., the same as saying that the interval is zero.

17-3 Past, present, and future

The space-time regions surrounding a given spacetime point can be separated into three regions, as shown in Fig. 17-1. In one region we have causal-like intervals, and in two regions, time-like intervals. Physically, these three regions also reflect space-time around a given point is caused have an interesting physical relationship to their points: a physical object or a light cone passes a point in region I in the event of moving along at a speed less than the speed of light. Therefore events in this region can affect the point O , and have an influence on it from the past. In fact, of course, an object at O , the spacetime point is precisely in the "past" with respect to O ; it is the same spacetime point, as O , only earlier. What happened there then affects O now. Unfortunately, that is all we can say. An object object at O can get to O by moving with a certain speed less than c , so if this object were to a speed slightly exceeding it would be again the past of the same spacetime point. That is, in another coordinate system, the curve of time might go through here O and O' . So all points of O are in the "past" of O , and any thing that happens in this region can affect O . Before region II is sometimes called the "future past," or affecting past; it is the locus of all events which can affect point O in any way.

Region II, on the other hand, is a region which we can affect now O , we can "hit" things he shows my "bullets" out at speeds less than c . So O is the world whereof it can be affected by us and we can still hit the object O . Now the remarkable thing about all the rest of spacetime, i.e., region III, is that we can neither affect it now/future O , nor can it affect us now in O , because nothing can go faster than the speed of light. Of course, what happens in O can affect us later; that is, if the sun is exploding ("right now"), it takes eight minutes before we know about it and it cannot possibly affect us before it has.

Why we mean by "right now" is a mysterious thing which we cannot define and we cannot affect, but it can affect us later, or we could have affected it if we had done something to enough in the past. When we look at the Alpha Centauri, we see it as it was four years ago; we might wonder what it is like "now," "now," as is the case here from our spacetime coordinate system. We can only see Alpha Centauri by the light that has come from the past, up to four years ago, but we do not know what is today "now"; it may take four years before what is today "now" can affect us. Alpha Centauri "now" is a idea or concept of our mind; it is not something that is really thinkable physically or the otherwise, because we have to wait to observe it, we cannot exceed its light "time." Furthermore, the "now" depends on the coordinate system. If, for example, Alpha Centauri were moving to please the x would no longer with us because he would not be fixed at an angle, and his "now" would be a different time. We have already talked about the fact that simultaneity is not a unique thing.

There are certain tellers of people who believe they can know the future and there is many wonderful stories about the "I" who suddenly discover a "secret" knowledge about the future. Well, this would be a paradoxical prediction by that because "we know something is going to happen, then we can make

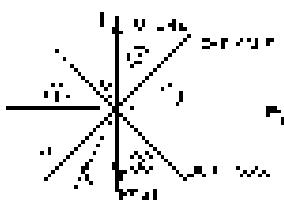


Fig. 17-1. The space-time regions surrounding a point of the origin

one we will avoid it by doing the right things. To right now, and so on. But actually there is no future tell who can even tell me if it's passed. That is to say who can tell us what is really happening right now, at any given moment. Because then is a real value. We might ask ourselves this question, which we have to the student to try to answer: Would any paradox be produced if a were suddenly to become possible to know this, that one is the spacetime coordinate system?"

17.4 More about four-vectors

Let us now return to our consideration of the analogy of the Lorentz transformation and rotations in three-space. We have been mainly to collecting up to other quantities which have the same transformation properties, as the coordinates, to both space and time, directed lines. In the case of ordinary rotations, there are many quantities in a translation, for example $\hat{x}, \hat{y}, \hat{z}$, and a linear invariant. For example, the velocity has three components, $\hat{u}_x, \hat{u}_y, \hat{u}_z$ component; when seen in a different coordinate system, none of the components is the same indeed, they will transform in this or in, say, somehow or other, however itself, has a greater variety than to any of its coordinate components, and we represent it by a directed line.

We therefore ask: Is it possible that there are quantities which transform, or which are related, in a moving system and to a non-moving system, in a way to $\hat{x}, \hat{y}, \hat{z}$ and \hat{u} ? From our acquaintance with vectors, we know that three of the quantities, like $\hat{u}_x, \hat{u}_y, \hat{u}_z$, constitute the three components of an ordinary space vector, but the fourth quantity would, "as it is," in some way be constant and three in space-time? We don't know whether there is, indeed, at least one such thing there are many of them, but that the three components of momentum and energy or the four components, therefore together to make what we call a "four-vector". In demonstrating this, since this quite accustomed to refer to \hat{u} it is everywhere, we shall use the same letter, writing in it's "the energy, the mass and the momentum, that we use in Eq. (17.1). Energy and mass, for example, differ only by a factor c^2 which is merely a question of units, so we can interchange it. In mass. Instead of trying to write it c^2 , we put $E = mc^2$, so then, of course, of this quantity it must we want to be larger amounts of also that the mass would straighten out in the last equation, but not in the intermediate ones.

This is the equation for energy and momentum are

$$\begin{aligned} E &= mc/\sqrt{1 - \hat{u}^2/c^2}, \\ p &= mv = m\hat{u}/\sqrt{1 - \hat{u}^2/c^2}. \end{aligned} \quad (17.6)$$

Again in these terms, we have

$$E^2 = p^2 + m^2c^2. \quad (17.7)$$

For example, if we measure energy in electron volts, what does a mass of an electron tell about it? If we take the mass which is energy in electron-volt, then it, we're told, is one electron volt. For example, the rest mass of an electron is 0.511×10^{-10} ev.

Now what would the momentum and energy look like in a new coordinate system? To find out, we shall have to transform Eq. (17.6) which we can do because we know how the velocity transforms. Suppose that we measure it, on this boat a velocity \hat{u} , but we look upon the same object from the point of view of a space ship which itself is moving with velocity \hat{v} , and in this system we have a point to design \hat{u}' the corresponding thing. In order to simplify things at least, we shall take the case that the velocity \hat{v} is in the direction of the \hat{x} axis, we consider the most general case. What is it, the velocity system from the space ship? It is the

component relative, the "difference" between v and v' . By the law which we worked out before,

$$v' = \frac{v - u}{1 - vu}. \quad (17.8)$$

Now let us calculate the new energy E' , the energy as it follows in the space ship would see it. He would use the same notations, of course, but he would use v' for the velocity. What we have to do is subtract u^2 , subtract it from one take the square root, and take the reciprocal:

$$\begin{aligned} v'^2 &= \frac{v^2 - 2uv + u^2}{1 - 2vu + u^2} \\ 1 - v'^2 &= 1 - 2uv + u^2 \frac{v^2 - 2uv + u^2}{1 - 2vu + u^2} \\ &= \frac{1 - v^2 - u^2 + u^2 v^2}{1 - 2vu + u^2} \\ &= \frac{(1 - u^2)(1 - v^2)}{(1 - vu)^2} \\ &= E^2/(1 - u^2). \end{aligned}$$

Therefore

$$\frac{1}{\sqrt{1 - v'^2}} = \frac{1 - vu}{\sqrt{1 - u^2}\sqrt{1 - v^2}}. \quad (17.9)$$

The energy E' is then simply m_1 , times the above expression. But we want to express the energy in terms of the original energy and momentum, and we know that

$$E = \frac{m_1 - m_1 u^2}{\sqrt{1 - u^2}\sqrt{1 - v^2}} = (m_1/\sqrt{1 - u^2}) - (m_1 u/\sqrt{1 - u^2})v,$$

or

$$E' = \frac{1 - vu}{\sqrt{1 - u^2}}, \quad (17.10)$$

which we recognize as being exactly of the same form as

$$v' = \frac{1 - vu}{\sqrt{1 - u^2}}.$$

Next we must find the new momentum p'_x . This is just the energy E' times v' , and is also simply expressed in terms of E and p_x :

$$p'_x = E'v' = \frac{m_1(1 - vu)}{\sqrt{1 - u^2}\sqrt{1 - v^2}} \frac{v - u}{(1 - vu)} = \frac{m_1(1 - vu)}{\sqrt{1 - u^2}\sqrt{1 - v^2}}.$$

Thus

$$p'_x = \frac{p_x - mE}{\sqrt{1 - u^2}}, \quad (17.11)$$

which we recognize as being of precisely the same form as

$$v' = \frac{v - u}{\sqrt{1 - u^2}}.$$

Thus the transformations for the new energy and momentum in terms of the old energy and momentum are exactly the same as the transformations for v in terms of v' , and v' in terms of v and u . All we have to do is, every time we see v in (17.4) substitute E , and every time we see v substitute p_x , and then the equations (17.4) will become the same as Eqs. (17.10) and (17.11). This would apply, if everything works right, at additional rate that $p'_x = p_x$ and that $v'_x = v_x$. To prove this would require you going back and studying the issue of motion up and down. Actually, we did study the case of rolling up and down in the last

chapter. We imagined a head-on collision and we noticed that, i.e., the components of momentum were changed when viewed from a moving system so we have already written that $\hat{p}_A' = p_A$ and $\hat{p}_B' = p_B$. The complete transformation, then, is

$$\begin{aligned} \hat{p}_A' &= p_A - \frac{vC}{\sqrt{1 - v^2}}, \\ p_A' &= p_A \\ \hat{p}_B' &= p_B \\ p_B' &= \frac{p_B + vp_A}{\sqrt{1 - v^2}}. \end{aligned} \quad (17.10)$$

In these transformations the time, v , has disappeared from p_A' which transforms like (x, t) , and which would be given by the equation $t = \gamma x$. Since the momentum is a four-vector, it can be represented on a space-time diagram as a moving vector as a "twice" four-vector, the path, as shown in Fig. 17.4. This arrow has a time component equal to the energy, and its space components represent a three-vector momentum, this arrow is more "real" than either the energy or the momentum, because there has to be a path in the diagram.



Fig. 17.4. The four-vector representation of a particle.

17.5 Four-vector algebra

The notation for four-vectors is different than that for three vectors. In the case of three-vectors, if we want to talk about the ordinary 3-vector momentum we could write it p . If we wanted to be more specific, we could say p_x , p_y , p_z , three components which are, for the case in question, p_x, p_y , and p_z , so we could simply say the general momentum p , or say that it is the 3-vector p , or that there are the three components p_x, p_y, p_z , imagine that p is any one of three directions, x, y , or z . The notation that we use for four-vectors is analogous to this. We write p , p_A , p_B , the four-vector, and p_{Ax} , p_{Ay} , p_{Az} for the four parallel directions x, y, z , or x, y, z .

We could, of course, use any notation we want; do not laugh all the time; just think, it's all possible. In fact, in other areas, we change our definition of four-vectors. The whole idea of a four-vector, in fact, is an improvement in notation so that the transformations can be completed easily. p_A , then, is a general four-vector, but for the specific case of energy, p_A , p_A is identified as the energy p_0 is the momentum in the x -direction, p_x , and it is p -vectorial, and p_1 is that in the y direction. Overall, for p_A , we add the corresponding components.

There is an equation among four vectors, when the energy is zero, for each component. For instance, the law of conservation of three-vector momentum is to be unchanged in a collision, i.e., if the sum of the momenta for a large number of interacting or colliding particles is to be a constant, and must mean that the sums of the momenta in the x -direction, in the y -direction, and in the z -direction, for all the particles must each be constant. This law alone would be impossible in relativity because c is a constant, it's like trying about only two of the components of p to be constant. I.e., it's impossible to do. If we take this case, we add the various components, so we must include all three components in our law. That is, in relativity, we must complete the law of conservation of momentum by exceeding it to include the time component. This is after all, however, to go with the other three, or that cannot be realistic importance. The conservation of energy is the fourth equation which goes with the same value of momentum to make a valid four-vector relationship in the geometry of space and time. That is the law of conservation of energy and momentum. In our discussion, condition is

$$\sum_{p \in p_A} \hat{p}_i = \sum_{p \in p_B} \hat{p}_i \quad (17.11)$$

or, in a slightly different notation

$$\sum_i p_{Ai} = \sum_j p_{Bj} \quad (17.12)$$

where $i = 1, 2, \dots$ refers to the particles going into the collision, $j = 1, 2, \dots$ refers to the particles coming out of the collision, and $\alpha = r, p, \pi$, or γ (you say "the weak ones?"). It makes no difference. The law is true for good occupation, using δ 's and ϵ 's.

In vector language we discussed one other thing, the dot product of two vectors. Let us even consider the corresponding thing in spacetime. In ordinary motion we discovered there was no invariant quantity $x^2 + y^2 - z^2$. In four dimensions, we find that the corresponding quantity is $t^2 - x^2 - y^2 - z^2$ (Eq. 17.5). How can we verify that? One way would be to represent four-dimensional things with a space-dot between, like $A_\mu \odot B_\nu$, and α^μ . In notation which is actually used is

$$\sum_i A_\mu A_\nu - \delta_{\mu\nu} = J_1^2 - J_2^2 - J_3^2. \quad (17.15)$$

The point on Σ where there is rest mass, the "time" term, is just $-c$, but the other three terms have mixed signs. This quantity then is to be the same in any coordinate system, and we may call it the square of the length of the four-vector. But how do we know what is the square of the length of the four-vector in spacetime, for a single particle? This will be equal to $p_1^2 - p_2^2 - p_3^2 = p_1^2$, or else we do $E^2 - p^2$, because we know $p_1^2 = E^2$. What is $E^2 - p^2$? It is the same long which is the same in every coordinate system. In a vacuum it must be the same for a nonrelativistic wave, which is moving right along with the particle, i.e., which is going right by standing still. If the particle is standing still, it would have no momentum. So in this coordinate system, it's particle its energy, which is zero as its rest mass. Thus $E^2 - p^2 = m^2$, so we see that the square of the length of this vector, the four-vector momentum, is equal to m^2 .

From the square of a vector, we can get in turn the "dot product," or the product which we want: " a_μ is one four-vector and b is another." Consider, for instance, the scalar product

$$\sum_i a_\mu b_\nu - \delta_{\mu\nu} = a_0 b_0 - a_1 b_1 - a_2 b_2 - a_3 b_3. \quad (17.16)$$

It is the same in all coordinate systems.

Finally, we shall mention something whose rest mass is 0, zero. A photon of light. For example, a photon is like a particle in that it has an energy and a momentum. The energy of a photon is proportional to its Planck's constant, times the frequency of the photon: $E = h\nu$. Each a photon also carries a momentum, and the momentum of a photon of some other frequency in fact is twice by the wavelength: $p = h/\lambda$. But, for a photon, there is a definite relation between the frequency and the wavelength: $\nu = c/\lambda$. The number of waves per second, times the wavelength of each, is the distance that the light goes in one second, which, of course, is c . This makes immediately that the energy of a photon, and its the momentum times c , are. If $c = 1$, the energy and momentum are equal. That is to say, the rest mass is zero. This is called in this way "zero mass." It is a particle of zero rest mass, with respect to it, it doesn't have any mass. It always goes at the speed c . The total formula for energy is $E = h\nu/c$. Now you see; $E = mc^2$, and $c = \infty$, so the energy is 0! We cannot say that it is zero, the photon really can't do any work, even though it has no mass, but the it possesses by nevertheless using the speed of light.

We also know that the momentum of any particle is equal to its total energy times its velocity if $c = 1$, $p = E$, or, in ordinary units, $p = mc^2/c$. For an electron moving at the speed of light $p = E/c = 1$. The Compton formula, etc. etc. etc. we seen from a moving system, of course given by Eq. 17.12, can tell the momentum we must subtract the energy times c (in this case $c = 1$) to keep the different energies after transmission means that there are different frequencies. This is called the Doppler effect, and one can calculate it easily from Eq. 17.12 using also $c = 1$ and $E = mc^2$.

As Minkowski said, "Space of itself, and time of itself will think and never understand, and only a live neuron between them shall survive."

Rotation in Two Dimensions

18-1 The center of mass

In the previous chapters we have been studying the mechanics of points, in small particles whose internal structure does not concern us. For the next few chapters we shall study the application of Newton's laws to more complicated things. When the word becomes more complicated, it also becomes more interesting, and we shall find that the phenomena associated with the mechanics of a more complex object than just a point are really quite striking. Of course these phenomena involve nothing but the kinematics of Newton's laws, but it is sometimes hard to believe that only *that* can be so much.

The more complicated objects we deal with can be of several kinds. Water flowing, galaxies, whirligigs, etc. The simplest "complicated" object is you, or, to start, is what we call a rigid body, a solid object that is moving as it moves again. However, even such a simple object may have a most complex motion, and we shall explore that outside the simplest aspects of such motion, in which an extended body rotates about itself. A given point on such a body then moves in a plane perpendicular to this axis. Such rotation of a body about a fixed axis is called *pure rotation*, *rotation in one dimension*. We shall later generalize the idea to three dimensions, but in doing so we shall find that, unlike the case of ordinary particle mechanics, rotations are subtle and hard to understand unless we first get a solid grounding in two dimensions.

The first interesting theorem concerning the motion of complicated objects can be observed at work if we throw an object made of a lot of blocks and gravel tied together by strings, into the air. Of course we know it goes in a parabola, because we studied that for a particle. But now our object is *not* a particle, it wobbles and it jiggles and so on. It goes up in a parabola. And, one sees now that this goes up a parabola! Certainly not the point on the corner of the brick, because that is jiggling about; rather it is the end of the wooden stick, or the middle of the wooden stick, or the middle of the block. But something goes up a parabola, there is an effective "center" which moves in a parabola. So our first theorem about complicated objects is to demonstrate that there is a mean position which is mathematically definite but not necessarily a point of the material itself, which goes in a parabola. That is called the theorem of the center of the mass, and the proof of it is as follows.

We may consider any object as being made of lots of little particles, the atoms, with various forces among them. Let i represent an index which denotes one of the particles, (there are millions of them, say 10^{23} , or something); Then the force on the i th particle is, of course, the sum since the acceleration of that particle:

$$\mathbf{F}_i = \sum_j (m_j \mathbf{a}_j) \quad (18.1)$$

In the next few chapters our moving objects will be such as which all the parts are moving at speeds very much lower than the speed of light, and we shall use the non-relativistic approximation for γ quantities. In these circumstances the mass is constant so that

$$\mathbf{F}_i = m_i \mathbf{a}_i / m_i^2 \quad (18.2)$$

Now if we add the forces on all the particles, that is, if we take the sum of all the \mathbf{F}_i 's for all the different indices, we get the total force \mathbf{F} . On the other side of the

18-2 The center of mass

18-2-1 Rotation of a rigid body

18-3 Angular momentum

18-4 Conservation of angular momentum

equation, we get the same thing as though we added before the differentiation:

$$\sum_i m_i \ddot{r}_i = \frac{d}{dt} (\sum_i m_i r_i), \quad (18.3)$$

Therefore the total force is the second derivative of the mass-times-their-position, added together.

Now the total force on all the particles is the same as the external force. Why? A. Through the air all kinds of forces act on the particles because of the strings, the wiggles, the pushing and pulling, and the elastic forces, and when it comes time to add we have to add all these together, we are reduced by Newton's Third Law. Between any two particles the elastic force tends to expand, so that when we add all the elastic forces together, if one particle has forces being exerted on it which cancel out in the sum, therefore the net result is only those forces which arise from other particles which are not included as separate objects; we should be sum. So P Eq. (18.3) is the second derivative of the center of mass, which together is called "the center of mass of the system of particles." The total object is equal to the sum of all the forces on all its constituent particles.

Now it would be nice if we could do in Eq. (18.3) as the total mass times some vector, say \vec{R} . Let us say it is the sum of all the masses, i.e., the total mass. Then \vec{R} will differ a certain amount. That is,

$$\vec{R} = \sum_i m_i \vec{r}_i / M, \quad (18.4)$$

Eq. (18.3) will be simply

$$\vec{F} = d^2(M\vec{R})/dt^2 = M(\vec{r}^2 \vec{R}/M^2), \quad (18.5)$$

since M is a constant. Thus we find that the external force is the total mass times the acceleration of an imaginary point whose position is \vec{R} . This point is called the center of mass of the body. It is a point somewhere in the "middle" of the object, a zone of average \vec{r} in which the different parts have weights or importances proportional to the masses.

We shall discuss this important theorem in greater detail in a later chapter, and we shall then see how our remarks fit two points. First, if $\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N$ are the positions of the various particles in an empty space ship - right side, and wiggly and twisty, and do all kinds of things. But, if the center of mass, this is definitely located, call it \vec{R} , at a definite position, somewhere in the middle. And now what is constant about \vec{R} ? In particular, if it is initially at rest, it will stay at rest. So, if we have come out of a box, perhaps a space ship with people in it, and we calculate the location of the center of mass and find that it is still, then the center of mass is constant, so no ship will fly out of control if forces are acting on it. Of course, the space ship may move a little in space, but that is because the people are walking back and forth inside; when one walks toward the front, the ship goes toward the back since to keep the average position, all the masses *at exactly the same place*.

For rocket propulsion therefore absolutely impossible because one cannot move the center of mass. Now, but of course we find that in propelling an interesting part of the rocket, the uninteresting part must be thrown away. In other words, if we can, with a rocket of zero velocity and we spit + one plus another rocket more than 1000 miles of gas per minute, as the rocket ship goes faster, but the center of mass is still exactly where it was before. So we simply move the part that we let unaccelerated. But part we want to is accelerated.

The second point is inserting the center of mass, which is the reason we introduced it into our discussion. This time, is this if any we need separately from the "inertial" motion of the object, and may therefore be ignored in our discussion of rotation.

18-2 Translation of a rigid body

Now let us discuss rotation. Of course an ordinary object does not simply rotate, it translates, pushes one hands, or to simply rotates we shall discuss the motion of a non-existent ideal object which we call a rigid body. This means no

object, in which the forces between the atoms are so strong, and of such character, that the little forces that are needed to move it do not have to act over any extended distance; it moves almost if we wish to study the motion of each particle in the body, and ignore the motion of its center of mass, there is only one thing left for it to do, and that is to turn. We have to describe that. Now, suppose there is some line in the body which stays put (perhaps it includes the center of mass and perhaps not), and the body is rotating about this particular line as we see. How can we define this rotation? That is easy enough, for if we mark a point somewhere on the object, say P , we can always tell exactly where the object is, if we only know where this point has gone to. The only thing needed to describe the position of P at t is an angle, so rotation consists of a study of the variations of the angle with time.

In order to study rotation, we observe the angle through which a body has turned. Of course, we are not referring to any particular angle inside the object itself; it is not, but we know some angle of the object. We are talking about the angular change of the position of the whole thing, from one time to another.

First, let us study the kinematics of rotation. The angle will change with time, and just as we talked about position and velocity in the case of linear motion, so may we talk about angular position and angular velocity in the case of rotation. In fact, there is a very interesting relationship between rotation in two dimensions and three-dimensional displacement, in which almost every quantity has its analog. Thus, we take the angle θ which defines how far the body has gone around; this replaces the distance s , which defines how far it has gone along. In the same manner, we have a velocity of turning, $\omega = d\theta/dt$, which tells us how much the angle changes in a second, just as $v = ds/dt$ describes how fast a thing moves, or $a = dv/dt$ gives us its acceleration. The angle is measured in degrees, then the angular velocity ω will be ω and ω many radians per second. The greater the angular velocity, the faster the object is turning. The faster the angle changes, we can go on; we can differentiate the angular velocity with respect to time, and we can call $\alpha = d\omega/dt$ ω -rate of the angular acceleration. This would be the analog of the ordinary acceleration.

Now of course we shall have to relate the dynamics of rotation to the laws of dynamics of the particles of which the object is made, or we must find out how a particular particle moves when the angular velocity as such and such. To do this, we take a system particle which is located at a distance r from the axis and say this is at a certain location $P(x, y)$ at a given instant, on the x -axis, say, at $(r, 0)$. If it is a moment $d\theta$ later the angle $\theta + d\theta$ the whole object has turned through $d\theta$, then this particle is carried with it. It is at the same radius away from O as it was before, but is rotated by $d\theta$. The first thing we would like to know is how much the x -displacement and y -displacement y changes. It will be clear, then, that x changes by $-r\omega d\theta$, because of the way angles are defined. The change in y , then, is simply the projection of $r d\theta$ in the y -direction, i.e.

$$dy = -r d\theta \sin \theta = -r \omega d\theta \quad (18.6)$$

Similarly,

$$dx = r \omega d\theta \quad (18.7)$$

If the object is turning with a given angular velocity ω , and, by writing both Eqs. (18.6) and (18.7) by $d\theta$, the velocity of the particle is

$$v_x = -ry = -rd\theta \quad v_y = rx \omega = rd\theta \quad (18.8)$$

Of course if we want to find the magnitude of the velocity, we add since

$$v = \sqrt{v_x^2 + v_y^2} = \sqrt{(rd\theta)^2 + (rd\theta)^2} = \omega(r^2 + r^2) = \omega r \quad (18.9)$$

It should not be exaggerated that the value of the magnitude of this velocity is not ωr , it would be self evident, because by definition r moves $\omega r d\theta$ and the distance it moves per second is $\omega r d\theta$, or ωr .

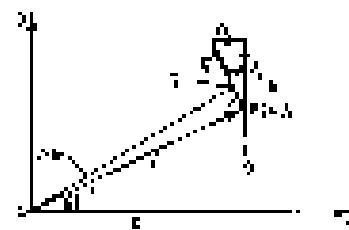


Fig. 18-1. Kinematics of two-dimensional rotation.

Let us now move on to consider the dynamics of rotation. There is one concept force must be introduced. We begin with the notion of work, which we first call the *work* (i.e., negative, re-twisted) when there is no displacement to rotate or to move due to linear movement. A force is the thing that is needed to make linear motion, and the thing that makes something rotate is a "rotary force" or a "rotating force," i.e., a torque. Quantitatively, a torque is a "force" applied to a torque about a pivot? We shall get to the theory of torque quantitatively by studying the work done in rotating an object. We can very easily way of defining a force is to say how much work it does when it acts through a given displacement. We are going to try to relate the analogy between linear and angular quantities by equating the work that we do when we turn something (i.e., in which case we have acting on it) to the torque times the angle it turns through. In other words, the definition of the torque is going to be $\tau = \text{work}/\theta$. And that the linear situation has an absolute analog: force times distance is work, and torque times angle is going to be work. That will be what torque is. Consider, for instance, a rigid body of some kind with various forces acting on it and an axis about which the body rotates. Let us at first concentrate on one force and suppose that this force is applied at a certain point (x, y). How much work would we do? This would be zero. The object through a very small angle! That is, zero. The work done is:

$$\Delta W = F_x \Delta x + F_y \Delta y \quad (13.10)$$

We do not only "resist" (see Eqs. (13.6) and (13.7)) but also add to obtain

$$\Delta W = (F_x^2 + F_y^2)^{1/2} \Delta s \quad (13.11)$$

This is, the amount of work that we have done is in fact equal to the angle through which we have turned the object, multiplied by a strange-looking combination of the force and its distance. This "strange-looking ratio" is what we call the torque. By defining the change in work as the torque times the angle, we have been the formula for torque in terms of the forces. Obviously, torque is not a completely new idea, independent of Newtonian mechanics—just now there is definite definition in terms of the forces.

When there is a vector? To do work, the work that is done is, of course, the sum of the works done by all the forces, plus the work will be a whole lot of work if added together, the sum the forces, each of which is proportional, however, to Δs . We can take the same procedure and therefore can say that the change in the work is equal to the sum of all the torques due to all the different forces that are acting, i.e., ΔW . This sum we might call the total torque, τ . Thus torque will be the moment of force of all the forces, but we shall find, see that this is only because we are working in a plane. It is the unopposed moment is, when other forces simply will affect directly, but only because they are all in the same direction. It is, then, compensated in three dimensions. Thus, for two-dimensional rotation,

$$\Delta W = \tau \Delta \theta_1 + \tau \Delta \theta_2 \quad (13.12)$$

and

$$\tau = \sum \tau_{ij} \quad (13.13)$$

It must be emphasized that the torque is the *in* a given axis. If a different axis is chosen, so that all the x_i and y_i are changed, the value of the torque is *not* by changed too.

Now we pause briefly to note that our foregoing introduction of torque, through the idea of work, gives us a much simpler result for an object in equilibrium: if all the forces on an object are in balance (i.e., if translation and rotation, not only is the net force zero, but the total of all the torques is also zero), *then* if an object is in equilibrium, it will rotate by the forces for a small displacement. Therefore, since $\Delta W = \tau \Delta \theta = 0$, the sum of all the torques must be zero. So there are two conditions for equilibrium: (1) the sum of the forces is zero, *and* then the sum of the torques is zero. Prove that it suffices to assume that the sum of torques need only one axis (in two dimensions) is zero.

Now let us consider a single force, and try to figure out exactly what this torque thing is. In Fig. 18-2 we see a force \mathbf{F} at a point P , where the object has rotated through a small angle $d\theta$. The total torque, of course, is the component of force in the direction of the displacement. In other words it is only the tangential component of the force that counts, and this must be multiplied by the distance $r \sin d\theta$. Therefore we see that the torque is also equal to the tangential component of force perpendicular to the radius times the radius. That makes sense in terms of our ordinary idea of the torque; indeed, if the force were completely radial, it would not produce any "twist" on the body; it is evident that the twisting effect should involve only the part of the force which is not pulling out from the center, and that means the tangential component. Furthermore, it is clear that a given force is more effective on a long arm than near the axis. In fact, in the case where we push right on the axle we do not twist it at all, so it makes sense that the amount of twist, or torque, is proportional both to the radial distance and to the tangential component of the force.

There is a 4.7 x 10⁻¹² coulombs for the torque which is very interesting. We have just seen that the torque is the force times the radius times the sine of the angle θ in Fig. 18-2. But if we extend the line of action of the force and draw the line OS , the perpendicular distance to the line of action of the force (the lever arm of the force) is twice that of the lever arm r shown; that is, in just the same proportion as the tangential part of the force is to the total force. Therefore the formula for the torque can also be stated as the magnitude of the force times the length of the lever arm.

The torque is also of course just the moment of the force. The origin of this word is obscure, but it may be related to the fact that "moment" is derived from the Latin *momentum*, and that the capacity of a force to move an object (using the word in its original sense) increases with the length of the lever arm. In addition, the "moment" means weighted by how far away it is from an axis.

18-3 Angular momentum

Although we have so far considered only the special case of a rigid body, the properties of torque and their mathematical relationships are interesting also even when an object is not rigid. In fact, we can prove a very remarkable theorem: just as external force is the rate of change of a quantity p , which we call the total momentum of a collection of particles, just as external torque is the rate of change of a quantity L which we call the angular momentum of the group of particles.

To prove this, we shall suppose that there is a system of particles on which there are some forces acting and find out what happens to the system as a result of the torque due to these forces. First, of course, we assume without just one particle. In Fig. 18-3 is an particle of mass m , and at axis O ; the particle is necessarily moving in a circle about O , it may be moving in an ellipse like a planet going around the sun, or in some other curve. It is moving somehow, and there are forces on it, and it accelerates according to the usual formula that the component of force is the derivative of the component of position, etc. But let us see what the torque does. The torque equals $\mathbf{r} \times \mathbf{F}_t$, and the force in the x -or y -direction is $m a_x$ times times the acceleration in the x -or y -direction:

$$\begin{aligned}\tau &= \mathbf{r} \times \mathbf{F}_t = m \mathbf{r}_t \\ &= m \omega^2 r \hat{i} \cos \theta - j m \omega^2 r \hat{i} \sin \theta\end{aligned}\quad (18.14)$$

Now, although this does not appear to be the derivative of any single quantity, it is in fact the derivative of the quantity $m \omega^2 r^2 \hat{i} = m \omega^2 \hat{i} / 2$.

$$\begin{aligned}\frac{d}{dt} \left[m \left(\frac{\partial \theta}{\partial t} \right) - i m \left(\frac{\partial x}{\partial t} \right) \right] &= m \left(\frac{d^2 \theta}{dt^2} \right) - \left(\frac{dr}{dt} \right) m \left(\frac{\partial y}{\partial t} \right) \\ &= m \omega^2 \left(\frac{\partial^2 \theta}{\partial t^2} \right) - \left(\frac{dr}{dt} \right) m \left(\frac{\partial^2 x}{\partial t^2} \right) = m \omega^2 \left(\frac{\partial^2 \theta}{\partial t^2} \right).\end{aligned}\quad (18.15)$$

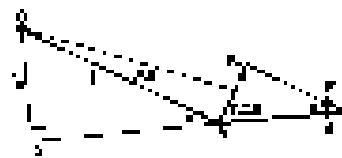


Fig. 18-2. The torque produced by a force.

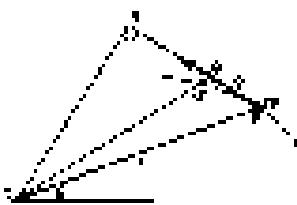


Fig. 18-3. A particle moves about an axis O .

So it is true that the torque is the rate of change of something with time! So we say again that the "something," we just gave a name, we call it L , the angular momentum.

$$L = \text{momentum} \times \text{position}/\text{mass}$$

$$\therefore \tau_{\text{ext}} = \partial L/\partial t \quad (18.16)$$

Although our present discussion is incomplete, the second form for L , given above is definitely correct. So we have found that there is also a conservation law for the angular momentum, that τ 's torque, the angular momentum, is given by an expression in terms of the components of linear momentum that is just like the formula for torque in terms of the force component! Thus, if we want to know the angular momentum of a particle about a axis, we take only the component of the moment arm that is perpendicular, and multiply it by the radius. In other words, what counts for angular momentum is not how far it is going away from the origin, i.e., how much it is going around the origin. Only the perpendicular part of the moment arm counts for angular momentum! Furthermore, the further out the line of the moment arm extends, the greater the angular momentum. And also, because the geometrical facts are the same whether the system is liberal τ or F , it is true that there is a lever arm (or the distance the lever arm of the force on the particle) which is obtained by extending the line of the moment arm, and dividing the perpendicular distance by the axis. But the angular momentum is the magnitude of the contribution times the moment arm lever arm. So we have three formulas for angular momentum, just as we have three formulas for the torque:

$$L = \tau_{\text{ext}} \times \Delta t$$

$$= \text{lever arm} \times \text{force}$$

$$= \text{lever arm} \times \text{mass} \times \text{velocity}. \quad (18.17)$$

Like torque, angular momentum depends upon the position of the axis about which it is to be calculated.

Before proceeding to a treatment of more exact and general, let us apply the above results to a planet going around the sun. In which direction is the force? The force is toward the sun. What, then, is the torque on the object? Of course, the object is not going to be at the axis, but we get a very simple result if we take it to the sun lines, for the torque is the force times the lever arm, or the component of force perpendicular to a times r . But there is an tangential force, so there is no torque about an axis other than. Therefore, the angular momentum of the planet going around the sun must remain constant. Let us see what that means. The tangential component of velocity, since the mass, m , is the same, will be constant, because that is the angular momentum, and the rate of change of angular momentum is the torque, and in this problem, the torque is zero. Of course, since the mass is also a constant, this means that the angular velocity times the radius is a constant. That is a something we already knew for the motion of a planet. Suppose we consider a small amount of time Δt . How far will the planet move when it moves from P to Q ? (Fig. 18-5) If we make Δt very small, we can sweep through the area QPF compared with the much larger area OPF , it is evidently half the base PQ times the height, OP . In other words, the area that is swept through in unit time will be equal to the velocity times the lever arm of the velocity plus one half r . Thus the rate of change of area is proportional to the angular momentum, which is constant. So Kepler's law alone, applied to equal areas in a short description of the statement of the law of conservation of angular momentum, when there is no longer granted to the laws.

18-4 Conservation of angular momentum

Now we shall give an example what happens when there is a large number of particles, such as object is made of many masses with many forces acting between them and on them from the outside. Of course, we already know that even for given τ we will have different angular particle (which is the mass or particle per its

comes from sum of $m_i \vec{r}_i \times \vec{\omega}$ due to the rate of change of the angular momentum of the particle, and that the angular momentum of the i particle is $\vec{m}_i \vec{v}_i \times \vec{r}_i$ (momentum lever arm). Now suppose we add the i particles, i.e. all the particles are still in the total torque τ . Then this will be the rate of change of the sum of the angular momenta of all the particles, and this defines a new quantity which we call the total angular momentum \vec{L} . Just as the total momentum of a object is the sum of the momenta of all the parts, so the angular momentum is the sum of the angular momenta of all the parts. Thus the rate of change of the total L is the total torque

$$\tau = \sum \tau_i, \quad \sum \frac{d\vec{L}_i}{dt} = \frac{d\vec{L}}{dt}. \quad (18.18)$$

Now it might seem that this total torque is a complicated thing. There are n of these in total, forces and n of the couplet forces to be considered. But, if we take Newton's law of action and reaction we see, not surprisingly, that the action and reaction are equal, but also that they are directed exactly oppositely along the same line. However, one of them may be zero, say $\vec{r}_i \times \vec{F}_{i\text{ext}}$, but we easily assume it. Then the two couplets in the reaction region due to other central interactions will be equal and opposite because the lever arms for any two are equal. Therefore, the external torque balance you pair by pair, and so we have: The remarkable theorem that the rate of change of the total angular momentum about any axis is equal to the external torque about that axis!

$$\tau = \sum \tau_i = \tau_{\text{ext}} = dL/dt. \quad (18.19)$$

Thus we have a very powerful theorem controlling the motion of large collections of particles, which permits us to study the overall motion without having to look at the detailed trajectory inside. The theorem is true for any collection of objects, whether they form a rigid body or not.

One extremely simple but nice case of this new theorem is the law of conservation of angular momentum: if no external couples act upon a system of particles, the angular momentum is thus constant.

A special case of great importance is that of a rigid body, that is, an object of a definite shape that is just turning around. Consider an object that is fixed to some axis of dimension, and which is rotating about a fixed axis. We can just set the object how the same body would do nothing at all. Note that it's difficult to tell what angular momentum of this object. If the mass of one of its particles is m , and its position in location is at (x_i, y_i) , then the problem is to find the angular momentum of that particle, because the total angular momentum is the sum of the angular momenta of all such particles in the body. For an object going around in a circle, the angular momentum of one is \rightarrow the mass times the velocity times the distance from the axis, and the velocity is equal to the angular velocity times the distance from the axis:

$$J_i = m v_i r_i = m \omega r_i \quad (18.20)$$

or, summing over all the particles i , we get

$$I = J_i. \quad (18.21)$$

where

$$I = \sum_i m v_i r_i. \quad (18.22)$$

This is the analog of the fact that the momentum is mass times velocity. Velocity is replaced by angular velocity, and we see that the mass is replaced by some thing which we call the moment of inertia I , which is analogous to the mass. Equations (18.21) and (18.22) mean that a body has inertia for twisting which depends not just on the mass, but on how far away they are from the axis. So, if we have two objects of the same mass, when we put the mass further away from the axis, the inertia for twisting will be higher. This is easily demonstrated by the apparatus

look in Fig. 18-4, where a weight M is kept from falling very fast because it has to turn. It being weighted now, it turns, causes us to use less by the units, and M speeds up at a constant rate. But when we change the moment of inertia by putting the two masses much farther away from the axis, then we see that M accelerates much less rapidly than it did before, because the body has much more inertia against turning. The amount of inertia is the inertia against turning, and is the sum of the contributions of all the masses, times their distance squared from the axis.

There is one important relation between mass and moment of inertia which is very dramatic. The mass of an object never changes, but its moment of inertia can be changed. If we stand on a转台 (a rotatable stand with our arms unextended, and hold some weight M in our hands) and rotate slowly, we may change our moment of inertia by drawing our arms in, but our mass does not change. When we do this, all kinds of wonderful things happen, because of the law of the conservation of angular momentum: if the net external torque is zero, then the angular momentum of the system remains constant. Initially we were rotating with a large moment of inertia I_1 at a low angular velocity ω_1 , and the angular momentum was $I_1\omega_1$. Then we changed our moment of inertia by pulling our arms in, say to a smaller value I_2 . Then the way in which has to go is the same because the total angular momentum has to stay the same, so $I_2\omega_2 = I_1\omega_1$. Using that, if we reduce the moment of inertia, we have to increase the angular velocity.

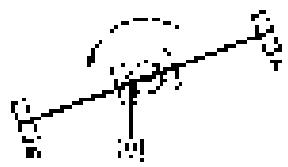


Fig. 18-4. The inertia for rotating depends upon the lever arm of the inertia.

Center of Mass; Moment of Inertia

19-1 Properties of the center of mass

In the previous chapter we found that if a great many forces are acting on a complicated mass of particles, whether the particles comprise a rigid or a nonrigid body, or a cloud of stars, or anything else, and we find the sum of all the forces (that is, of course, the external forces, because the internal forces balance out), then if we consider the body as a whole, and say it has a total mass M , there is a certain point "inside" the body, called the center of mass, such that the net resulting external force produces an acceleration of this point, just as though the whole mass were concentrated there. Let me now discuss the center of mass in a little more detail.

The location of the center of mass (abbreviated CM) is given by the equation

$$y_{CM} = \frac{\sum m_i y_i}{\sum m_i}. \quad (19.1)$$

This is, of course, a vector equation which is really three equations, one for each of the three directions. We shall consider only the x -direction, because if we can understand that, we can understand the other two. What does $X_{CM} = \sum m_i x_i / \sum m_i$ mean? Suppose for a moment that the object is divided into little pieces, all of which have the same mass m ; then the total mass is simply the number N of pieces times the mass of one piece, say one gram, or any unit. Then this equation simply says that we add all the x 's, and then divide by the number of pieces that we have added: $X_{CM} = m \sum x_i / mN = \sum x_i / N$. In other words, X_{CM} is the average of all the x 's, if the masses are equal. But suppose one of them were twice as heavy as the others. Then in the sum, that x would come in twice. This is easy to understand, for we can think of the total mass as being split into two equal parts, just like the others; then in taking the average, of course, we have to count that a twice heavier there are two masses there. Thus X is the average position, in the x -direction, of all the masses, the' mass being counted a number of times proportional to the mass, as though the object were divided into "little grams." From this it is easy to prove that X must be somewhere between the largest and the smallest x , and, therefore lies inside the envelope enclosing the entire body. It does not have to be in the material of the body, for the body could be a hoop, and the center of mass is in the center of the hoop, not in the hoop itself.

Of course, if an object is symmetrical in some way, for instance, a rectangle, so that it has a plane of symmetry, the center of mass lies somewhere on the plane of symmetry. In the case of a rectangle there are two planes, and that balances it adequately. But if it is just any symmetrical object, then the center of gravity lies somewhere on the axis of symmetry, because in those circumstances there are as many positive as negative x 's.

Another interesting proposition is the following very obvious one. Suppose that we imagine an object to be made of two pieces, A and B (Fig. 19-1). Then the center of mass of the whole object can be calculated as follows. First, find the center of mass of piece A, and then of piece B. Also, find the total mass of each piece, M_A and M_B . Then consider a new problem, in which a point mass M_A is at the center of mass of object A, and another point mass M_B is at the center of mass of object B. The center of mass of these two point masses is then the center of mass of the whole object. In other words, if the centers of mass of various parts

19-1 Properties of the center of mass

19-2 Locating the center of mass

19-3 Finding the moment of inertia

19-4 Rotational kinetic energy



Fig. 19-1. The CM of a compound body lies on the line joining the CM's of the two composite parts.

of an object have been worked out, we do not have to start all over again to find the center of mass of the whole object; we just have to put the pieces together, treating each one as a point mass situated at the center of mass of that piece. Let us see why that is. Suppose that we wanted to calculate the center of mass of a composite object, some of whose particles are considered to be members of object A and some members of object B . The total sum $\sum m_i x_i$ can then be split into two pieces—the sum $\sum_{i \in A} m_i x_i$ for the A object only, and the sum $\sum_{i \in B} m_i x_i$ for object B only. Now if we were computing the center of mass of object A alone, we would have exactly the first of these sums, and we know that this by itself is $M_A X_A$, the total mass of all the particles in A times the position of the center of mass of A , because that is the theorem of the center of mass, applied to object A . In the same manner, just by looking at object B , we get $M_B X_B$, and of course, adding the two yields $M X$:

$$M X_{\text{tot}} = \sum_i m_i x_i + \sum_n m_n x_n \\ = M_A X_A + M_B X_B. \quad (19.2)$$

Now since M is evidently the sum of M_A and M_B , we see that Eq. (19.2) can be interpreted as a special example of the center of mass formula for two point objects, one of mass M_A located at X_A and the other of mass M_B located at X_B .

The theorem concerning the motion of the center of mass is very interesting, and has played an important part in the development of our understanding of physics. Suppose we assume that Newton's law is right for the small component parts of a much larger object. Then this theorem shows that Newton's law is also correct for the larger object, even if we do not study the details of the object, but only the total force acting on it and its mass. In other words, Newton's law has the peculiar property that if it is right on a certain small scale, then it will be right on a larger scale. If we do not consider a baseball as a tremendously complex thing, made of myriads of interacting particles, but study only the motion of the center of mass and the external forces on the ball, we find $F = ma$, where F is the external force on the baseball, m is its mass, and a is the acceleration of its center of mass. So $F = ma$ is a law which reproduces itself on a larger scale. (There ought to be a good word, out of the Greek, perhaps, to describe a law which reproduces the same law on a larger scale.)

Of course, one might suspect that the first laws that would be discovered by human beings would be those that would reproduce themselves on a larger scale. Why? Because the actual scale of the fundamental gears and wheels of the universe are of cosmic dimensions, which are at much finer than our observations that we are moreover near that scale in our ordinary observations. So the first things that we would discover must be true for objects of no special size relative to all atomic scales. If the laws for small particles did not reproduce themselves on a larger scale, we would not discover these laws very easily. What about the reverse problem? Must the laws on a small scale be the same as those on a larger scale? Of course it is not necessarily so in nature, that at an atomic level the laws have to be the same as on a large scale. Suppose just the true laws of motion of atoms were given by some strange equation which does not have the property that when we go to a larger scale we reproduce the same law, but instead has the property that if we go to a larger scale, we can approximate it by a certain expression such that, if we extend that expression up and up, it keeps reproducing itself on a larger and larger scale. That is possible, and in fact that is the way it works. Newton's laws are the "last cut" of the atomic laws extrapolated to a very large size. The actual laws of motion of particles on a fine scale are very peculiar, but if we take large numbers of them and compound them, they approximate, but only approximately, Newton's laws. Newton's laws then permit us to go on to a higher and higher scale, and it still seems to be the same law. In fact, it becomes more and more accurate as the scale gets larger and larger. This self-reproducing factor of Newton's laws is thus really not a fundamental feature of nature, but is an important historical feature. We would never discover the fundamental laws of the atomic particles or first observation because the first observations are much too crude. In fact, in this

out that the fundamental atomic laws, which we call quantum mechanics, are quite different from Newton's laws, and are difficult to understand because all our direct experiences are with large-scale objects and the small-scale atoms behave like nothing we see once is so scale. So we cannot say, "An atom is just like a planet going around the sun," or anything like that. It is like nothing we are familiar with because there is nothing like it. As we apply quantum mechanics to larger and larger things, the laws about the behavior of atoms together do not reproduce themselves, but produce new laws, which are Newton's laws, which then continue to reproduce themselves from, say, micro-micromgram size, which still is billions and billions of atoms, on up to the size of the earth, and above.

Let us now return to the center of mass. The center of mass is sometimes called the center of gravity, for the reason that, in many cases, gravity may be considered uniform. Let us suppose that we have small enough dimensions that the gravitational force is not only proportional to the mass, but is everywhere parallel to some fixed line. Then consider an object in which there are gravitational forces on each of its constituent masses. Let m be the mass of one part. Then the gravitational force on that part is $m g$. Now the question is, where can we apply a single force to balance the gravitational force on the whole thing, so that the entire object, if it is a rigid body, will not turn? The answer is that the force must go through the center of mass, and we show this in the following way. In order that the body will not turn, the torque produced by all the forces must add up to zero, because if there is a torque, there is a change of angular momentum, and thus a rotation. So we want calculate the total of all the torques on all the particles, and see how much torque law is about any place and it should be zero if this axis is at the center of mass. Now, examining τ horizontally and vertically, we know that the torques are the forces in the x -direction times the lever arm x (that is to say, the lever times the lever arm around which we want to calculate the torque). Now the total torque is the sum

$$\tau = \sum m_i g x_i = g \sum m_i x_i. \quad (19.5)$$

so if the total torque is to be zero, the sum $\sum m_i x_i$ must be zero. But $\sum m_i x_i = Mx$, the total mass times the distance of the center of mass from the axis. Thus the x distance of the center of mass from the axis is zero.

Of course, we have checked the result only for the x -distance, but if we use the true center of mass the object will balance in any position, because if we turned it 90 degrees, we would have y 's instead of x 's. In other words, when an object is supported at its center of mass, there is no torque on it because of a parallel gravitational field. In case the object is so large that the nonparallelism of the gravitational forces is significant, then the center where one must apply the balancing force is not simple to describe, and is deflected slightly from the center of mass. That is why one must distinguish between the center of mass and the center of gravity. The fact that an object supported entirely at the center of mass will balance in all positions has another interesting consequence. If, instead of gravitation, we have a pseudoforce due to acceleration, we may use exactly the same mathematical procedure to find the position to support it so that there are no torques produced by the inertial force of acceleration. Suppose that the object is held in some instant inside a box, and that the box, and everything contained in it, is accelerating. We know that, from the point of view of someone at rest relative to this accelerating box, there will be an effective force due to inertia. That is, to make the object go along with the box, we have to push on it to accelerate it, and this force is "balanced" by the "force of inertia," which is a pseudoforce equal to the mass times the acceleration of the box. To the man in the box, this is the same situation as if the object were in a uniform gravitational field whose " x " value is equal to the acceleration a . Thus the inertial force due to accelerating an object has no torque about the center of mass.

This fact has a very interesting consequence. In an inertial frame that is not accelerating, the torque is always equal to the rate of change of the angular momentum. However, about an axis through the center of mass of an object which

is accelerating, it is still true that the torque is equal to the rate of change of the angular momentum. Even if the center of mass is accelerating, we may still choose one special axis, namely, one passing through the center of mass, such that it will still be true that the torque is equal to the rate of change of angular momentum around that axis. Thus the theorem that torque equals the rate of change of angular momentum is true in two general cases: (1) in fixed axis in inertial space; (2) an axis through the center of mass, even though the object may be accelerating.

19-2 Locating the center of mass

The mathematical techniques for the calculation of centers of mass are in the province of a mathematical tourist, and such problems provide good exercise in integral calculus. After you have learned calculus, however, and want to know how to locate centers of mass, it is nice to know certain tricks which can be used to do so. One such trick makes use of what is called the theorem of Pappus. It works like this: if we take any closed area in a plane and generate a solid by moving it through space such that each point is always moved perpendicular to the plane of the area, the resulting solid has a total volume equal to the area of the cross section times the distance that the center of mass moved. Certainly this is true if we move the area in a straight line perpendicular to itself; but if we move it in a circle or in some other curve, then it generates a rather peculiar volume. For a curved path, the outside goes around further, and the inside goes around less, and these effects balance out. So if we want to locate the center of mass of a plane sheet of uniform density, we can remember that the volume generated by spinning it about an axis is the distance that the center of mass goes around, times the area of the sheet.

For example, if we wish to find the center of mass of a right triangle of base B and height H (Fig. 19-2), we might solve the problem in the following way. Imagine an axis along H , and rotate the triangle about that axis through $2\pi/3$ degrees. This generates a cone. The distance that the center of mass has moved is $\pi H/3$. The area which is being moved is the area of the triangle, $\frac{1}{2}BH$. So the distance of the center of mass times the area of the triangle is the volume swept out, which is of course $\pi B^2 H/2$. Thus $(\text{area})(\frac{1}{2}\pi H)$ = $\pi B^2 H/2$, or $\pi = B/3$. In a similar manner, rotating about the other axis, or by symmetry, we find $\rho = H/3$. In fact, the center of mass of any uniform triangular area is where the three medians, the lines from the vertices through the centers of the opposite sides, all meet. That point is $1/3$ of the way along each median. Check! Since the triangle splits into 6 pieces, each parallel to a base. Note that the median line bisects every piece, and therefore the center of mass lies on this line.

Now let us try a more complicated figure. Suppose that it is desired to find the position of the center of mass of a uniform semicircular disc—a disc sliced in half. Where is the center of mass? For a full disc, it is at the center, of course, but a half-disc is more difficult. Let r be the radius and x be the distance of the center of mass from the straight edge of the disc. Spin it around this edge as axis to generate a sphere. Then the center of mass has gone around 2π , the area is $\pi r^2/2$ (because it is only half a circle). The volume generated is, of course, $4\pi r^3/3$, from which we find that

$$(2\pi r)(\frac{1}{2}\pi r^2) = 4\pi r^3/3,$$

or

$$r = 4r/3.$$

There is another theorem of Pappus, which is a special case of the above one, and therefore equally true. Suppose that instead of the solid semicircular disc, we have a semicircular piece of wire with uniform mass density along the wire, and we want to find its center of mass. In this case there is no cone in the interior, only on the wire. Then it turns out that the area which is swept by a plane curved line, when it moves as before, is the distance that the center of mass moves times the length of the line. (The line can be thought of as a very narrow area, and the previous theorem can be applied to it.)

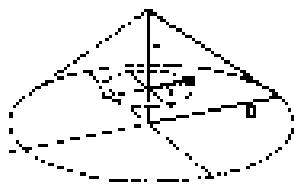


Fig. 19-2. A right triangle and a right circular cone generated by rotating the triangle.

19.3 Finding the moment of inertia

Now let us discuss the problem of finding the moments of inertia of various objects. The formula for the moment of inertia about the center of an object is

$$I = \sum m_i(x_i^2 + y_i^2)$$

or

$$I = \int (x^2 + y^2) dm = \int (x^2 + y^2) \rho dV. \quad (19.4)$$

That is, we must sum the masses, each one multiplied by the square of its distance ($x_i^2 + y_i^2$) from the axis. Note that it is not the three-dimensional distance, only the two-dimensional distance squared, even for a three-dimensional object. For the present part, we shall restrict ourselves to two-dimensional objects, but the formula for rotation about the z -axis is just the same in three dimensions.

As a simple example, consider a rod rotating about a perpendicular axis through one end (Fig. 19-3). Now we must sum all the masses times the x distances squared (the y 's being all zero in this case). What we mean by "the sum" of course, is the integral of x^2 times the little elements of mass. If we divide the rod into small elements of length dx , the corresponding elements of mass are proportional to dx , and if L were the length of the whole rod the mass would be M . Therefore

$$dm = M dx/L$$

and so

$$I = \int_0^L x^2 \frac{M dx}{L} = \frac{M}{L} \int_0^L x^2 dx = \frac{ML^2}{3}. \quad (19.5)$$

The dimensions of moment of inertia are always mass times length squared, so all we really had to work out was the factor $1/3$.

Now what is I if the rotation axis is at the center of the rod? We could just do the integral over again, letting x range from $-\frac{L}{2}$ to $\frac{L}{2}$. But let us notice a few things about the moment of inertia. We can imagine the rod as two rods, each of mass $M/2$ and length $L/2$; the moments of inertia of the two small rods are equal, and are both given by the formula (19.5). Therefore the moment of inertia is

$$I = \frac{2(M/2)(L/2)^2}{3} = \frac{ML^2}{12}. \quad (19.6)$$

Thus it is much easier to turn a rod about its center than to swing it about an end.

Of course, we could go on to compute the moments of inertia of various other bodies of interest. However, while such computations provide a valuable source of important exercise in the calculus, they are not basically of interest to us as such. There is, however, an interesting theorem which is very useful. Suppose we have an object, and we want to find its moment of inertia around some axis. This means we want the kinetic needed to carry it by rotation about that axis. Now if we support the object or pivot at the center of mass, so that the object does not turn as it rotates about the axis (because there is no torque on it, linear inertia effects, and therefore it will not turn when we start moving it), then the forces needed to swing it around act the same as though all the mass were concentrated at the center of mass, and the moment of inertia would be simply $I = MR_{CM}^2$, where R_{CM} is the distance from the axis to the center of mass. But of course that is not the right formula for the moment of inertia of an object, which is really being rotated as it revolves, because not only is the center of it moving in a circle, which would add twice an amount I_C to the moment of inertia, but also we must find I about its center of mass. So it is not unreasonable that we must add to I_C the moment of inertia I_C about the center of mass. So it is a good guess that the total moment of inertia about any axis will be

$$I = I_C + MR_{CM}^2 \quad (19.7)$$

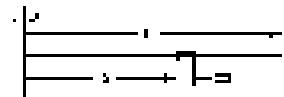


Fig. 19-3. A straight rod of length L rotating about a z -axis through one end.

This theorem is called the *parallel-axis theorem*, and may be easily proved. The moment of inertia about any axis is the same times the sum of the x_i^2 's and the y_i^2 's, each squared: $I = \sum (x_i^2 + y_i^2)m_i$. We shall concentrate on the x 's, but of course the y 's work the same way. Now x' is the distance of a particular point from the origin, but let us consider how it would look if we measured x' from the CM, instead of x from the origin. To get ready for this analysis, we write

$$x_i = x'_i + X_{CM}.$$

Then we just square this to find

$$x_i^2 = x'^2 + 2x'_iX_{CM} + X_{CM}^2.$$

So, when this is multiplied by m_i , and summed over all i , what happens? Taking the constants outside the summation sign, we get

$$I_x = \sum m_i x_i^2 + 2X_{CM} \sum m_i x'_i + X_{CM}^2 \sum m_i.$$

The third sum is easy; it's just MX_{CM}^2 . In the second sum there are two pieces, one of them is $\sum m_i x'_i$, which is the total mass times the displacement of the center of mass. But this certainly has nothing to do with the center of mass, and in these axes, the average position of all the particles, weighted by the masses, is zero. The first sum, of course, is the x part of I . Thus we arrive at Eq. (19.7), just as we guessed.

Let us check (19.7) for one example. Let us just see whether it works for the rod. For an axis through one end, the moment of inertia should be $ML^2/3$, for we calculated that. The center of mass of a rod, of course, is at the center of the rod, at a distance $L/2$. Therefore we should find that $ML^2/3 = ML^2/12 + M(L/2)^2$. Since one-quarter plus one-twelfth is one-third, we have made no fundamental error.

Incidentally, we did not really need to use an integral to find the moment of inertia (19.5). If we simply assume that it is ML^2 times γ , an unknown coefficient, and then use the argument about the two halves to go from (19.6), then from our argument about transferring the axis we could prove that $\gamma = \frac{1}{3}$. So γ must be $1/3$. There is always another way to do it!

In applying the parallel axis theorem, it is of course important to remember that the axis for I , must be parallel to the axis about which the moment of inertia is wanted.

One further property of the moment of inertia is worth mentioning because it is often helpful in finding the moment of inertia of certain kinds of objects. This property is that if one has a plane figure and a set of coordinate axes with origin in the plane and z -axis perpendicular to the plane, then the moment of inertia of this figure about the x -axis is equal to the sum of the moments of inertia about the x - and y -axes. This is easily proved by noting that

$$I_x = \sum m_i(x_i^2 + z_i^2) = \sum m_i x_i^2$$

(since $z_i = 0$). Similarly,

$$I_y = \sum m_i(x_i^2 + z_i^2) = \sum m_i y_i^2$$

but

$$\begin{aligned} I_z &= \sum m_i(x_i^2 + y_i^2) = \sum m_i x_i^2 + \sum m_i y_i^2 \\ &= I_x + I_y. \end{aligned}$$

As an example, the moment of inertia of a uniform rectangular plate of mass M , width w , and length l , about an axis perpendicular to the plate and through its center is simply

$$I = Mlw^2 + l^2w^2/12,$$

because the moment of inertia about an axis in its plane and parallel to its length is $Mw^2/12$, etc., just as for a rod of length w , and the moment of inertia about the other axis in its plane is $Ml^2/12$, just as for a rod of length l .

To summarize, the moment of inertia of an object about a given axis, which we shall call the *axis of rotation*, has the following properties:

- (a) The moment of inertia is

$$I_z = \sum m_i(r_i^2 + y_i^2) = \int (r^2 + y^2) dm.$$

- (b) If the object is made of a number of pieces, each of whose moment of inertia is known, the total moment of inertia is the sum of the moments of inertia of the pieces.
- (c) The moment of inertia about any given axis is equal to the moment of inertia about a parallel axis through the CM plus the total mass times the square of the distance from the axis to the CM.
- (d) If the object is a plane figure, its moment of inertia about any perpendicular to the plane is equal to the sum of the moments of inertia about any two mutually perpendicular axes lying in the plane and intersecting at the perpendicular axis.

The moments of inertia of a number of elements of shapes having uniform mass densities are given in Table 19-1, and the moments of inertia of some other objects, which may be deduced from Table 19-1, using the above properties, are given in Table 19-2.

Table 19-1

Object	axis	I_z
Thin rod, length L	end or center	$ML^2/12$
Thin, concentric circular ring, radii r_1 and r_2	loop at center	$M(r_1^2 + r_2^2)/2$
Sphere, radius r	through center	$2Mr^2/5$

Table 19-2

Object	axis	I_z
Rect. sheet, sides a, b	both ends	$Ma^2/12$
Rect. sheet, sides a, b	center	$M(a^2 + b^2)/12$
Thin annulus, radii r_1, r_2	any diameter	$M(r_1^2 + r_2^2)/4$
Rect. parallelepiped, sides a, b, c	c , through center	$M(a^2 + b^2)/12$
Right cyl., radius r , length L	$\perp L$, through center	$Mr^2/2$
Right cyl., radius r , length L	$\perp L$, through center	$Mr^2/4 + L^2/12$

19-4 Rotational kinetic energy

Now let us go on to discuss dynamics further. In the analogy between linear motion and rotation just on that we discussed in Chapter 18, we used the work theorem, but we did not talk about kinetic energy. What is the kinetic energy of a rigid body, rotating about a certain axis with an angular velocity ω ? We can immediately guess the correct answer by using our analogies. The moment of inertia corresponds to the mass, angular velocity corresponds to velocity, and so the kinetic energy ought to be $(1/2)\omega^2$, and indeed it is, as will now be demonstrated. Suppose the object is rotating about some axis so that each point has a velocity whose magnitude is ωr , where r is the radius from the particular point to the axis.

Then if ω is the ω of that point, the total kinetic energy of the wheel is just the sum of the kinetic energies of all of the little pieces.

$$T = \frac{1}{2} \sum m_i v_i^2 = \frac{1}{2} \sum m_i (\omega r_i)^2$$

Now v^2 is a constant, the same for all points. This

$$T = \frac{1}{2} \omega^2 \sum m_i r_i^2 = \frac{1}{2} I \omega^2. \quad (19.8)$$

At the end of Chapter 28 we put that out, but there are some interesting phenomena associated with an object which is not rigid, but which changes from one rigid condition with a definite moment of inertia to another one condition. Namely, in our example of the turntable, we had a certain moment of inertia I_{turn} with our arms stretched out, and a certain angular velocity ω_0 . When we pulled our arms in, we had a different moment of inertia I_f , and a different angular velocity ω_f . In again we were "right." The angular momentum, $I \omega$, remained constant, since there was no torque due to external forces of the turntable. This means that $I \omega_0 = I_f \omega_f$. Now what about the energy? That is an interesting question. With our arms pulled in, we turn faster, but our moment of inertia is less, and so obviously there's more energy right? But they aren't because what does decrease is $I \omega$, not ω . So the answer is, the kinetic energy before and after the kinetic energy before is $\frac{1}{2} I \omega_0^2 - \frac{1}{2} I_f \omega_f^2$, where $I = I_{\text{turn}} = I_f + I$ the angular momentum. Afterward, by the same argument, we have $\frac{1}{2} I_f \omega_f^2$, and since $\omega_f > \omega_0$, the kinetic energy of rotation is greater than it was before. So we had a certain energy when our arms were out, and when we pulled them in, we were turning faster and had more kinetic energy. What happened to the theorem of conservation law of energy? So clearly you never do any work. We did work when did we do any work? When we move a weight horizontally, we do not do any work. If we have a thing and pull it in, we do not do any work. But this is when we are not rotating! When we are rotating, there is a centrifugal force on the weights. They are trying to fly out so when we are going around we have to pull the weights in against the centrifugal force, i.e., the work we do against the centrifugal force F_c , is equal with the difference in rotational energy, and of course it does. This is where the conservation of energy comes from.

This is still another interesting fact, which we can treat only intuitively, as a matter of general knowledge. This figure is a little more abstract, but it works pointing out because it is quite curious and produces many interesting effects.

Consider first the experimental setup. Consider the body and the ring separately from the point of view of the man who is turning. After the weights are pulled in, the wheel object is spinning faster, but observe, the overall part of the body is not changing, yet it is turning faster than the ring. So, if we were to draw a circle around the inner body, and consider only ω_0 to make the circle, their angular momentum would change, they are going faster. Therefore there must be a torque exerted on the body while we spin our arms. No torque can be exerted by the centrifugal force, because that is radial. So that means that among the forces that are developed in a rotating system, centrifugal force is the business story, there is another force. The other force is called *Gyroscopic force*, and it has the very strong property that when we move something in a rotating system, it seems to be pushed sideways. Like the centrifugal force, it is an apparent force. It is the law of a reason. It is rotating, and moves something radially out, and outwards, and also push it sideways to move it radially. This is why just what we have to learn is what turned the body around.

Now let us develop a formula to show how the Coriolis force really works. Suppose Max is sitting on a carousel that is going to be stationary. But from the point of view of Joe, who is standing on the ground and who knows the right laws of mechanics, the is moving it going around. Suppose that we have drawn a radial line on the carousel, and that Max is moving some mass steadily along this line. We would like to determine what is the greatest force he is required to do this. We can do this by paying attention to the angular momentum of the mass. It is

always going around with the same tangential velocity v , so that the angular momentum is

$$L = m v_{\text{tang}} r = m v r = m v^2 r$$

so when the mass is close to the center, it has relatively little angular momentum, but if we move it to a new position farther out, if we increase r , it has more angular momentum, so a torque must be exerted in order to move it along the radius (to pull along the radius) in accordance with the conservation of angular momentum along the radius. If an object only moves on a radius, no torque is constant, so that the torque is

$$\tau = I \alpha = \frac{dI}{dt} = \frac{J_{\text{orbital}}^2}{r^3} = J_{\text{orbital}} \frac{dr}{dt},$$

where J_{orb} is the Orbital Momentum. We actually want to know is what sideways force has to be exerted by Moon in order to move its orbital speed $v = dr/dt$. This is $F_c = \tau/r = 2m\omega^2$.

Now let us have a friend for the Coriolis force let instant at the θ angle. Let's move carefully, to see whether we are understanding the origin of this force from a more elementary point of view. We note that the Coriolis force is the same in every radius, and is evidently present even at the equator. But it is especially easy to understand it at the origin, just by looking at what happens from the orbital equation of the Moon, who is standing on the ground. Figure 19-4 shows three successive views of just as it passes the origin at $t = 0$. Because of the rotation of the earth, we see that it does not move in a straight line, but in a curved path tangent to a radius of the earth at $r = 0$. In order for it to go in a curve, there must be a force to accelerate it in physical space. It is the Coriolis force.

This is not the only case in which the Coriolis force occurs. We can also show this. If an object is moving with constant speed v , and the circumference of a circle, there is also a Coriolis force. Why? Moon has a velocity say around the circle. On the other hand, he sees an object around the circle with the velocity $v = v_0 - \omega r$, because it is also carried by the rotation. Therefore we know who the force really is, namely, the total centripetal force due to the velocity v_0 , or v_0^2/r ; that is the centripetal force. Now from Moon's point of view, the centripetal force has three pieces. We may write it all out as follows:

$$F_c = \frac{v_0^2}{r} = \frac{m v_0^2}{r} = 2m\omega v_0 + m\omega^2 r.$$

Now, $m\omega$ is the force that Moon weight has. Let me try to understand it. Would she appreciate the first term? "Yes" he would say, "but if I were not rotating, there would be no centrifugal force; if I were to go around a circle with velocity v_0 ." This is simply the centripetal force that Moon would expect, having nothing to do with rotation. In addition, Moon's would notice that there is another centripetal force; but would he see on objects which are standing v_0 ? on his is moved. This is the centrifugal force. There is another term in addition to these, namely the second term which is again $2m\omega v_0$. The Coriolis force F_c is tangential when the velocity v_0 exists, and note it is radial when the velocity is tangential. i.e. v_0 , one expresses v_0 as a radius $r\omega$ relative to the other. The force is always at the same direction, relative to the velocity, no matter in which direction the velocity v_0 . The force is at right angles to the velocity, and of magnitude $2m\omega v_0$.

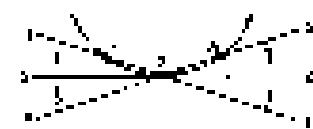


Fig. 19-4. Three successive views of a point moving rectely on a rotating turntable.

Newtonian mechanics

20-1 Torque in three dimensions

In this exercise, we shall discuss one of the most remarkable and amazing consequences of mechanics: the behavior of a rotating wheel. In order to do this we must first extend the mathematical framework of rotational motion. The principles of angular momentum, long known so far to three-dimensional space. We will see how they reduce in all cases to a very simple form. All the "cross" terms, because the word take away from us, and we must now turn to ω in subscripts. In an absolute sense we can proceed only by the fundamental laws and apply them in a very few situations of special interest.

First, we notice that if we have rotation in three dimensions, whether of a rigid body or a rigid system, what we deduced for two dimensions is still right. That is, it is still true that $\Delta E_k = \Delta E_\tau$ is the torque "in the xy-plane," or the torque "around the z-axis." That is, the net torque is still equal to the sum of the two " $\tau_{xy} = p_{xy}$, for if we go back over the derivation of Eq. (14.1a) from Newton's laws we see that we did not have to assume that the rotation was in a plane. When we differentiate $p_{xy} = p_{xy}$, we get $\Delta E_k = \Delta E_\tau$, so this theorem is still right. The quantity $p_{xy} = p_{xy}$ then, is still the angular momentum belonging to the xy-plane, or the angular momentum about the z-axis. It is being true, we can use the other part of this and get another equation. For instance, we can use the zy-plane, and it is clear from symmetry that, for just this理由 the x and z for τ we would find $\Delta E_z = \Delta E_\tau$ for the torque and $p_{zy} = p_{zy}$, would be the angular momentum associated with the zy-plane. Of course we could have another plane, the xz-plane, and for this we would find $\Delta E_x = \Delta E_\tau$, $\delta E_x^2(p_{xz}) = \Delta E_\tau$.

That these three equations can be deduced for the moment of a triple you will is quite clear. Furthermore, if we added such things as $p_{yz} = p_{yz}$, ΔE_y^2 for many variables and called it the total angular momentum we would have three terms for the three planes p_{xy} , p_{yz} and p_z , and if we did the same with the forces, we would obtain the torque in the xy, yz, xz, and z axes. Thus we would have shown that the external torque associated with any plane is equal to the rate of change of the angular momentum associated with that plane. This is just a generalization of what we wrote in two dimensions.

Let now take one step, "Ah, but there are more planes; after all, can we not take some other fixed angle, and calculate the torque in that plane from the forces?" Since we would have to write a rather lot of equations for every such plane, we would have a lot of "useless" terminology enough. It turns out that it is easier to work in the combination $\Delta E_\tau = \Delta E_\tau$. We can do this, measuring only ΔE_τ , etc., in each plane. It is useful, in a certain sense, some combination of the three existing ones for the xy, yz, xz and z-planes. There is nothing new. To this point, if we know what the forces are in the xz, xy, etc., and z-planes are, then the torque in any other plane, and vice versa, the angular momentum can also be written as a combination of these six percent of six and thirty-two percent of another, and so on. This property is called *closure* to you.

Suppose that in the system we find forces and all the torques due to angular momentum in the planes. But also ΔE_x^2 , ΔE_y^2 , ΔE_z^2 in some other dimension. To make it a little easier, we shall suppose that only the τ -terms have been listed. Most of ΔE_x^2 , etc., are zero, but this is supposed to be the case. That is, he has no planes to worry about and ΔE . He therefore has six ΔE 's and angular momentum which he would work out. For example his term in the xy-plane would be equal to $\Delta E_\tau = p_{xy}$, and ΔE for b. What we must do is to find all the relationship between the new angles and the old angles, as we will be able to make a

20-1 Torque in three dimensions

20-2 The relation between the cross products

20-3 The gyroscope

20-4 Angular momentum of a solid body

mention from Chapter 17 says "the cross product is just like vector multiplication with vectors." And indeed, that is exactly what we are looking for here. Then he may say, "Well, isn't a vector just a scalar?" A scalar cannot fail to be a vector, but we do not know that right away without making a analysis. So if the following steps we take make no analysis, We can't" because it's just a step by step, since we only want nothing but how it works. The two equations that follow are

$$\begin{aligned} \mathbf{r}_{xy} &= \mathbf{x}\mathbf{r}_y - \mathbf{y}\mathbf{r}_x \\ \mathbf{r}_{xz} &= \mathbf{x}\mathbf{r}_z - \mathbf{z}\mathbf{r}_x \\ \mathbf{r}_{yz} &= \mathbf{y}\mathbf{r}_z - \mathbf{z}\mathbf{r}_y \end{aligned} \quad (20.1)$$

We discuss the first one and it is enough to discuss among the three since the same quantity of the components is not needed in the right way. We're not calculating \mathbf{r}_{xy} , \mathbf{r}_{xz} , \mathbf{r}_{yz} . The problem depends on the fact that x can have either two values "right-handed" or "left-handed". To say clearly which hand it is right for, we say "left" the second reference to the other two quantities may cause a trouble introducing the later stage in our analysis.



This now calculates the reference in 3D system:

$$\begin{aligned} \mathbf{r}_{xy} &= \mathbf{x}'\mathbf{r}_y - \mathbf{y}'\mathbf{r}_x \\ \mathbf{r}_{xz} &= \mathbf{x}'\mathbf{r}_z - \mathbf{z}'\mathbf{r}_x \\ \mathbf{r}_{yz} &= \mathbf{y}'\mathbf{r}_z - \mathbf{z}'\mathbf{r}_y \end{aligned} \quad (20.2)$$

Now we suppose that our coordinate system is rotated by a fixed angle θ , such that the local axes are the same. (Triangle 2 has nothing to do with rotating objects or what is going on, but in the coordinate system, it is there.) The relationship between the axes and by our memory, the axes used by one other, and is supposedly constant. Thus, the coordinates of the two systems are related by

$$\begin{aligned} x' &= x \cos \theta + y \sin \theta, \\ z' &= -y \cos \theta + x \sin \theta \\ y' &= y \end{aligned} \quad (20.3)$$

Therefore, we see from y' is a vector it transforms into the new system in the same way as do x' and z' . Just as being a vector if and only if its various components is transformed linearly by x , y , and z .

$$\begin{aligned} \mathbf{r}_{xy} &= \mathbf{r}_x \cos \theta + \mathbf{r}_y \sin \theta, \\ \mathbf{r}_{xz} &= \mathbf{r}_x \sin \theta - \mathbf{r}_y \cos \theta \\ \mathbf{r}_{yz} &= \mathbf{r}_y \end{aligned} \quad (20.4)$$

Now we can find out how the x -axis transforms by merely substituting for x , y , and z the x -vectors (20.3) into (20.1). For \mathbf{r}_{xy} , \mathbf{r}_{xz} , \mathbf{r}_{yz} these given by (20.4) all make (20.4). So, we have a rather long string of terms for \mathbf{r}_{xy} , and (rather surprisingly) an additional term than it comes up in \mathbf{r}_{xz} and \mathbf{r}_{yz} , which we need to be the component of the x -plane:

$$\begin{aligned} \mathbf{r}_{xy} &= (\mathbf{x} \cos \theta + \mathbf{y} \sin \theta)(\mathbf{x} \cos \theta + \mathbf{y} \sin \theta) \\ &\quad - (\mathbf{x} \cos \theta + \mathbf{y} \sin \theta)(\mathbf{z} \cos \theta - \mathbf{z} \sin \theta) \\ &\quad + \mathbf{x}^2(\cos^2 \theta + \sin^2 \theta) - \mathbf{y}^2(\sin^2 \theta + \cos^2 \theta) \\ &\quad - \mathbf{z}(\mathbf{x} \cos \theta + \mathbf{y} \sin \theta) + \mathbf{z}^2(\cos^2 \theta) \\ &\quad + \mathbf{z}^2(\sin^2 \theta) + \mathbf{z} = \mathbf{z}(\cos^2 \theta) \\ &\Rightarrow \mathbf{r}_{xy} = \mathbf{y} \mathbf{r}_x + \mathbf{x} \mathbf{r}_y \end{aligned} \quad (20.5)$$

The result is clear, but if we only took the axes in the plane, the twist around a fixed point is insufficient; then it goes back to hexagonal in the same sense. What will be more interesting is the expression for τ_{xy} , because that is a new plane. We know the answer, the same thing with the vectorial case, but it is not so simple:

$$\begin{aligned}\tau_{xy} &= \tau_x \cos \theta - \tau_z \sin \theta \\&= \tau_x^2 \cos^2 \theta - \tau_z^2 \sin^2 \theta \\&= (\tau_x^2 - \tau_z^2) \cos^2 \theta + \tau_x \tau_z \cos \theta \\&= \tau_x \cos \theta + \tau_z \sin \theta\end{aligned}\quad (20.6)$$

Finally, we consider τ_{xz} :

$$\begin{aligned}\tau_{xz} &= 2(\tau_x \cos \theta - \tau_y \sin \theta) \\&= -(\tau_x + \tau_y + \tau_z) \sin \theta \\&= (\tau_x - \tau_z) \cos \theta - (\tau_x^2 - \tau_z^2) \sin \theta \\&= \tau_x \cos \theta - \tau_z \sin \theta\end{aligned}\quad (20.7)$$

We wanted to get a rule for finding torques in these cases in terms of torques in old sense, and now we have the rule. This can also be understood that rule. If we look carefully at (20.6), (20.7), we see that there's a close relationship between these equations and the equations (20.4), (20.5), and (20.6). Sometimes, we could call τ_x the component of summing τ_1 + τ_2 + τ_3 , component of τ_1 , then it would be all right, we would understand (20.5) is a vector transformation, once the assignment would be unchanged, as it should be. Otherwise, if we associate that the yz -plane the x -component of our study invariant vector, and with the exception, the yz -component, that these transformable expressions would not.

$$\begin{aligned}\tau_x &= \tau_1 \\&= \tau_x \cos \theta + \tau_z \sin \theta \\&= \tau_y - \tau_z \cos \theta - \tau_x \sin \theta\end{aligned}\quad (20.8)$$

where θ is just the angle for vector τ_1 .

There is a we have proved that we can identify the combination of $\tau_1 + \tau_2 + \tau_3$, with what we ordinarily call the x -component of a vector, as it is really a real vector. Although a torque is a twist of a surface, and it has no physical vector character, and most really it does not have linear character. This vector is along it angles to the axis of the twist, and its length is proportional to the strength of the twist. The dimension, probably of such a quantity like τ_1 , transforms like a real vector.

So we represent a torque by a vector; with each plane π , what the torque is supposed to be acting, we represent a point at right angles to a plane. But the "right angles" carries the sign uncancelled. To get the x -axis right, we must adopt a rule which will tell us that the torque were in clockwise sense on the z -axis. This was not required to associate with τ_1 in the "right" orientation. That is, somebody has to define "right" and "left" for us. Supposing that the coordinate system is x, y, z in a right-hand system, then the rule will be the following: if we think of the twist as if we were turning a screw having a right-hand thread, then the direction of the twist is the "new" coordinate, with that x is 1. Consideration that the screw would advance.

What is torque in reality? It is a miracle of good luck that we can associate a single axis with a plane, and therefore that we can associate a vector with the torque; it is a special property of three-dimensional space. In two dimensions, the torque is an ordinary vector, and there need be no relation associated with it. In less dimensions, it is a vector. If we had four dimensions we would be in great difficulty, because if we are using, for example, at the fourth dimension we would not only xy -plane like x, y, z, u , we would have to x, y, z, w components. There would be x, y, z, w there, and one cannot represent x, y, z, w as a vector in four dimensions.

We will be thinking in three dimensions for a long time, so it is well to realize that the triggering moment and the reaction did not depend upon the fact that a

has position and force tensor, it only depends on the transformation laws for vectors. Therefore if instead of a we had the x -component of some other vector, it is not going to make any difference. In other words, if we were to calculate $a_1 b_1 + a_2 b_2$, where a and b are vectors and call it the x -component of some new quantity c , then these two quantities form a vector c . We need a mathematical notation for the relationship of the new vector with its three components a_1, a_2 to the vectors a and b . The notation has been devised to this is $c = a \times b$. We have seen, in addition to the ordinary scalar product in the theory of vector spaces, a new kind of product called the cross product. Thus, if $c = a \times b$, this is the componentship

$$\begin{aligned} c_1 &= a_2 b_1 - a_1 b_2, \\ c_2 &= a_3 b_1 - a_1 b_3, \\ c_3 &= a_2 b_3 - a_3 b_2. \end{aligned} \quad (20.8)$$

If we reverse the order of a and b , calling a from b , a , we would have the components of c now become c_1 would be $b_2 a_1 - b_1 a_2$. The same commutative property is in like ordinary multiplication, where $ab = ba$; in the cross product, $b \times a = -a \times b$. From this, we can prove at once that if $b = 0$, the cross product is zero. Thus, $a \times a = 0$.

The cross product is very important for representing the features of rotation and it is important that we understand the geometrical realization of the three vectors a , b , and c . By means of relationships among planes is given in 17 (20.9) and 20m that one can determine what the relationship is in geometry. One knows by first that the vector c is perpendicular to both a and b . (The b vector sits in a and c if they not reduce to zero.) Second, the magnitude of c turns out to be the magnitude of a times the sine of the angle between the two. To which I continue with a point: Imagine a string from a into b through an angle less than 180° ; if you pull with a right-hand thread running in this way will produce in the direction of c . The fact that we say a right-hand screw instead of a left-hand screw is a convention, and is a geometrical consideration. If a and b are "frozen" vectors in the ordinary sense, the new kind of "vector" which we have created by $a \times b$ is not this, or slightly different to its structure from a vector, however it was made up with a special rule. If we had three ordinary vectors, we have a special name for them, which are their real vectors. Examples of such vectors are the gravitational vector \mathbf{g} , momentum \mathbf{p} , velocity \mathbf{v} , electric field \mathbf{E} , etc.; these are ordinary vector vectors. Vectors which involve just one cross product to their definition are called axial vectors or pseudovectors. Examples of pseudovectors are, of course, torque τ and the angular momentum L . It also turns out that the angular velocity ω is a pseudovector as is the magnetic field \mathbf{B} .

In order to complete the mathematical properties of vectors, we shall know all the rules for their multiplication, using the various products. In our applications to mechanics, we will use very little of this, but for the sake of completeness, we shall write down all of the rules for vector multiplication. Let us first list the results here. These are

$$\begin{aligned} (a) \quad a \times (b + c) &= a \times b + a \times c \\ (b) \quad (ab) \times b &= a(b \times b), \\ (c) \quad a \cdot (b \times c) &= (a \times b) \cdot c, \\ (d) \quad a \times (b \times c) &= (b \cdot a) c - (c \cdot a) b, \\ (e) \quad a \times a &= 0, \\ (f) \quad a \cdot (a \times b) &= 0. \end{aligned} \quad (20.10)$$

20-3 The rotation equations using cross products

Now let us ask whether any equations in physics can be solved using the cross product. The answer, of course, is that a great many equations can be written. For instance, we can immediately that the torque is equal to the position

$$\tau = \mathbf{r} \times \mathbf{F}. \quad (20.11)$$

This is a vector summary of the 1D...equations $\tau_x = r_x \dot{\theta}$, $= -r_y \dot{\theta}$, etc. By the same token, the angular momentum vector, if there is only one particle present in the system from the origin multiplied by the vector momentum.

$$\mathbf{L} = \mathbf{r} \times \mathbf{p}. \quad (20.12)$$

For three-dimensional system rotation, the dynamical law analogous to the law $\mathbf{R} = \mathbf{ap}/dt$ of Newton, is that the unique vector is the rate of change with time of the angular momentum vector:

$$\tau = d\mathbf{L}/dt. \quad (20.13)$$

Over $n=30$, 000 many particles, the average torque on a system is the rate of change of the total angular momentum:

$$\tau_{av} = d\mathbf{L}_{av}/dt. \quad (20.14)$$

Angular momentum: If the total angular momentum is zero, then the total vector angular momentum of the system is a constant. This is called the **axis of conservation of angular momentum**. If there is no torque on a given system, its angular momentum is conserved.

What about angular velocity? Is it a vector? We have already discussed turning a solid object about a fixed axis, but for a moment suppose that we are turning simultaneously the "fixed axis". It might be turning itself or it might be a box, where the box is rapidly about some other axis. The net result of such combined motions is that the object simply turns about some new axis. The **angular velocity** along this new axis is then it can be figured out this way. If the rate of turning in the xz plane is written as a vector in the xz -direction, whose length is equal to the rate of rotation in the plane, and if another vector is drawn in the yz -direction, say, where is the rate of rotating in the yz plane, then if we add these together as a vector, the magnitude of the resultant is how fast the object is turning, and the direction tells us in what plane by the rule of the parallelogram. That is to say, simply, angular velocity is a vector, whereas we drew the magnitudes of the rotations in the three planes to project them at right angles to those planes.*

An simple application of the preceding angular velocity vector, we may evaluate the power being expended by the turn of a large solid rigid body. The power, P , of course, is the rate of change of work with time; in three dimensions, the power turns out to be $P = \tau \cdot \omega$.

All the formulas that we learned for plane rotation can be generalized to 3D motion. For example, if a rigid body is turning about a certain axis with angular velocity ω , we might ask, "What is the velocity of a point of a certain *fixed* coordinate?" We can't leave it up to problem for the student to solve. Let the velocity of a particle in a rigid body be given by $\mathbf{v} = \omega \times \mathbf{r}$, where ω is the angular velocity and \mathbf{r} is the position. Also, in another example of fixed coordinates, we had a formula for Coriolis force, which can also be written using these products of \mathbf{T} , \mathbf{A} , and \mathbf{B} . That is, if a particle is moving with velocity \mathbf{v} in a coordinate system which is, in fact, rotating with angular velocity ω , and we want to find \mathbf{F} in terms of the rotating coordinate system, then we have to take the previously-formula.

20-3 The gyroscope

Let us now return to the law of conservation of angular momentum. This law may be demonstrated with a rapidly spinning wheel, or gyroscope, as follows (see Fig. 20-1). If we sit on a swivel chair and hold the spinning wheel with its axis horizontal, the wheel has an angular velocity ω . And we set it in this situation of rotating in our chair;

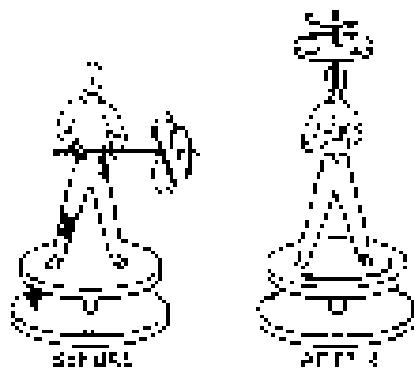


Fig. 20-1. Before and after swiveling, moment of inertia about vertical axis. (It's better to use a vertical axis, so it's easier to see what's happening.)

* That this is true can be shown by commanding the displacement of the particles of the body during an infinitesimal time Δt . If we call \mathbf{v}_i the velocity and \mathbf{r}_i the position of the i th particle, then $\mathbf{v}_i = \omega \times \mathbf{r}_i$.

Angular momentum around a vertical axis (center of the frictionless pivot of the chair) or if we turn the rest of the wheel into the vertical, the free-wheel would have angular momentum about the vertical axis, because it is now spinning about this axis. But the system (wheel, cartell, and chair) cannot have a vertical component, so we and the cartell have to turn in the direction opposite to the spin of the wheel, to balance it.

Just let us analyze it more fully the things we have just described. What is surprising and new we must understand, is the origin of the forces which can send the chair around as we are in the case of the gyroscope around the vertical. Figure 20-2 shows the wheel spinning steadily about the pivot. Therefore its angular velocity is about the axis and it is constant in angular momentum + linear motion in that direction. Now suppose that we wish to rotate the wheel about the x -axis at a small angular velocity α ; what does this do? After a short time Δt , the axis has turned to a new position, rotated an angle $\Delta\theta$ with the horizontal. Since the center part of the angular momentum is due to the spin of the wheel (that will be counteracted by the side friction), we see that the angular momentum vector has changed. What is the change in angular momentum? The x -component does not change in magnitude, but it does change its direction by an amount $\Delta\theta$. The magnitude of the vector $\Delta\theta$ is thus $\Delta\theta = \omega \Delta t$, so that the torque, which is the same rate of change of the angular momentum, is $\tau = 2I\omega\omega - I\omega^2 \Delta\theta = L\omega$. Taking the cosine law of the various vectors at the intersection, we see that

$$\tau = I\omega \times L. \quad (20.15)$$

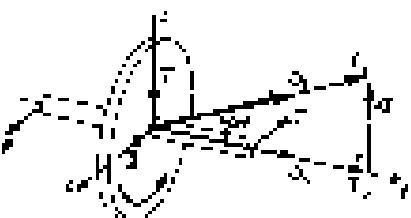


Fig. 20-2. A gyroscope.

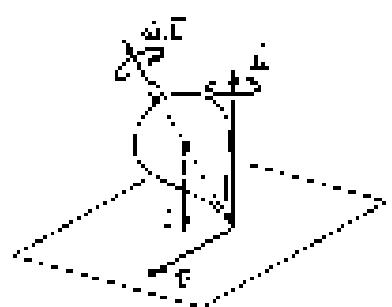


Fig. 20-3. A rapidly spinning top. Note that the direction of the torque vector is the direction of the precession.

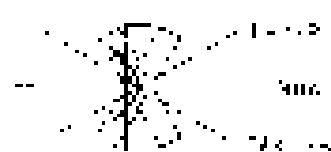


Fig. 20-4. The motion of particles in the spinning wheel of Fig. 20-2, where ω is zero, is in curved lines.

Thus, if ω and L_x are both horizontal, as shown in the figure, τ is zeroed. To produce such a torque, momenta forces F and $-F$ must be applied to the ends of the rod. How are these forces supplied? By our hands, as we turn the ends of the axis of the top off in the vertical direction. But Newton's Third Law demands that equal and opposite forces (and equal and opposite momenta) act on us. This causes us to rotate in the opposite sense around the vertical axis.

This result can be generalized to a rapidly spinning top. In the similar case of a spinning top, gravity acts on its center of mass (unless it is suspended) and points of contact with the floor (see Fig. 20-4). This torque is in the horizontal direction, and causes the top to precess with the direction of a circular path around the vertical. If Ω is the (vertical) angular velocity of rotation, we again find that

$$\tau = I\Omega\dot{\omega} = \Omega \times L.$$

Thus, when we apply a torque to a rapidly spinning top, the direction of the precessional motion is in the direction of the torque, τ , at right angles to the forces producing the torque.

We may now claim to understand the dynamics of gyroscopes, and indeed we do, rather nicely. However this is a rather "direct" way which, in a sense, appears sort of "intuitive." In real truth, we can go to more and more advanced physics, that usually comes later, can be exerted mathematically more easily than they can be really understood in a didactical or simple sense. This is a stumbling block, and as we get into more and more advanced work the more terminology will which usually will produce results we take as new but really can't be understood in any direct fashion. An example is the gyroscope, which appears in a very simple and beautiful form, but whose actual analysis is hard to understand. In our particular case, the precession is in big loops like those kind of orbits involving right angles and circles and twigs and right angles. What we should try to do is to understand it in a more physical way.

How can we explain this precession in terms of the real forces and accelerations? We note that when the wheel is precessing, the particles that are going around the wheel are not really moving in a plane because the wheel is precessing (see Fig. 20-4). As we explained previously (Fig. 19-4), the particles which are crossing through the precession axis are having a centripetal, and this requires application of a force. That is supplied by our pushing on the axis which can come up to

caused by the forces it has through its spokes. "Wait," you may say, "what about the forces that are going outside on the outer side?" It does not take long to realize that the τ must be a force in the opposite direction on that side. The net force that we have to apply is then $\tau + mg$. The forces balance out, but one of them must be applied at one side of the wheel, and the other must be applied at the other side of the wheel. We could apply τ to the left directly, but because the wheel is solid we can't just do it by pushing on the side, since forces can be applied only through the spokes.

What we learn is that if the wheel is precessing, it can balance the torque due to gravity or some other applied torque. But what we have shown is that τ is a function of an equation. That is, if the torque is given, and τ is due to gravity started right, then the wheel will precess sinusoidally and uniformly. But we have not stated that this is not true; that a uniform precession is the most general motion a spinning body can make as the result of a pure torque. The general motion involves then a "wobbling" about the mean precession. This "wobbling" is called nutation.

Some people like to say that the moment of inertia is constant, and so on and so forth, and that the torque precludes precession. It is very strong then when one suddenly lets go of a gyroscope, it does not fall under the action of gravity, but moves sideways instead. Why? Well, the component toward the axis of rotation, which we have analyzed, makes it go sideways! All the formulas in the usual like (20.9) are not going to be valid, because (20.15) is a static condition, valid only after the gyroscope is precessing steady. What really happens, in detail, is the following. If we were to let the axis accelerate down, so that it cannot precess in any manner (in the hope it's spinning there is no torque acting), and even a torque from gravity, because it is horizontal by now, but if we suddenly let go, then the axis will instantaneously be a torque from gravity. Anyone in our right mind would think that the top would fall, and that is what would occur, unless if the top is not spinning initially.

The axis actually does fall, as we would expect. But as soon as it falls, it then starts up and if this starting up were to continue, a torque would be required. In the absence of a torque in this direction, the gyroscope will fall. If the direction is again in the direction of the missing a torque. This gives the gyro a component of motion precessing around its vertical in steady precession. But because initial "over-shoots" the steady precessional velocity, and the axis actually rises again to the height from which it started. This, with followed by the end of the axis is precessing (the gyro followed by a point that is stuck in the end of an automobile tire). Call this the "gyroscopic effect". The eye follows the eye to follow, and it drops out quickly because of the friction in the optical bearings. In fact, the steady precessional drift (Fig. 20-2) - the gyro can't when spins the mass around the horizontal axis.

When the motion set in down, it's evident the gyro is pulled back to the axis of the star. Why? (They are not normal to each other, but we bring them in because we don't want the effect to affect the idea of the gyroscopic effect.) Just because it is a wonderful thing, and it is not a miracle! Two wheels rolling absolutely horizontally, and suddenly let go, then the simple precession equation would tell us that it precesses. But it was around a horizontal plane. This is impossible! Although we neglected friction, it is in the "horizontal" because it is of inertia about the rotation axis, and as it is moving about that axis, even slowly, it has a fixed angle measurement about the axis. We are difficult come round. If the gyro is precessing, then it must precess around the vertical axis. Hence then that it just continues if there is no change in the angular momentum! The answer is then the gyroscopic motion of the end of the wheel precesses to the average steady motion of the center of the equivalent rotating circle. That is, it oscillates in a vertical line. Because the law of the law of motion, it has a vertical component which is used by which is needed for the precession. As we see it has to go down a little, it needs to go up. It has to yield a little bit to the gravity; by turning its axis there a little bit it maintains the rotation about the vertical axis. Thus, the gyroscopic effect works.

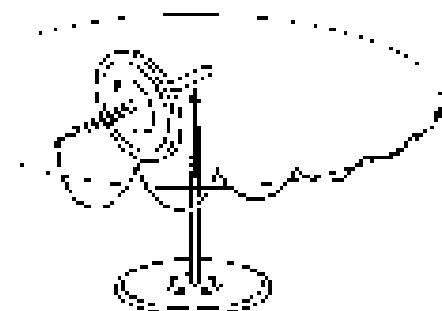


Fig. 20-5. Nutational motion of a gyroscope under gravity just after releasing and previously balanced.

20.4 Angular momentum of a solid body

Before we begin the subject of rotation in three dimensions, we shall consider, more qualitatively, a few effects that occur in two dimensions, regarding rotating rigid bodies. The main effect is that, in general, the angular momentum of a rigid body is not necessarily in the same direction as the center-of-mass velocity. Consider a wheel that is free to rotate about its shaft in a horizontal direction, so that the axis through the center of gravity, to be given \vec{v}_c (Fig. 20.6). When we spin the wheel around the axis, anybody knows that there will be shaking at the bearings because of the so-called τ_{fr} we have learned of. Qualitatively, we know that in the rotating system there is one tangential force acting on the wheel, trying to draw all mass toward the center. This tends to bring up the wheel and the wheel axis is perpendicular to the shaft. To resist this tendency, a torque is exerted by the bearings. This is a torque exerted by the bearings, causing a rate of change of angular momentum. How can there be a rate of change of angular momentum when we are simply turning the wheel about the shaft? Suppose we bring the angular velocity or ABC components ω_x , ω_y , ω_z perpendicular and parallel to the axis of the wheel. What is the angular momentum? The moment of inertia about these two axes are different, so the angular momentum components, which in this particular case of course only is equal to the moments of inertia times the corresponding angular velocity components, are at a different rate than are the angular velocity components. The effect of the angular momentum vector is at a distance from the center of the shaft. When we sum the effects, we have to sum the tangential momentum vectors in space, so we can result compute on the shaft.

Although it is much too complicated to prove here, there is a very important, and interesting property of the moments of inertia which is very descriptive and important, and which is the focus of our above analysis. This property is the following: Any rigid body, even an irregular one like a potato, possesses three mutually perpendicular axes through its C.G., such that the moment of inertia about one of these axes has the greatest possible value for any axis through the C.G. of the body; another of the axes has the minimum possible value, and the moment of inertia about the third axis is intermediate between these two (or equal to one of them). These axes are called the principal axes of the body, and they have the important property that if the body is rotated about one of them, its angular momentum is in the same direction as the angular velocity. If a body having axes of symmetry, the principal axes are along the symmetry axes.

If we take the x , y , and z axes along the principal axes, and call the corresponding principal moments of inertia I_x , I_y , and I_z , we may easily compute the angular momentum and the kinetic energy of rotation of the body by any angular velocity ω . If we resolve ω into components ω_x , ω_y , and ω_z , we find the components of the unit vectors \hat{i} , \hat{j} , \hat{k} , and along with ω we may readily calculate the components of

$$\vec{L} = I_x \omega_x \hat{i} + I_y \omega_y \hat{j} + I_z \omega_z \hat{k} \quad (20.16)$$

The kinetic energy of rotation is

$$K_E = \frac{1}{2}(I_x \dot{\omega}_x^2 + I_y \dot{\omega}_y^2 + I_z \dot{\omega}_z^2) \quad (20.17)$$

The Harmonic Oscillator

21-1 Linear differential equations

In the study of physics, usually the course is divided into a series of subjects such as mechanics, elasticity, optics, etc., and one subject after the other. For example, this course has so far dealt mainly with mechanics. But a strange thing occurs again and again: The same laws often appear in different fields of science, and even in other sciences, are often almost exactly the same, so that many phenomena have analogs in these different fields. This is the simplest example, the propagation of sound waves, in many ways analogous to the propagation of light waves. If we study something in great detail we discover that in a lot of our work it is no time so it would be if we were studying up in great detail. So the study of a science, even in one field may permit an extension of its knowledge. In fact, it is hard to realize from the first that such extensions are possible, for otherwise one might not understand the reason for spending a great deal of time and energy on what appears to be only a small part of mechanics.

The harmonic oscillator, which we are about to study, has close analogs in many other fields; although we start with a mechanical example of a weight on a spring, or a pendulum with a small swing, or certain other mechanical vibrations, we are really studying a certain differential equation. This equation appears again and again in physics and in all sciences, and in fact it is a sort of so many phenomena that its close study is well worth our while. Some of the phenomena involving this equation are the oscillations of a mass on a spring; the oscillations of charges flowing, laws and teeth in an electric circuit; the vibrations of a tuning fork which is generating sound waves; the analogous vibrations of the electrons in an atom, which generate light waves. The equation for the vibration of a system, such as a thermometer trying to adjust a temperature, complicated in reactions in chemical reactions, the growth and decay of bacteria in interaction with the food supply and the parasites the bacteria produce; losses eating rotting fruits, apples, and so on; there are phenomena follow equations which are very similar to one another, and this is the reason why we study this important oscillation in such detail. The equation is one called linear differential equations with constant coefficients. A linear differential equation with constant coefficients is a differential equation containing at most of terms, each term being a derivative of the dependent variable with respect to the independent variable, and multiplied by some constant. Thus

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_1 y' + a_0 y = f(t) \quad (2-1)$$

is called a linear differential equation of order n with constant coefficients (with a_i is constant).

21-2 The harmonic oscillator

Perhaps the simplest mechanical system whose motion follows a linear differential equation with constant coefficients is a mass on a spring. First the spring is stretched to balance the gravity; once it's balanced, we then displace the vertical displacement of the mass from its equilibrium position (Fig. 21-1). We will call the upward displacement y , and we also suppose that the spring is perfectly linear, that is, when we displace pulling back when the spring is stretched is precisely proportional to the amount of stretch. That is, the force is $-kx$ (with a

21-1 Three differential equations

21-2 The harmonic oscillator

21-3 Harmonic motion and circular motion

21-4 Initial conditions

21-5 Forced oscillations



Fig. 21-1. A mass on a spring: a simple example of a harmonic oscillator.

in. (21.2) is named as the "differential". Thus the mass takes the acceleration measured earlier:

$$m \frac{d^2x/dt^2}{dt} = -kx. \quad (21.3)$$

For simplicity, suppose it bounces (so we change our unit of time measurement) that the ratio $x_0/m = 1$. We shall then study the equation

$$\frac{dx/dt}{dt} = -\omega. \quad (21.4)$$

In other words come back to Eq. (21.2) with the x and m explicitly given.

We have already analyzed Eq. (21.3) in detail numerically when we first introduced the subject of mechanics; see section 10. Equations like Eq. (21.3) in fact oscillate. By numerical integration we found a curve (Fig. 9-4) which showed that if x was initially displaced, but $v=0$, it would bounce down and go through zero; we do not then follow it any further, but of course we know that x will keep going up and down—forever. When we calculated the motion numerically, we found that it went through the equilibrium point at $x = 1.5708$. The length of the whole cycle is four times the long, or $T = 0.2\pi/\omega = \pi$. This was found numerically, before we knew much calculus. We assume that in the meantime the Mathematics Department has brought for us a function, which, when differentiated twice, is equal to itself with a negative sign. (There are, of course, many ways of getting at this function in a direct fashion, but there are many ways of doing it, already knowing what the answer is.) The function is $\cos t$. If we differentiate this we find $d^2x/dt^2 = -\sin x$ and $d^2x/dt^2 = -\cos t = -x$. The function $x = \cos t$ starts at $t = 0$ with $x = 1$, and no initial velocity, that is, $v = 0$, exactly where which we started when we did our numerical work. Now that we know that $x = \cos t$, we can calculate a general value for the time at which it should zero, $x = 0$. The answer is $t = \pi/2$, or 1.5708. We were wrong in the last figure because of the errors of numerical analysis, but it was very close!

Now let us finish with the original problem. We reduce the time units to real seconds. What is the solution then? First of all, we might think that we can get the constant A involved by multiplying each by m (canceling). We let us do the equation, $x = A \cos t$; then we find $dx/dt = -A \sin t$, and $d^2x/dt^2 = -A \cos t = -x$. This we expect to occur because (as we did not succeed in solving Eq. (21.3)) now we get Eq. (21.3) again! That is, the \cos function one of the most important properties of the differential equations if we multiply a variable of the equation by any constant, it is still a solution. The mathematical reason for this is clear. If you substitute a constant multiple of the equation, say by A , we see that all derivatives cancel—multiplied by A , and therefore Ax is just a solution of the original equation as x was. The physics is this. In the following, if we have a weight on a spring, and pull it down a certain far, the force is twice as much; the resulting acceleration is twice as great; the velocity it acquires in a given time is twice as great; the distance covered in a given time is twice as great; and it has to cover twice as great a distance in order to get back to the origin because it is pulled down twice as far. So it takes the same time to get back to the origin irrespective of the initial displacement. To start with with a linear equation, the motion has to have some power, no matter how "strong" it is.

That was the wrong thing to do—but only taught us that we can multiply the solution by anything and it satisfies the same equation, but with different amplitude. After multiplying and dividing by m , we replace x with a constant multiple of x , we find that we must take the scale of time. In other words, Eq. (21.3) has a solution of the form

$$x = A \cos \omega t. \quad (21.5)$$

It is important to realize that ω is positive, since ω is the angular velocity of a spinning body, but we can only take it if we are not allowed to use our same letters for more than one thing! (because we put a subscript "0" on $x_0 > 0$, we are going to have more omega before long; i.e., we remember the ω_0 refers to the initial motion of this oscillator. Now why is (21.5) just? Just because we are more successful, because $d^2x/dt^2 = -A \omega^2 \cos \omega t$ and $d^2x/dt^2 = -\omega^2 x$.

So, if we have solved the equation that we really wanted to solve. The equation $\ddot{x} + \omega_0^2 x = -\sqrt{A} \cos(\omega t + \phi)$ reduces to Eq. (21.2) if $\omega_0^2 = \omega^2$.

The next thing we must investigate is the physical significance of ω_0 . By now, you know that the spring force is proportional to the angle it makes with \vec{x} , so if we neglect air resistance, it will go through a complete cycle, when the "angle" changes by 2π . The quantity ω_0 is often called the phase of the motion. In order to change ω_0 by 2π , the time must change by an amount τ_0 , called the period of one complete oscillation, of course τ_0 must be such that $\omega \tau_0 = 2\pi$. That is, ω_0 must account for one cycle of the angle, and that's why τ_0 will repeat itself—if we increase t by τ_0 , we add 2π to the phase. That is,

$$\omega_0 = 2\pi/\tau_0, \quad \lambda = \sqrt{\omega_0^2}, \quad (21.5)$$

Thus if we had a heavier mass, it would take longer to oscillate back and forth on a spring. That is because there are more inertia, and so, although forces are the same, it takes longer to get the mass moving. Or, if the spring is stiffer, it will move more quickly, and that is right because it less of the energy is stored.

Note that the period of oscillation of a mass on a spring does not depend in any way on whether it has started, how far from $x=0$ it is. The period is determined, but the amplitude of the oscillation is determined by the amount of energy $E(1/2)$. The amplitude is determined, in fact, by how we let go of the initial position, or starting conditions.

Actually, we have not quite found the most general solution of Eq. (21.2). There are other solutions. It should be clear why: because \ddot{x} is the force exerted by $x = \omega_0 \sin(\omega t + \phi)$ with an initial displacement and initial velocity. But it is possible, for instance, for the mass to start at $x = 0$ and $\dot{x} = 0$ and yet still give $\ddot{x} = 0$ impulse kick, so that it has some speed at $t = 0$. Such a motion is not represented by a sine—it is represented by a sine. To put it another way, if $x = 0$ is not a solution, then is it not obvious that if we want to happen to walk into the room at exactly x which we want (say, $x = 0^\circ$) and see the mass at $t = 0$ passing, $\dot{x} = 0^\circ$, we will keep on going past the same? Therefore, $x = 0$ is not just the most general solution, it must be possible to shift the beginning of time, or to speak, to an example where if we write the solution this way, $x = A \cos(\omega_0 t + \phi_0)$, where ϕ_0 is some constant. This also corresponds to shifting the origin of time to some new time t_0 . For this reason we may expand

$$x(t) = A \cos(\omega_0 t + \phi) = A \cos(\omega_0 t_0 + \phi_0) \cos(\omega_0(t - t_0))$$

and write

$$x = A \cos(\omega_0 t_0 + \phi_0) \cos(\omega_0 t),$$

where $A = \text{const}$ and $\phi_0 = -\omega_0 t_0$. Any one of these terms is a possible way to write the harmonic general solution of (21.2), that is, every solution of the differential equation $\ddot{x} + \omega_0^2 x = 0$, the solution in the world can be written as

$$(a) \quad x = A \cos(\omega_0 t + \phi), \quad (21.6)$$

or

$$(b) \quad x = A \cos(\omega_0 t + \phi_0), \quad (21.6)$$

or

$$(c) \quad x = A \cos(\omega_0 t_0 + \phi) \cos(\omega_0 t).$$

Some of the quantities in (21.6) have names: ω_0 is called the angular frequency, λ is the number of times by which the phase changes in a second. This is determined by the differential equation. The other constants are not determined by the equation, but by how the motion is started. Of these constants, A means the maximum displacement attained by the mass, and is called the amplitude of oscillation, ϕ is constant ϕ_0 is sometimes called the phase of the oscillation, but can be confusing, because ϕ_0 is people call it — a true phase, and not the phase change — with time. We might say that ϕ_0 is a phase shift from some definite zero. Let me put it differently. Different ϕ_0 is a constant of motion at different phases. That is to say, the value for ϕ_0 we want to start the phase at ϕ_0 , is completely specific.

21-3 Harmonic motion and circular motion

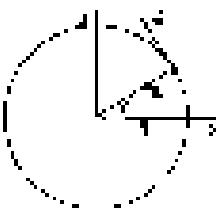


Fig. 21-2. A particle moving in a circular path at constant speed.

The last two sections are involved in the solution of Eq. (21.2) except for one, there might be some relationship to circles. This is artificial, of course, because there's no circle explicitly involved in the linear motion—it just goes up and down. We may point out that we have, in fact, already solved that differential equation when we were studying the mechanics of circular motion. If a particle moves in a circle with a constant speed v , the radius vector from the center of the circle to the particle turns through an angle whose size is proportional to the time. That we call the angle $\theta = \omega t$ (Fig. 21-2), for which $\omega = v/R$. We know that there is an angle between x and θ : $\alpha = \theta - x$ (Fig. 21-2). Now we also know that the position x , y is given by $x = R \cos \theta$, $y = R \sin \theta$.

$$x = R \cos \theta, \quad y = R \sin \theta.$$

Now what about the acceleration? What is the x -component of acceleration, $a_x = d^2x/dt^2$? We have already worked that out previously; it's the magnitude of the acceleration times the cosine of the projection angle, with a minus sign because it's toward the center. Now we also know that

$$a_x = -R \cos \theta = -\omega^2 R \cos \theta = -\omega^2 x. \quad (21.7)$$

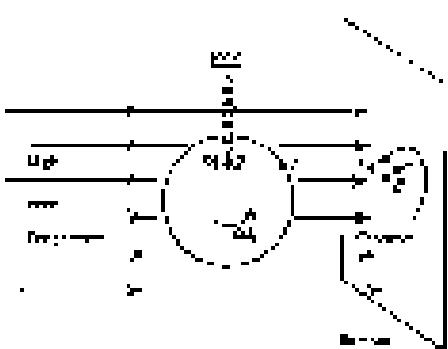


Fig. 21-3 Demonstration of the equivalence between simple harmonic motion and uniform circular motion.

In other words, when a particle is moving in a circle, the horizontal component of its velocity is an acceleration which is proportional to the horizontal displacement from the center. Of course we can have the solution for motion in a circle: $x = R \cos \omega t$. Equation (21.7) does not depend upon the radius of the circle, so for a circle of any radius, one has the same equation for a given ω . Thus for several reasons, we expect that the displacement of a mass in a ring will turn out to be proportional to the x and will, in fact, be exactly the same function as we would expect if we looked at the x -component of the position of an object rotating in a circle with angular velocity ω . And that is this: one can do this an experiment to show that the up-and-down motion of a mass on a spring is the same as that of a point going around in a circle. In Fig. 21-3 one light is projected on a screen, casts a shadow of a small sphere at x_0 and of a vertically oscillating mass, side-by-side. If we let go of the mass at the eight o'clock position from the right place, and if the stair speed is carefully adjusted so that the sphere is just, each should follow the other exactly. One can then check the numerical solution we obtained earlier with the engine function, and see whether that agrees very well.

Here we may point out that because uniform motion in a circle is so closely related mathematically to oscillatory, periodic motion, we can analyze most of our motion in a simpler way if we imagine it to be a projection of something going in a circle. In other words, although the distance y —from nothing to the oscillating problem, we may still artificially supplement Eq. (21.2) with another equation using y , and put the two together. If we do this, we will be able to analyze the one-dimensional oscillations with circular motions, which is a lot easier than having to solve a difference equation. The trick in doing this is to use complex numbers, a procedure we shall introduce in the next chapter.

21-4 Initial conditions

Now let us consider who determines the constants A and B , or α and β . Of course these are determined by how we start the motion. If we start the motion with just a small displacement, that is one type of way to do it; if we start with the initial displacement zero, then push up when we let go, we get still a different motion. The constants A and B , or α and β , or any other way of getting it, are determined, of course, by the way we initially started, not by any other degrees of the system. These are called the initial conditions. We would like to say that the initial conditions will be constants. Although this can be done using any one of the forms (21.6), it turns out to be easiest if we use Eq. (21.3c). Suppose that at $t = 0$ we have started with an initial displacement x_0 , and a zero velocity. At

This is the most general way we can start the motion. (We cannot specify the initial time t_0 at which it started, because that is determined by the spring, once we specify v_0 .) Now, we have two initial values A and B . We start with the equation for x :

$$x = A \cos(\omega_0 t) + B \sin(\omega_0 t).$$

Since we shall later need the velocity v_0 , we differentiate x and obtain

$$v = -\omega_0 A \sin(\omega_0 t) + \omega_0 B \cos(\omega_0 t).$$

These are again two initial conditions, but we have *zero* knowledge about A and B , and $v_0 = 0$. So if we put $v_0 = 0$ in these equations, on the left we get A_0 and B_0 . Because that is what A and B did when $v_0 = 0$; also, we know that the cosine of zero is unity, and anything of zero is zero. Therefore we get

$$A_0 = 0 \cdot 1 = 0 \cdot 0 = 0,$$

and

$$B_0 = -\omega_0 A_0 \cdot 0 + \omega_0 \cdot 0 \cdot 1 = \omega_0 \cdot 0.$$

So for simple harmonic motion we find that

$$A = A_0, \quad B = v_0/\omega_0.$$

From these values of A and B , we can now work x if we wish.

That is *too bad* for our purpose, but there is one physically interesting thing to check, and that is the conservation of energy. Since there was no friction present, energy ought to be conserved. Let us see. We know that

$$x = A \cos(\omega_0 t) + B \sin(\omega_0 t)$$

then

$$v = -\omega_0 A \sin(\omega_0 t) + \omega_0 B \cos(\omega_0 t).$$

Now let us look at what the kinetic energy, T , and what the potential energy are. The potential energy at any moment is $(1/2)kx^2$, where x is the displacement and k is the constant of the spring. If we substitute for x , using our expression above, we get

$$E = \frac{1}{2}kx^2 = \frac{1}{2}k(A \cos(\omega_0 t) + B \sin(\omega_0 t))^2.$$

Of course the potential energy is *not* constant; the potential never becomes negative, because x always has energy in the spring, but the amount of energy fluctuates with x . The kinetic energy, on the other hand, is $(1/2)m v^2$ and by a similar argument we get

$$T = \frac{1}{2}mv^2 = \frac{1}{2}m\omega_0^2(A \cos(\omega_0 t) + B \sin(\omega_0 t))^2.$$

Now the kinetic energy is zero when x is at its maximum, because then v is zero (velocity); on the other hand, it is maximal when x is passing through zero, because then v is moving fastest. This variation of the kinetic energy is just the opposite of that of the potential energy. If we let the total energy E equal to be a constant, if we note that $E = m\dot{x}^2$, we see that

$$\dot{x} + 0 = \sqrt{m\omega_0^2}[(A \cos(\omega_0 t) + B \sin(\omega_0 t)) - \sin(\omega_0 t)] = \sqrt{m\omega_0^2}x'$$

The energy is dependent on the sign of the amplitude. If we lower twice the amplitude, we get an oscillation which has four times the energy. The average potential energy is $(1/2)A^2$. The maximum and, therefore, half the total, and the average kinetic energy is likewise $(1/2)$ the total energy.

21-5 Forced oscillations

Next we shall discuss the forced harmonic oscillation, *i.e.*, one in which there is an external driving force on top. The equation then is the following:

$$m\ddot{x} + \omega_0^2 x = -Av + F(t). \quad (21.8)$$

We would like to find out what happens in these circumstances. The external driving force can have various kinds of functional dependences on the time; the first one that we shall analyze is very simple—*i.e.*, suppose that the force is oscillating:

$$F(t) = F_0 \cos \omega t. \quad (21.5)$$

We find, however, that this ω is not necessarily ω_0 , we have to make some assumptions; the forcing may be done at different frequencies. So we try to solve Eq. (21.4) with the special force (21.5). What is the solution of (21.4)? One special solution, we shall discuss the more general cases later), is

$$x = C \cos \omega t, \quad (21.6)$$

where the constant is to be determined. In other words, we might suppose that if we kept pushing back and forth the mass would *oscillate* back and forth in step with the force. We could try it anyway. So let's put (21.6) into (21.4), and get

$$m\omega^2 C \cos \omega t = -m \omega^2 C \cos \omega t + M \ddot{x} \cos \omega t. \quad (21.7)$$

We have also put in $k = m\omega_0^2$, so that we will understand the equation better at the end. Now because the cosine object is zero, we can divide it out, and that shows that (21.7) is, in fact, a solution, provided we pick C just right. The answer is that $C = M/m$:

$$C = M/m(\omega_0^2 - \omega^2). \quad (21.8)$$

That is, an oscillates at the same frequency as the force, but with an amplitude which depends on the frequency of the force, and so upon the frequency of the mass vibration of the oscillator. It means, then, that if ω is very much compared with ω_0 , then the displacement and the force are in the same direction. On the other hand, if we shake it back and forth very fast, then (21.8) tells us that C is negative. If ω is above the natural frequency ω_0 , the harmonic oscillator ω . (We will call ω the *natural frequency* of the harmonic oscillator, and ω the applied frequency.) At very high frequencies the displacement may become very large and there is then *resonance* and *amplification*.

Of course the solution we have found is not sufficient only in things described just right, for otherwise there is a part which usually does not enter at all. The other part is called the *transient response* to $F(t)$, while (21.6) and (21.8) are called the *steady-state response*.

According to our form, ω (21.8), a very remarkable thing should also occur: if ω is almost exactly the same as ω_0 , then C could approach infinity. So we require the frequency of the force to be "in exact" with the natural frequency. Then we should get an enormous displacement. This is well known, especially when we pushed a child on a swing. It does not work very well unless our eyes and just our seat stay at a column. If we happen to get the right timing, then the swing goes very high, but if we lose the swing timing, the swing is over, may be pushing after we should be pulling, and so on, and it makes her work.

If we make ω exactly equal to ω_0 , we find that it should oscillate at us infinite amplitude, which is, of course, impossible. Therefore it follows that something goes wrong with the equation. There are some other external forces, and other forces, which are not in (21.8) but which occur in the real world. So the amplitude does not necessarily do some reason, it may be that the spring has got

Algebra

22-1 Addition and multiplication

In our study of oscillating systems we shall have occasion to use one of the most remarkable, or most remarkable formulas in all of mathematics. From the physicist's point of view we could bring up the Pythagorean theorem in two minutes or so, and be done with it. But, as you know, much fun intellectual enjoyment is to be had in multiplying our instead of just spending a few minutes on this amazing result, we shall pursue the issue by its proper setting, in the grand design of that branch of mathematics which is called elementary algebra.

Now you may ask, "What is mathematics doing in a physics book?" We have several possible answers. And, at some level, mathematics is an important tool. In fact, we did only now begin giving the formula in two minutes. On the other hand, in theoretical physics we discover that all the laws can be written in mathematical form, and that this has a certain simplicity and beauty about it. So, it is only, in order to understand nature, it may be necessary to have a deeper understanding of mathematical relationships. But the real reason is that the subject is enjoyable, and it is enough to turn six and seven up in different ways, and we have a dozen courses in different departments. Since compartmentalization is really artificial, and we should take our intellectual pleasures where we find them.

Another reason for treating them carefully is algebra now, even though most of the actual algebra in high school is that part was the first time we studied it. The equations were therefore, and it was the work, not as physicists now. Every so often it is a good pleasure to look back to see what territory has been covered, and who has gone map or not of the whole thing is. Perhaps we can contribute to the Mathematics Department will present this as one mechanics in such a way as to give when I say we were trying to work on the physics course.

The subject of algebra will not be developed from the point of view of symmetry, condition, exactly, because the mathematicians are mainly interested in how various properties and facts are demonstrated, and new assumptions are obviously required, and what is not required. They are not interested in the how of why they prove. For example, we may find the Pythagorean theorem quite interesting, that the sum of the squares of the sides of a right triangle is equal to the square of the hypotenuse; this is an extremely fine, a very simple thing which may be appreciated without discussing the question of how to prove it, or what axioms are required. So, in the same spirit, we will describe qualitatively. This may put in that way, the system of elementary algebra. We say elementary algebra, because there is a branch of mathematics called modern algebra, in which some of the ideas such as $a^b = ab$, are abandoned, and it is not called algebra, but we shall not discuss that.

To continue the subject we start in the middle. We suppose that we already know what integers are, what zero is and what it means to increase a number by one unit. You may say, "That is not in the middle!" But it is considerable from a mathematical standpoint, because we could go even further back and describe the theory of sets in order to derive some of these properties of integers. But we are not going in that direction, the discussion of mathematical philosophy and mathematical logic, but rather in the other direction, from the assumption that we know what integers are and we know how to count.

Consider n a certain number of an integer, and we count successively n a number of times, the number we call $n+1$. And that defines addition of integers.

22-1 Addition and multiplication

22-2 The inverse operations

22-3 Abstraction and generalization

22-4 Approximation, irrational numbers

22-5 Complex numbers

22-6 Imaginary numbers

Once we have defined addition, then we can consider this: if we start with nothing and add a to it, a times in succession, we call the result multiplication of integers, we call it a times a .

Now we can also have a succession of multiplications: if we start with 1 and multiply by a, b times in succession, we call that relative to a power of.

Now as a consequence of these definitions it can be easily shown that all of the following relationships are true:

$$\begin{array}{ll}
 \text{(a)} \quad a + b = b + a & \text{(b)} \quad a + (b + c) = (a + b) + c \\
 \text{(c)} \quad ab = ba & \text{(d)} \quad a(b - c) = ab - ac \\
 \text{(e)} \quad (ab)c = a(bc) & \text{(f)} \quad (ab)^n = a^n b^n \\
 \text{(g)} \quad a^m \cdot a^n = a^{m+n} & \text{(h)} \quad (a^m)^n = a^{mn} \\
 \text{(i)} \quad a^{-1} \cdot a = 1 & \text{(j)} \quad a^0 = 1 \\
 \text{(k)} \quad a^{-1} = a^{-1} &
 \end{array} \tag{2.1}$$

These results are well known and we shall not belabor the point, we merely list them. Of course, 1 and 0 have special properties; for example, $a + 0 = a$, $a \cdot 1 = a$, and $0 \cdot 0 = 0$ is the first proof of 0.

In this chapter we will also look at a few other properties like continuity and ordering, which are very hard to define; we will let the reader carry on it. Furthermore, it is definitely true that we have written down too many "definitions"; some of them may be redundant from the others but we shall not worry about such details.

23-3. The inverse operations

In addition to the conventional four of addition, multiplication, and raising to a power, we have also two inverse operations which are defined as follows. Let us assume that a and b are given, and that we wish to find other values of b satisfying such equations as $a + b = c$, or $ab = c$. If $a + b = c$, b is defined as $c - a$, which is called subtraction. The operation called division is also clear: if $ab = c$, then $b = c/a$ defines division. A solution of the equation $ab = c$ "backwardly" shows if we have a power $b^m = c$ and we ask ourselves, "What is b ?" it is called the *mth root of c* : $b = \sqrt[m]{c}$. For instance, if we are considering the following question, "What integer, raised to the third power, equals 8?" the one answer is called the *cubic root of 8*: $\sqrt[3]{8} = 2$. These are direct and inverse equalities, these are the inverse problems associated with powers, and the other inverse problem would be, "To what power must we raise 2 to get 8?" This is called taking the logarithm. If $2^x = 8$, we write $x = \log_2 8$. The fact that it has a unique solution relative to the value x does not mean that it is any less elementary, it just applies to integers, like the other processes. Although logarithms are a little in an algebra class, in practice they are, of course, just as simple as roots; they are just a different kind of solution of a algebraic equation. The direct and inverse operations are summarized as follows:

(a) addition	(b) subtraction
$a + b = c$	$b = c - a$
(c) multiplication	(d) division
$ab = c$	$b = c/a$
(e) power	(f) root
$b^m = c$	$b = \sqrt[m]{c}$
(g) power	(h) logarithm
$b^x = c$	$x = \log_b c$

Now here is the idea. These relationships are rules one cannot break, since they follow from the definitions of addition, multiplication, and raising to a power. We are going to discuss whether or not we can broaden the class of systems

which a , b , and c represent so that they will obey those same rules, although the processes for $a = b - c$, and $bc = 0$, will not be deductible in terms of the operation of adding 1. For instance, an associative multiplication by induction.

21-3 Abstraction and generalization

When we try to solve simple algebraic equations, taking all these definitions, we soon discover some insoluble problems, such as the following: Suppose that we try to solve the equation $a = b - c$. That means, according to our definition of subtraction, that we must find a number which, when added to c , gives b . And of course there is no such number, because we consider only positive integers; this is an insoluble problem. However, the plan, the great idea, is that: abstraction and generalization. From the whole structure of algebra, rules plus integers, we restrict the original definitions of addition and multiplication, but we leave cases like (22.1) and (22.2), and assume these to be true in general on a wider class of numbers, even though they are originally defined on a smaller class. Thus, in characterizing integers systematically by definition of rules, we use the rules as the definition of the symbols, which then represent a more general kind of number. As an example, by working with fractions alone we can show that $3 - 5 = 0 - 2$. So that we can see that one can make all subtractions, provided we define a whose set of non members is $0, -1, 0, -2, 0, -3, 0, -4$, and so on, called the negative integers—but we may use $+$ —the other rules, like $a + b = ab$ —are not enough to find what the rules are for multiplying negative numbers, and we will discover, in fact, that “ $+$ ” of the rules can be associated with negative as well as positive integers.

So we have increased the range of objects over which the rules work, but the meaning of the symbols is different.

One cannot say, for instance, that -2 times 3 really means to add 3 together successively -2 times. That means nothing. But nevertheless everything will work out all right according to the rules.

An interesting problem comes up in taking powers. Suppose that we wish to discover what $a^{1/2}$ means. We know already that $a^{1/2}$ is a solution of the problem, $a = x^2 + y^2$, i.e., knowing that, we know that $a^{1/2} = \sqrt{a}$. Therefore $a^{1/2} \cdot a^{1/2} = a^{1/2}a^{1/2}$, by the definition of division. With a little more work, this can be reduced to a . So we find that the negative powers are the reciprocals of the positive powers, but $a^{1/2}$ is a meaningless symbol, because if a is a positive or negative integer, the square of it is greater than 1 , and we do not yet know what we mean by “divided by a number greater than 1 ”.

Onward! The great plan is to continue the process of generalization; whenever we find another problem that we cannot solve we extend our realm of numbers. Consider division: we cannot find a number which is an integer when a large integer, which is equal to the result of dividing 3 by 5 . But if we suppose that all fractional numbers satisfy the rules, then we can talk about multiplication and adding fractions, and everything works as well as it did before.

Take another one step of progress: what is $a^{3/2}$? We know only that $(1/a)^2 = 1$, since this is the definition of $1/a$. So we know also that $(a^{1/2} \cdot a)^2 = a^{1/2} \cdot a^{1/2} = a^2$, because this is one of the rules. Then by the definition of roots we find that $a^{3/2} = \sqrt[3]{a^2}$.

In this way, then, we can refine what we mean by putting fractions in the number system, by using the rules themselves to help us determine the meaning—it is not arbitrary. It is a new rule, but that all the “old” work for positive and negative integers, as well as for fractions.

We go on in the process of generalization. Are there any other equations we cannot solve? Yes, for example, it is impossible to solve this equation, $x = 2^{-x} = \sqrt{x}$. It is impossible to find a number which is natural (a fraction) whose square is equal to 2 . It is very easy for us in modern days to answer this question. We knew the decimal system, and as we have no difficulty in specifying the meaning of an ordinary decimal as a type of expression relative to the expansion of 2 . Historically, this idea presented great difficulty to the Greeks. To say

define $\sqrt[n]{x}$, which is meant by it requires that we add some substance of continuity and ordering, and it is, in fact, what we have done at this point. It was done formally and rigorously, by Dedekind. However, without worrying about the mathematical rigor of the thing, it is quite easy to understand that what we mean is that we are going to find a whole sequence of approximating fractions, perfect fractions (from now on they'll be called), which stop at somewhere, is of course rational, which goes on going, getting closer and closer to the desired result. That is good enough for what we wish to discuss, and it permits us to make ourselves in numerical numbers and to calculate things like the square root of 2 to any accuracy that we desire, with enough work.

23.4 Approximating irrational numbers

The next problem comes with what happens with the irrational powers. Suppose that we want to extract, for instance, $10^{1/3}$. In principle, the answer is simple enough. If we express this as the sum of 10 to a certain number of decimal places, then the power is a fixed, and we can then the approximation, just using the above method, and get an answer which is to $10^{1/3}$. Then we may just repeat a few more decimal places (it is again rational), to be as approximate as we like. This time it's much higher now because there is a much bigger denominator in the fraction, and get a better approximation. Of course we are going to get some extremely high numbers involved here, and the work is quite difficult. How can we cope with this problem?

In the computations of square roots, cube roots, and other small roots, there is some arithmetical process available to which we can go one decimal place at a time. But the situation of higher powers is rather different, irrational powers and the logarithms that go with them, the inverse relation. The reason that there is no simple arithmetical procedure with these. The first tables were built up which permit us to calculate these powers, and these are called the tables of logarithms. At the tables of powers, depending on which way the table is set up. It is only a question of looking up; if we take our same number to an irrational power, we can look it up easier than having to compute it. Of course, with computation, it is just a technical problem, but it is an interesting one, and a most historical study. In the first place, not only do we have the problem of extracting $x = 10^{1/3}$, but we also have the problem of solving $10^x = 2$ or $x = \log_{10} 2$. This is not a number where we have to define a new kind of number for the root, it is merely a nonreal number. The answer is simply an irrational number, or, writing decimal, will be a new kind of a number.

Let us now discuss the problem of calculating solutions of such equations, the power idea is really very simple. If we could calculate 10^1 , and $10^{1/10}$, and $10^{1/100}$, and $10^{1/1000}$, and so on, and multiply the n all together, we would get $10 \cdot 10^{1/10} \cdot 10^{1/100} \cdots 10^{1/10^n}$, and then is the general answer which things were. Our instead of calculating $10^{1/10}$ and so on, we shall calculate $10^{1/2}$, $10^{1/4}$, and so on. Before we start, we should explain why we make so much work with 10, instead of some other number. Of course we realize that logarithm tables are of great practical utility, quite aside from the mathematical problem of taking roots, since with any base at all.

$$\log_b(x^n) = \log_b x + \log_b n. \quad (23.1)$$

We are all aware that the fact that one can use this is that is a practical way to calculate numbers if we have a table of logarithms. There's quite less is with a two, which we compute? It makes no difference what base is used; we can use the same although b , the time, and if we are using long division, or any particular base, even if it's not large division or any other way, namely by a change in scale, a multiplying factor. If we multiply $10 \cdot (23.1)$ by 61, it is just as true, and if we had a table of \log_6 with a base 6, and similarly eliminating all of our entries by 61, then we'd have no essential difference. A priori then we know the logarithms of all the numbers to the base 6. In other words, we can solve the equation $6^x = 61$ for x by a table of \log_6 .

we have a table. The problem is to find the logarithm of the state number, or to some other base, for every number x . We would like a table of $\log_b x$. It is easy to do, because we can always write $x = b^y$, which means $\log_b x = y$. As a matter of fact, $b = \exp x$. Then, if we put $x = 1$ and take $\log_b 1$, we see that $\log_b 1 = \log_b b^0 = 0$. In other words, 0 is the base column in base b . Thus $\log_b x = \log_b b^y = y$, which is a constant, times the entry in the base b . Therefore any log table is equivalent to any other log table if we multiply by a constant and then, for convenience, if $y < 0$. This permits us to choose a particular base, and for convenience we take the base ten. (The question may arise as to whether there is any normal base, one base in which things are still simpler, and we shall try to find an answer to that after we, by means, shall introduce the base 10.)

Table 22-1

Square Root of Ten

Number x	$10\log x$	10^x	$10^{10\log x} = x$
1	0.00	1.000000	1.00
1.12	0.12	3.14278	3.14
1.25	0.25	3.77328	3.77
1.38	0.38	4.37177	4.37
1.51	0.51	4.84179	4.84
1.64	0.64	5.34457	5.34
1.77	0.77	5.81137	5.81
1.90	0.90	6.28243	6.28
2.03	1.03	6.70370	6.70
2.16	1.16	7.13497	7.13
2.29	1.29	7.57621	7.57
2.42	1.42	8.02745	8.02
2.55	1.55	8.48869	8.48
2.68	1.68	8.95997	8.96
2.81	1.81	9.43121	9.43
2.94	1.94	9.90245	9.90
3.07	2.07	10.37370	10.37
3.20	2.20	10.84502	10.84
3.33	2.33	11.31634	11.31
3.46	2.46	11.78766	11.79
3.59	2.59	12.25897	12.26
3.72	2.72	12.73029	12.73
3.85	2.85	13.20161	13.20
3.98	3.00	13.67293	13.67
4.11	3.11	14.14425	14.14
4.24	3.24	14.61557	14.61
4.37	3.37	15.08689	15.09
4.50	3.50	15.55821	15.56
4.63	3.63	16.02953	16.03
4.76	3.76	16.49985	16.50
4.89	3.89	16.97017	16.97
5.02	4.00	17.44150	17.44
5.15	4.15	17.91282	17.91
5.28	4.30	18.38414	18.38
5.41	4.41	18.85546	18.86
5.54	4.54	19.32678	19.33
5.67	4.67	19.79810	19.80
5.80	4.80	20.26942	20.27
5.93	4.93	20.74074	20.74
6.06	5.06	21.21206	21.21
6.19	5.19	21.68338	21.68
6.32	5.32	22.15470	22.15
6.45	5.45	22.62602	22.63
6.58	5.58	23.09734	23.10
6.71	5.71	23.56866	23.57
6.84	5.84	24.03998	24.04
6.97	5.97	24.51130	24.51
7.10	6.10	24.98262	24.98
7.23	6.23	25.45394	25.46
7.36	6.36	25.92526	25.93
7.49	6.49	26.39658	26.40
7.62	6.62	26.86790	26.87
7.75	6.75	27.33922	27.34
7.88	6.88	27.81054	27.81
8.01	7.00	28.28186	28.28
8.14	7.14	28.75318	28.75
8.27	7.27	29.22450	29.23
8.40	7.40	29.69582	29.69
8.53	7.53	30.16714	30.17
8.66	7.66	30.63846	30.64
8.79	7.79	31.10978	31.11
8.92	7.92	31.58110	31.58
9.05	8.05	32.05242	32.05
9.18	8.18	32.52374	32.52
9.31	8.31	32.99506	32.99
9.44	8.44	33.46638	33.47
9.57	8.57	33.93770	33.94
9.70	8.70	34.40902	34.41
9.83	8.83	34.88034	34.88
9.96	9.00	35.35166	35.35
10.09	9.19	35.82298	35.82
10.22	9.38	36.29430	36.29
10.35	9.55	36.76562	36.77
10.48	9.72	37.23694	37.24
10.61	9.89	37.70826	37.71
10.74	10.00	38.17958	38.18
10.87	10.17	38.65090	38.65
10.99	10.34	39.12222	39.12
11.12	10.51	39.59354	39.59
11.25	10.68	40.06486	40.06
11.38	10.85	40.53618	40.54
11.51	11.00	41.00750	41.01
11.64	11.17	41.47882	41.48
11.77	11.34	41.95014	41.95
11.90	11.51	42.42146	42.42
12.03	11.68	42.89278	42.89
12.16	11.85	43.36410	43.36
12.29	12.00	43.83542	43.84
12.42	12.17	44.30674	44.31
12.55	12.34	44.77806	44.78
12.68	12.51	45.24938	45.25
12.81	12.68	45.72070	45.72
12.94	12.85	46.19202	46.19
13.07	13.00	46.66334	46.66
13.20	13.17	47.13466	47.13
13.33	13.34	47.60598	47.61
13.46	13.51	48.07730	48.08
13.59	13.68	48.54862	48.55
13.72	13.85	49.01994	49.02
13.85	14.00	49.49126	49.49
13.98	14.17	49.96258	49.97
14.11	14.34	50.43390	50.44
14.24	14.51	50.90522	50.91
14.37	14.68	51.37654	51.38
14.50	14.85	51.84786	51.85
14.63	15.00	52.31918	52.32
14.76	15.17	52.79050	52.79
14.89	15.34	53.26182	53.27
15.02	15.51	53.73314	53.74
15.15	15.68	54.20446	54.21
15.28	15.85	54.67578	54.68
15.41	16.00	55.14710	55.15
15.54	16.17	55.61842	55.62
15.67	16.34	56.08974	56.09
15.80	16.51	56.56106	56.57
15.93	16.68	57.03238	57.04
16.06	16.85	57.50370	57.51
16.19	17.00	57.97502	57.98
16.32	17.17	58.44634	58.45
16.45	17.34	58.91766	58.92
16.58	17.51	59.38898	59.39
16.71	17.68	59.86030	59.87
16.84	17.85	60.33162	60.34
16.97	18.00	60.80294	60.81
17.10	18.17	61.27426	61.28
17.23	18.34	61.74558	61.75
17.36	18.51	62.21690	62.22
17.49	18.68	62.68822	62.69
17.62	18.85	63.15954	63.16
17.75	19.00	63.63086	63.64
17.88	19.17	64.10218	64.11
18.01	19.34	64.57350	64.58
18.14	19.51	65.04482	65.05
18.27	19.68	65.51614	65.52
18.40	19.85	65.98746	65.99
18.53	20.00	66.45878	66.46
18.66	20.17	66.92010	66.93
18.79	20.34	67.39142	67.39
18.92	20.51	67.86274	67.87
19.05	20.68	68.33406	68.34
19.18	20.85	68.80538	68.81
19.31	21.00	69.27670	69.28
19.44	21.17	69.74802	69.75
19.57	21.34	70.21934	70.22
19.70	21.51	70.69066	70.69
19.83	21.68	71.16198	71.17
19.96	21.85	71.63330	71.64
20.09	22.00	72.10462	72.11
20.22	22.17	72.57594	72.58
20.35	22.34	73.04726	73.05
20.48	22.51	73.51858	73.52
20.61	22.68	73.98990	73.99
20.74	22.85	74.46122	74.47
20.87	23.00	74.93254	74.94
20.99	23.17	75.40386	75.41
21.12	23.34	75.87518	75.88
21.25	23.51	76.34650	76.35
21.38	23.68	76.81782	76.82
21.51	23.85	77.28914	77.29
21.64	24.00	77.76046	77.77
21.77	24.17	78.23178	78.24
21.90	24.34	78.70310	78.71
22.03	24.51	79.17442	79.18
22.16	24.68	79.64574	79.65
22.29	24.85	80.11706	80.12
22.42	25.00	80.58838	80.59
22.55	25.17	81.05970	81.06
22.68	25.34	81.53102	81.54
22.81	25.51	81.99234	82.00
22.94	25.68	82.46366	82.47
23.07	25.85	82.93500	82.94
23.20	26.00	83.40632	83.41
23.33	26.17	83.87764	83.88
23.46	26.34	84.34896	84.35
23.59	26.51	84.82028	84.83
23.72	26.68	85.29160	85.29
23.85	26.85	85.76292	85.77
23.98	27.00	86.23424	86.24
24.11	27.17	86.70556	86.71
24.24	27.34	87.17688	87.18
24.37	27.51	87.64820	87.65
24.50	27.68	88.11952	88.12
24.63	27.85	88.59084	88.59
24.76	28.00	89.06216	89.07
24.89	28.17	89.53348	89.54
25.02	28.34	90.00480	90.01
25.15	28.51	90.47612	90.48
25.28	28.68	90.94744	90.95
25.41	28.85	91.41876	91.42
25.54	29.00	91.88008	91.89
25.67	29.17	92.35140	92.36
25.80	29.34	92.82272	92.83
25.93	29.51	93.29404	93.29
26.06	29.68	93.76536	93.77
26.19	29.85	94.23668	94.24
26.32	30.00	94.70800	94.71
26.45	30.17	95.17932	95.18
26.58	30.34	95.65064	95.66
26.71	30.51	96.12196	96.13
26.84	30.68	96.59328	96.59
26.97	30.85	97.06460	97.07
27.10	31.00	97.53592	97.54
27.23	31.17	97.99724	98.00
27.36	31.34	98.46856	98.47
27.49	31.51	98.93988	98.94
27.62	31.68	99.41120	99.42
27.75	31.85	99.88252	99.89
27.88	32.00	100.35384	100.36

* This is a definite arithmetic procedure, but the easiest way to find the square root of any number N is to choose some fairly close, but not too large, starting value $a_0 = g_0 + (\delta/2)$, and now the average of a_0 and N/a_0 . The next choice is $a_1 = \frac{1}{2}(a_0 + N/a_0)$. Convergence is very rapid— π numbers, for example, double each time.

10¹⁰th power of 10¹⁰⁻⁰⁰³ to get back to 10, so we had better not start with too big a number; it has to be close to 1. What we notice is that the small numbers that we started in 1 begin to look as though we are merely dividing by 2 each time: we see: 16 it becomes 8, then 16, 223; so it is clear that our result is approximately. If we take another look, we shall see 1000 is something, and rather than actually take all the steps to 1000, we guess at the ultimate limit. When we take a step of fraction 5 of 1000 as a quotient we ask what will the answer be? Of course it will be some number close to 1000¹⁰⁻⁰⁰³; but exactly 100022311.3, however—we can get a better answer by the following trick. We subtract the 1, and then divide by the power 5. This ought to convert all the numbers to the same value. We see that they are very closely spaced. At the top of the table they are not equal, but as they come down, they get closer and closer to a constant value. What is the value? Again we look to see how the last is changing, and it has changed next to 1000 divided by 211, by 24, by 51 by 10. These changes are obviously half of each other, very closely, as we have seen. Therefore, if we kept going, the changes would be 13, 7, 3, 2 and 1, minus or less, or a total of 26. Thus we have only 26 more to go, and so we find that the final number is 1.33023. Actually, we shall have to add the error, which should be 1.33023, but to keep it realistic, we shall not alter anything in the arithmetic.) From this table we can now calculate any power of 10 by multiplying the power, call it 100404.

i.e. we have easily calculated logarithms, because the process we shall use is where logarithmic values actually come from. The procedure is shown in Table 22-2, and the numerical values are shown in Table 22-3 (columns 2 and 3).

Table 22-2

Calculation of a logarithm $\log_{10} 2$

$$\begin{aligned} 1 &= 1.00000 \quad 1.124682 \\ 1.124682 &= 1.00461 \dots 1.000000, \text{ etc.} \\ \dots 2 &= (1.124682)(1.00461) \dots 1.000000(1.000000) \\ &= 10^{\left[\frac{1}{1024} (223 - 12 - \ln + 7 + 0.25) \right]} = 10^{\left[\frac{18.754}{1024} \right]} \\ &= 1.000000 \quad \left(\frac{18754}{1024} = 0.25 \right) \\ 10^{\frac{1}{1024}} &= 1.000201 \end{aligned}$$

Suppose we want the logarithm of 2. That is, we want to know to what power we must raise 10 to get 2. Can we raise 10 to the 1/2 power? No; that is too big. In other words, we can see that the answer is going to be bigger than 1/2, and less than 1/2. So we take the factor 10^{1/2} out; we divide 2 by 1.333 . . . and get 1.333 . . . , and so on, and now we know that we have taken away 0.22000001 from the logarithm. Then, number 1.124 . . . is zero the number which log₁₀ we have. When we are finished we shall add back the 1/4, or 25% there. Now we look at the table for the next number just below 1.124 . . . and that is 1.004603. We therefore divide by 1.004603 and get 1.000201. From that we observe the 2 can be made up of a product of numbers that are in Table 22-1, as follows:

$$2 = (1.124682)(1.004603)(1.000000)(1.000000)$$

The 1.000000 factor (1.000000) fell over naturally, which is beyond the range of our table. To get the logarithm of this factor, we see in the result that $10^{1/2} = 1 = 1.33023$ or 1/1024. We find $1 = 0.25 \times 1$. Therefore our answer is 10 to the following power: 1/26 = 12 - 12 - 1/4 - 0.25/1/1024. Adding these together we get .08251/1024. Dividing we get 0.00001, so we know that the $\log_{10} 2 = 0.30103$, which I happen to be right to 5 digits!

This is how logarithms were originally computed by Mr. Briggs of Hullux, in 1623. He said, "I computed approximately 10 square roots of 20.7. We know he was

we've computed only the first 27, because the rest of them can be obtained by this trick with $a = 2$. It's worth noticing that square root of 10 means seven digits, which is no more than twice the six digits we did; however, it was more work because he calculated to sixteen decimal places, and then reduced his answer to fourteen digits by putting off the point. And there were corresponding errors. He used tables of 'logarithms' to fourteen decimal places by John Napier, which is quite curious. But all logarithmic tables for common logarithms were calculated from Mr. Briggs' tables by reducing the number of decimal places. Only in modern times with the BPPA and computing machines have new tables been independently computed. There are such 'one-liner' methods for computing logarithms today, using certain series expansions.

In the above process, we discovered something not so interesting: one thing that you've heard before is we can calculate \log_e easily; we have discovered that $10^x = 1 + 2.30258 \dots$ by sheer numerical analysis. Of course this also means that $\log_{10} x = 1 + \text{something}$ where $x > 1$. Now log₁₀ is to any other base a merely a multiple of log₁₀ from the result that $\log_a x = \log_{10} x / \log_{10} a$. So 10 was used only because we have 10 fingers, and the arithmetic is nice, but 2 is also really not a bad base, one that has nothing to do with the number of fingers on human beings. We might say it's a change of base of logarithms is quite convenient and useful because any the method which people have chosen is to calculate the logarithms by multiplying all the logarithms in the base 10 by "1.397...". This then can't be able to using some other base, and this is called the natural base or base e. Note that $\log_e(1 + x) \approx x$, or $e^x \approx 1 + x$ as $x \rightarrow 0$.

It is easy enough to find out who x is: $y = 0.101025 \dots$ or $10^{0.101025} \dots$ an irrational power of 10. Both of the successive square roots of 10 can be used to compute, not just logarithms, but also 10 to any power, so let me start to calculate this natural base e. For convenience we transform $0.101025 \dots$ into $441.73/4096$. Now 441.73 is $2^6 + 120 + 32 + 16 + 2 = 0.73$. Therefore x , since it is an exponent of a sum, will be a product of the numbers

$$0.73428(0.33853)(0.674402)(0.28774)(0.15153)(0.08085)(0.001642) = 2.3184$$

(The only problem is the last one, which is 0.001642, which is not in the table, but we know that it is small enough, the answer is $1 + 2.30258 \dots$). When we multiply all the \log_e 's together, we get 2.3184 (it should be 2.3183, but it is good enough). The use of such tables, then, is the way in which irrational powers and the logarithms of irrational numbers can actually be computed bases other than the binomials.

20-5 Complex numbers

Now it turns out that after all that work we cannot solve every equation! For example, what's the square root of -1 ? Suppose we have to find $x^2 = -1$. The equation $x^2 = -1$ is forced to be irrational, otherwise that we have discussed, so x^2 , is equal to -1 . So we again have to go outside our numbers to a new number. I.e., as suggested, there's specific solution to $x^2 = -1$ is called *imagination*, we shall call it i ; it has the property, by definition, that its square is -1 . That is about all we can say; to say anything else, there is more than one root of the equation $x^2 = -1$. Surprised? Well, our teacher could say, "No, I'm the -1 , why is minus your i ?" Let just go ahead and say that by definition that i does it, $i^2 = -1$, now let's try another very equation we can write a basically form of the equation $x^2 = -1$ is changed everywhere. This is called taking the complex conjugate. Now we are going to make up numbers by adding a complex part, and multiplying it by numbers and adding other numbers, and so on, according to all of our rules. In this way we find that numbers $x + yi$ can be written in the form $p + qi$, where p and q are what we call real numbers, i.e., the numbers we were talking about in the beginning. The number i is called the unit imaginary number. Any real multiple of i is called pure imaginary. The most general number, as it is at the bottom of the line, is called a complex number. Things may get very worse if, for instance, we multiply two such numbers, let us say $(x + yi)(u + vi)$. Then

Using the rules, we get

$$\begin{aligned} (r - s)(r + s) &= r^2 - s^2 + 2rs - 2rs \\ &= r^2 - s^2 + 2rs - 2rs \\ &= r^2 - s^2 = rs - rs, \end{aligned} \quad (32.4)$$

since $r^2 - s^2 = 1$. Therefore all the numbers that now belong to the ring (32.3) have this most convenient form.

"Now you say, "This can go no further. We have defined powers of logarithms and all the rest, and when we are finished, somebody else will come along with another equation, which can not be solved, like $x^2 + 1x^2 = -x$." Then we have to generalize all over again." But it turns out that one has one more or less, not the square root of -1, every algebraic equation can be solved! This is a far more fine. What we must leave to the Mathematics Department to prove. The proofs are very long, it is, and very necessarily, but certainly not difficult. In fact, the most obvious supposition is that we are going to have to invent again and again one again. But the easiest method of all is that we do not. This is the last situation. After this invention of complex numbers, we find that the rules still work with computers, and we are finished inventing new things. When we find the same idea general for any complex number, we can solve any equation and do it algebraically, in terms of a finite number of these symbols. We do not find any new symbols. The logarithm of a, for example, has a definite result, it is not curved in, nor is it something. We will discuss that now.

We have already discussed multiplication, and addition is easy; if we have two complex numbers, $(x - iy) + (j - iz)$, we simply $\rightarrow (j - i) + (y - z)$. Now we can add and multiply complex numbers. But the real problem, however, is to compute complex powers of complex numbers. It turns out that the problem is actually no more difficult than computing complex powers of real numbers. So let us concentrate now on the problem of calculating 10^a to a complex power, not just at integral powers, but $10^{0.1}$ etc. Of course, we may also times use our rule (32.1) and (32.2). Thus

$$10^{x+iy} = (10^x)^i e^{iy}. \quad (32.5)$$

But if we already know how to compute 10^x we can always multiply anything by anything else; therefore the problem is to determine $i10^x$. Let us call some complex number, $r = iy$. To obtain, however, look at y . Now if

$$10^y = x + iy,$$

then the complex conjugate of this equation must also be true, so that

$$10^{-y} = x - iy.$$

(This we see that we can, sketchily, a number of things without actually computing anything, by using our rules.) The deduction which is resulting using by multiplying these together:

$$10^y 10^{-y} = (x + iy)(x - iy) = x^2 + y^2. \quad (32.6)$$

Thus if we find x , we have y .

Now the problem is how to compute 10^x to an imaginary power. What grade is i in $i10^x$? We may work out our ratio until we can go no further, but here is a reasonable grade. If we can compute it for any particular x , we can get it for the rest. If we know 10^x for arguments x and $x + \Delta$, we want x for twice that x , we can double the number, and so on. But how can we find 10^x for even the special value of $\pi/2$? To do so we shall make one additional assumption, which is not specific to the economy of all the other rules, but which leads to reasonable results and permits us to make progress. When the power is small, we shall suppose that the "true" $10^x = 1 + 2.3025x$ is right, as a generalization of the only "true" a , for complex numbers. Therefore, we begin, with the assumption that this a is true in general, and this tells us that $10^x = 1 + 2.3025 \cdot x$, for $x \ll 1$. So we assume that if x is very small, any part in 10^x , we have a rather good approximation to 10^x .

Now we make a table by which we can compute all the imaginary powers of 10, that is, complex $e^{i\theta}$. It is shown as follows. The first power we want is the 1/1000 power, which as you know is very nearly $1 - i \cdot 0.0001000$. Thus we start with

$$10^{-0.001} = 1.0000 + 0.0003489i, \quad (22.3)$$

and if we keep on multiplying by itself, we can get to a higher imaginary power. In fact, we just reverse the procedure we used in making our logarithmic and exponential law square, 4th power, 8th power, etc., of (22.7), and thus build up the values shown in Table 22-3. We notice an interesting thing, that the numbers are positive at first, but then owing to $i^2 = -1$, they become negative. We shall look it up. That is how we get to a moment. For first, we may be curious to find for what number θ the real part of 10^θ is zero. The θ -value should be i , and we would have $10^i = e^{i\theta} = \cos \theta + i \sin \theta$. As an example, if we want to be able just as we calculated before, let us now use Table 22-3 to find $\log_{10} i$.

What is i ? The number i is Table 22-3 does we have to multiply together to get a pure imaginary result? After a little trial and error, we discover that to recover a 10 times, it is best to multiply "312" by "123." This gives 0.12056 + 0.993486. Then we discover that we should multiply this by a number whose imaginary part is about equal to the size of the real part we are trying to remove. Thus we choose "74" since 74^2 is 0.14848, since that is closest to 0.12056. This then gives $-0.01163 + 0.98907$. Now we have $e^{i\theta} = 1$, and must divide by $0.98907 - 0.01163i$. How do we do this? By inverting the sign of i and multiplying by $0.98907 + 0.01163i$, which leaves $i^2 = -1$. Continuing in this way, we find that the extra power to which 10 must be raised to give i is 0.512 + 138 = 64 = $2 + 6(20) + 100$, or 693,200/1000. If we take 100 in that power, we come up with $e^{i\theta} = 0.98907$.

22-4 Imaginary exponents

To further investigate the subject of taking complex imaginary powers, let us look at the powers of 10 taking various exponents, not doubling the power each time, in order to follow Table 22-3 better and to see what happens to those minus signs. This is shown in Table 22-4, in which we take $10^{i\theta}$, and just keep multiplying. We see that a decrease passes through zero, owing to $\theta = 0$ if we could just lie between $\theta = 10$ and $\theta = 11$ it would obviously swing to -10 and swing back. The value is going back and forth.

In Fig. 22-1 the dots represent the numbers cast up in Table 22-1, and the lines are just drawn to help you visually. You see that the numbers are not oscillating; 10^{iθ} represents itself, it is a periodic thing, and so even i is very analogous to ω , because if a certain power is k then the fourth power of that would be i^4 squared. It would be plus one, and therefore, since $i^4 = 1$ is equal to by taking the fourth power we discover that $10^{i4\theta}$ is equal to $+1$. Therefore, if we wanted $10^{i20\theta}$, for instance, we could write $+1 \cdot 10^{i2\theta} \cdot 10^{i18\theta}$. In other words, it has a period, it repeats. Of course we recognize what a sine wave looks like. They look like the sine and cosine, and we shall see them, but while the algebraic sine and also basic cosine. However, instead of using the base 10, we shall put them into our natural base, which only changes the horizontal scale; so we denote $e^{i\theta}$ by c , and write $10^{i\theta} = c^{\theta}$, where c is a real number. Now $c^0 = 1$, up and down, with this as the general course of c plus i times, along the circumference of c . Thus

$$c^{\theta} = \cos \theta + i \sin \theta. \quad (22.6)$$

What are the properties of $\cos \theta$ and $\sin \theta$? First, we know, for instance, that $x^2 + y^2 = 1$, but it is also proved that $\cos^2 \theta + \sin^2 \theta = 1$ just as for the base 10. Therefore $\cos^2 \theta + \sin^2 \theta = 1$. We also know that $\cos \theta \sin \theta = \sin \theta \cos \theta = 0$ and the sine and cosine are nearly 1, and $\sin \theta$ is nearly θ and $\cos \theta$ is $1 - \theta$. But all of our previous properties of these trigonometric functions, when carried from being imaginary numbers, are the same as the new ones of this situation.

Table 22-3

Successive Powers of

$$10^{-0.001} = 1 - 0.0003489i$$

Power i	$10^{i\theta}, \theta$	10^θ
0.001	1	1.00000 + 0.00000i*
0.012	2	1.00000 + 0.000489i
0.025	4	0.99999 - 0.000500i
0.037	8	0.99999 - 0.000500i
0.048	16	0.99999 - 0.000500i
0.062	32	0.99999 - 0.000500i
0.075	64	0.99999 - 0.000500i
0.089	128	0.99999 - 0.000500i
0.102	256	0.99999 - 0.000500i
0.115	512	0.99999 + 0.000500i
0.128	1024	0.99999 + 0.000500i

* Should be 0.99999

Table 22-4

Successive Powers of $10^{i\theta}$

θ	power i	$10^{i\theta}$
0	0	1.00000 + 0.00000i
1	1	0.99999 - 0.000489i
2	2	0.99999 - 0.000978i
3	3	0.99999 - 0.001467i
4	4	0.99999 - 0.001956i
5	5	0.99999 - 0.002445i
6	6	0.99999 - 0.002934i
7	7	0.99999 - 0.003423i
8	8	0.99999 - 0.003912i
9	9	0.99999 - 0.004399i
10	10	0.99999 - 0.004885i
11	11	0.99999 - 0.005371i
12	12	0.99999 - 0.005856i
13	13	0.99999 - 0.006341i
14	14	0.99999 - 0.006825i
15	15	0.99999 - 0.007309i
16	16	0.99999 - 0.007793i
17	17	0.99999 - 0.008277i
18	18	0.99999 - 0.008761i
19	19	0.99999 - 0.009245i
20	20	0.99999 - 0.009729i
21	21	0.99999 - 0.010213i
22	22	0.99999 - 0.010697i
23	23	0.99999 - 0.011181i

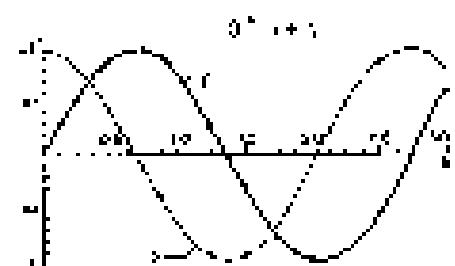


Figure 22-1

Is this good enough? Let us find out to what power we need to e . What \rightarrow the logarithm of the base e ? We stacked it, as before. In the case 10 it was 2.3026, but when we change our logarithm's base to e , we have to multiply by 2.3026, and if we divided it comes out 1.3863. So this will be called "algebraic π ." But we see, π differs from the regular π ' by only one place at the last point, and that, of course, \rightarrow the result of others at the arithmetic. So we have created two new functions in a purely algebraic manner, trigonine and the one which refers to algebra and only to circles. We wake up at the end to discover the very identities that are natural to geometry. So there is a connection, you might say between algebra and geometry.

We summarize with this, the sine and cosine formula in mathematics:

$$e^{it} = \cos t + i \sin t \quad (2.9)$$

This is our new:

We now close the geometry to the algebra by representing complex numbers as points; the horizontal position of a point x is the vertical position of a point $i y$ (Fig. 22-2). We represent every complex number, $x + iy$, then if the radial distance to the point of origin x and the angle is called t , the algebraic law is that $x + iy$ is written in the form e^{it} , where the geometrical relationship between x , y , r , and t are as shown. This, then, is the unification of algebra and geometry.

When we begin this chapter, we did only with the basic notion of integers and counting, we had little idea of the power of the processes of the method and generalization. Using the art of algebraic "tricks," or properties of numbers, Eq. (2.8), and the definitions of inverse operations (2.6), we can handle here, otherwise, to our absolute not only numbers but also all things like ratios of lengths, powers, and trigonometric functions (for these are with the simple powers of the numbers (2.6) all unity by connecting to successive higher roots or root).

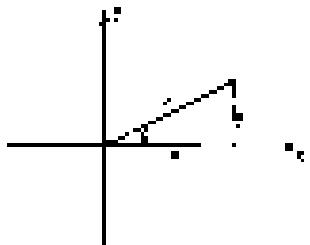


Fig. 22-2 $x + iy = re^{it}$.

Harmonics

13-1 Complex numbers and harmonic motion

In the present chapter we shall continue our discussion of the harmonic oscillator and, in particular, of forced harmonic oscillation, using a new technique called analysis. In the preceding chapter we introduced the idea of complex numbers, which have real and imaginary parts and which can be represented on a diagram in which the horizontal axis is the imaginary part and the vertical axis represents the real part. If a is a complex number, we may write it as $a = a_r + ia_i$, with the subscript r telling us the real part of a and the subscript i telling the imaginary part of a . Referring to Fig. 27, we see that we may also write a complex number $a = x + iy$ in the form $x + iy = re^{i\theta}$, where $r^2 = x^2 + y^2$ ($r = |a|$) and $\theta = \arg(a)$. (The complex conjugate of a , written a^* , is obtained by reversing the sign of $i\theta$.) So we shall represent a complex number a by two figures: a real axis (the imaginary part), a magnitude r and a phase angle θ , described. Here x and y and r are clearly *real* and $\sin \theta$ and $\cos \theta$ give a complex number $x + iy$, i.e. $\sqrt{r^2} = j\sin \theta$ and $a = jy/r$, the ratio of the imaginary to the real part.

We are going to apply complex numbers to the analysis of physical phenomena by the following rule. We have complex to change the basis set; the real values may have a driving force which is a certain constant times cos ωt . Now note it since, $F = F_r$, we can write the solution as the real part of a complex number $F = F_r e^{i\omega t}$ because $e^{i\omega t} = \cos \omega t + j \sin \omega t$. The reason we do this is that it is easier to work with exponential functions than with a cosine. So the whole trick is to represent all oscillatory functions as the real part of certain complex functions. The complex number F that we have so defined is not a real physical force, because it is in physics, in reality, complex, i.e. it has both its imaginary part, only a real part. We shall, however, speak of the "force" F , just, for convenience, the actual force is the *real part* of that expression.

To take another example, suppose we want to represent a force which is a cosine wave that is out of phase with a delayed phase ϕ . This, after all, would be the real part of $F_r e^{i(\omega t - \phi)}$, but exponents being what they are, we can write $F_r e^{i(\omega t - \phi)} = F_r e^{i\omega t} e^{-i\phi}$. Thus we see that the only law of causality is that ϕ is the sum of sines and cosines; this is the reason we choose to use complex numbers. We shall often write

$$F = F_r e^{i(\omega t - \phi)} = F_r e^{i\omega t}. \quad (13.1)$$

We write a little $e^{i(\omega t - \phi)}$ over the F to remind ourselves that this quantity is a complex number; here the number is

$$e^{i\omega t} = \rho e^{i\omega t}$$

Now let us solve an equation, using complex numbers, to see whether we can work out a problem for some real case. For example, let us try to solve

$$\frac{d^2x}{dt^2} + kx = F_r \cos \omega t, \quad (13.2)$$

where F is the force which drives the oscillator and x is the displacement. Now, about the right way seems to suggest that x and F are actually complex numbers, for a mathematical purpose only. That is to say, x has a real part and an imaginary part, and F has a real part and an imaginary part, too!

13-1 Complex numbers and harmonic motion

13-2 The forced oscillator with damping

13-3 Electrical resonance

13-4 Resonance in nature

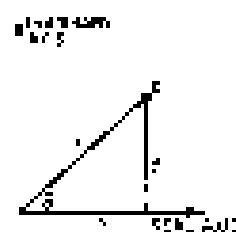


Fig. 27. A complex number may be represented by a point in the "complex plane".

Now if we had a solution of (21.2) with complex numbers, and substituted the complex numbers in the equation, we would get

$$\frac{d^2x_1}{dt^2} + \frac{k(x_1 - x_2)}{m} = \frac{\alpha_r + i\omega}{m}$$

or

$$\frac{d^2x_1}{dt^2} + \frac{kx_1}{m} + \left(\frac{d^2x_2}{dt^2} + \frac{kx_2}{m} \right) = \frac{\alpha_r + i\omega}{m}$$

Now, since if two complex numbers are equal, their real parts must be equal and their complex parts must be equal, we deduce that the real part of x_1 satisfies the equation with the real part of the force. We may emphasize, however, that this assumption introduces real x_1 and its imaginary part is not zero in general, but is only zero for equations which are linear, that is, for equations in which ω appears in every term only in the last power or the tenth power. For instance, if there were in the equation a term kx^2 then when we substitute $x_1 = A_r e^{i\omega t}$, we can't get $x_1 = kx^2$, but when separated into real and imaginary parts this would yield $kA_r^2 - \omega^2$ as the real part and $2kA_r \omega$ as the imaginary part. So we see that the real part of the equation which does involve just ω^2 , but ω not ω^2 . In this case we get a different result so that the two we will be able to solve, with x_1 , the complex x_2 being the independent variable that the two we will be able to solve, with x_2 , the complex x_1 being the independent variable.

Let us now try our new method for the problem of the damped mass-spring, that we already know how to solve. We want to solve Eq. (21.2) now, but we say that we are going to do it in a new way.

$$\frac{d^2x}{dt^2} + \frac{kx}{m} = \frac{F_0 e^{i\omega t}}{m} \quad (21.3)$$

where $F_0 e^{i\omega t}$ is complex number. Of course x will also be complex, but remember the rule: take the real part to find x_1 , what is really going to be easy to solve (21.3) is the forced solution; we shall discuss other solutions later. The function x then has the same frequency as the applied force, and has same amplitude of oscillation and same phase, and so it can be represented also by some complex number A whose magnitude represents the ratio of x to x_1 whose phase represents the difference in the same way as for the force. Now a wonderful feature of an exponential function is that $d(e^{i\omega t})/dt = i\omega e^{i\omega t}$. When we differentiate an exponential function, we bring down the exponent as a simple multiple. The second derivative does the same thing, if (dx/dt) is multiplied by $i\omega$, then it is very simple to prove immediately, by inspection, that the equation is true if every time we see a differentiation we simply multiply by $i\omega$. Differentiation is now as easy as multiplication. The idea of using exponentials in such a Tertius' logarithms is almost as great as the invention of logarithms, in which multiplication is replaced by addition. Non-differentiation is replaced by multiplication. This we inquire here:

$$(d^2x/dt^2) = (d^2x_1/dt^2) - i\omega \dot{x}_1 \quad (21.4)$$

We have introduced the common factor $e^{i\omega t}$. See how simple it is! Differential equations are immediately converted, by sight, into more algebraic eq's where you only have the solution by sight, that,

$$x = \frac{F_0/m}{(k/m) - \omega^2} e^{i\omega t}$$

since $(i\omega)^2 = -\omega^2$. This may be slightly simplified by substituting $k/m = \omega_0^2$ which gives

$$x = F_0/m \cdot \frac{1}{\omega_0^2 - \omega^2} e^{i\omega t} \quad (21.5)$$

This, of course, is the situation which before for since $m(\ddot{x} - \omega^2 x) = 0$ is a real number, the phase angles of \dot{x} and of x are the same (as we have seen $\omega^2 > \omega_0^2$ as mentioned previously). The angle $\angle(\dot{x})$, which requires how fast x oscillates, is related to the angle of the F by the factor $\angle(F_0 e^{i\omega t}) = \omega^2 t$, and this factor becomes $\omega_0^2 t$.

encounters which ω is nearly equal to ω_0 . So we get a very strong resonance effect. At such the right frequency ω_0 if we add a pendulum at the end of a string and shake it at just the right frequency, we can make it swing very high!

23.3 The forced oscillator with damping

Now, let's look at how we analyze oscillatory motion with the more elegant mathematical technique. But the elegance of the technique is not at all exhibited in such a problem that can be solved easily by other methods. In our problem, often one appears to be more difficult problems. Let us consider some another, more difficult problem, which is common and a relatively realistic situation. In the previous unit (Equation 23.9), we saw that if the frequency ω were exactly equal to ω_0 , we would have no damped response. Actually, of course, no such infinite response occurs because some other damping-like friction, which we have set for ignored, limits the response. So we therefore set in Eq. 23.9 a friction term.

Obviously such a problem is very difficult because of the oscillator and complexity of the frictional term. That is... however, every circumstance in which the frictional force is proportional to the speed with which an object moves, an example of such friction is the friction for slow motion of an object in oil or a thick liquid. There is no force when it is not moving $v < 0$, but the faster it moves, the harder the oil tries to go past the object, and the greater is the resistance. So we will assume that the F_f is in addition to the term in (23.2), we have some proportional proportionality to the velocity: $F_f = -\gamma v$, it may be convenient, in our mathematical analysis, to write the constant γ from Eq. 23.2 to simplify the equation a little. This is just the same trick we use with ϵ when we replace it by $m\ddot{x}$ to simplify the physics. This circumstance will be

$$m(d^2x/dt^2) + \gamma dx/dt + kx = F \quad (23.6)$$

or, setting $\omega_0^2 = k/m$ and $\gamma = m\beta/\omega_0$ and canceling out the mass m ,

$$(d^2x/dt^2) + \gamma dx/dt + \omega_0^2 x = F/m. \quad (23.6a)$$

Now we're at the equation in the most convenient form possible. If γ is very small, and represents very little friction; if γ is very large, there is a tremendous amount of friction. How does $x(t)$ in this linear differential equation? Suppose that the driving force is constant in time ($F = A$; we could pick this from (23.2) and try to solve it, but we want instead solve it by our new method). Thus we write F as $F = A e^{i\omega t}$, A is F and ω is the real part of ω , and substitute this into (23.6a). It is not even necessary to do the actual substituting, for we can see at a glance that the equation would become

$$(i\omega)^2 x + i\gamma \omega x - \omega_0^2 x = (F/m)e^{i\omega t}. \quad (23.7)$$

As a matter of fact, if we tried to solve Eq. 23.6a by our old straightforward way, we would really complicate the analysis of the "complex" method. If we divide by $e^{i\omega t}$ on both sides, then we can obtain the necessary function, given force F ; it is

$$x = (F/m)(\phi - \omega^2 - i\gamma\omega) \quad (23.8)$$

This result x is given by F times a certain factor. There is no additional factor for the factor, we've already taken care of it, b . We may call B for "resonant process":

$$B = \frac{m\omega_0^2}{m\omega_0^2 - \omega^2 - i\gamma\omega}$$

and

$$\phi = F/m \quad (23.9)$$

(Although the letters ϕ and ω_0 are in very common use, this B has no particular meaning.) This factor B can either be written as $B = \phi$, or as a certain \rightarrow quantity \times times ω^2 . If it is written as a certain magnitude times ω^2 , let us see what it means

Now $\hat{F} = F_0 e^{i\omega t}$, and the real part of F is the real part of $F_0 e^{i\omega t}$. That is, $F_0 \cos(\omega t - \phi)$. Now, Eq. (23.9) tells us that ϕ is equal to $\pi/2 - \theta$, making $R = \rho^2$ as another name for R , we get

$$R = \rho^2 = \omega^2 F_0^2 / (\omega^2 + \gamma^2).$$

Finally, going even further back, we see that the phasor \vec{x} , which is the real part of the complex \vec{x} , is equal to the real part of $\vec{F}_0 e^{i\omega t}$. But ρ and θ are real, and the real part of $e^{i\omega t}$ is simply $\cos(\omega t + \theta)$. Thus,

$$\vec{x} = \rho \vec{F}_0 \cos(\omega t + \theta) = R \vec{F}_0. \quad (23.10)$$

This tells us that the amplitude of the response is the magnitude of the force F multiplied by a certain magnification factor, ρ . This gives us the "resonance" of oscillation. It also tells us, however, that \vec{x} is not oscillating in phase with the force, which has the phase θ , but is shifted by $\pi/2$ extra current θ . Therefore ρ and θ represent ratios of the response and the phase shift of the response.

Now let's work out what ρ is. If we have a complex number, the square of the magnitude is equal to the number times its complex conjugate; thus,

$$\begin{aligned} \rho^2 &= \frac{1}{m^2(\omega^2 - \omega^2 - 2\omega i\omega\gamma) + \omega^2 + \gamma^2} \\ &= \frac{1}{m^2((\omega^2 - \omega^2)^2 + \gamma^2\omega^2)}. \end{aligned} \quad (23.11)$$

In addition, the phase-angle θ is easily found, for if we write

$$1/\rho = 1/\rho e^{i\theta} = (1/\rho) e^{-i\theta} = m\omega i(-\omega^2 + \gamma^2\omega),$$

we see that

$$\tan \theta = -m\omega(\omega^2 - \gamma^2). \quad (23.12)$$

If ω equals ω_0 , then $\tan \theta = -\infty$. A negative value for θ results for all ω , and this corresponds to the displacement x lagging the force F .

Figure 23.2 shows how ρ^2 varies as a function of frequency (ω^2) is physically more interesting than ρ , because it is proportional to the square of the amplitude. Of course, this is the curve that is developed in the oscillator by the forces. We see that if ω^2 is very small, then $1/\rho(\omega^2 - \omega^2)^2$ is the most important term, and the response lies high up toward infinity when ω equals ω_0 . Now, the "infinity" is not actually infinite because if $\omega = \omega_0$, then $1/\rho^2\omega^2$ is still there. The shape of ρ^2 is shown in Fig. 23.1.

In certain circumstances we get a slightly different form to curve (23.12), also called a "resonance" formula, and one in ρ^2 linking ω and ω_0 . Representing a different phenomenon, but it does not. The reason is that if ω is very small, the most important part of the curve is $1/\rho(\omega^2 - \omega_0^2)$, and we may replace $(\omega^2 - \omega_0^2)$ by an approximate formula which is very accurate if ω is small and very near ω_0 . Since $\omega_0^2 = \omega_0^2(\omega_0 - \omega_0 + \omega)$, this is nearly the same as $1/\omega_0(\omega_0 - \omega)$ and the two terms are nearly the same as each other. Using these in (23.12), we see that $1/\rho^2 = \omega^2 - \omega_0^2/(\omega_0^2 - \omega^2)$, or that

$$\rho = \sqrt{2\omega_0/(\omega_0^2 - \omega^2 - 2\omega_0^2)} \quad \text{if } \omega \ll \omega_0 \quad \text{and} \quad \omega \neq \omega_0. \quad (23.13)$$

It is easy to find the corresponding form θ for ρ^2 . It is

$$\theta^2 = 1/(2\omega_0^2)(\omega_0^2 - \omega^2 + \gamma^2/2).$$

We shall leave it as an exercise to show the following: if we call the maximum height of the curve of ρ^2 as ω_0 units, and we ask for the width $\Delta\omega$ of the curve, at one-half the maximum height, i.e., at half the maximum height of the curve is $\Delta\omega = \gamma$, supposing that γ is small. The resonance is sharper and sharper as the friction factors are made smaller and smaller.

As another measure of the width, some people use a quantity Ω which is defined as $\Omega = \omega_0/\omega$. The higher is the resonance, the higher the Ω ; $\Omega = 100$ means a resonance whose width is only 10% of the frequency scale. The Ω of the capacitor shown in Fig. 29-2 is 1.

The importance of the resonance phenomenon is that it occurs in many other circumstances, and on the rest of this chapter will describe some of these other circumstances.

29-3 Electrical resonance

The simplest and most widespread application of resonance is inductively. In the electrical work there are a number of objects which can be connected in simple electric circuits. These parts are circuit elements, as they are often called, and of three main types, although each one has a little bit of the other two mixed in. Before discussing them in greater detail, let us note that the weight of an inductor usually depends mainly on the end of a spring, only an approximation. All the mass is not necessarily at the "mass"; some of the mass is in the helical coil spring. Similarly, all of the spring is not at the "spring"; the mass itself has a little elasticity, and so though it may appear as if it is not absolutely rigid, and as it goes up and down, it never goes so slightly under the action of the spring pullings. The same thing is true in electricity. There is an experimental, though we can't jump things into "circuit elements" which are assumed to have particular characteristics. To set the proper time to discuss the experimental part, we shall simply assume that it is true in the circuit elements.

The three main kinds of circuit elements are the following. The first is called a capacitor (Fig. 29-4); an example is two plane metallic plates spaced a way or d , distance apart, by an insulating material. When the plates are charged there is a certain voltage difference between them; a certain difference in potential between them. The same difference of potential appears between the terminals alone, because if there were any difference along the connecting wire, electricity would flow right away. So there is a certain voltage difference V between the plates. If there is a certain electric charge q and $-q$ on them, respectively. Between the plates there will be a certain electric field; which we can find a formula. In it (Chap. 13 and 14):

$$V = qd/\epsilon_0 A, \quad (29-1)$$

where d is the spacing and A is the area of the plates. Note that the potential difference is a linear function of the charge. If we do not have insulated wires, but insulated electrons which are of any sort of use, the difference in potential is still directly proportional to the charge, but the constant of proportionality may not be so easy to compute. However, all we need to know is that the potential difference per unit capacity is proportional to the charge. If $C = q/V$, the proportionality constant is $1/C$, where C is the capacity of the object.

The second kind of circuit element is called a resistor; it offers resistance to the flow of electrical current. It turns out that metallic wires and many other substances resist the flow of electricity in this manner; a resistor is a voltage difference across a piece of something where there exists an electric current. And this is proportional to the electric voltage difference:

$$V = IR = R \text{ (current).} \quad (29-2)$$

The proportionality constant is called the resistance R . This relationship is already familiar to you; it is Ohm's law.

If we think of the charge and a capacitor as being analogous to the displacement x of a harmonic system, we see that the current, $I = dq/dt$, is analogous to velocity, (dx/dt) is analogous to a spring constant k , and R is analogous to the resistive coefficient γ . Now it is very interesting that there exists another circuit element which is the analog of m . This is the coil which builds up a magnetic field W in association to a direct current i . A changing magnetic field builds up the voltage V which is proportional to di/dt (this is how a transformer works).

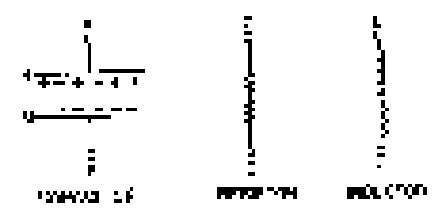


Fig. 29-4. The three passive circuit elements.

(21) The magnetic field is proportional to current, and the induced voltage (so-called inductance) is proportional to the rate of change of current:

$$V = L \frac{di/dt}{dt} = L \frac{di}{dt} \quad (21.16)$$

Let us return to the ω -circuit, and let's always assume a sinusoidal voltage input:



Fig. 21-5 An oscillatory electrical circuit with resistance, inductance, and capacitance.

Suppose we have a circuit in which we have connected the three circuit elements in series (Fig. 21-5); then the voltage across the whole thing (from 1 to 2) is the work done in carrying charge through, and it consists of the sum of several pieces: across the inductor, $V_L = L \frac{di^2/dt^2}{dt}$; across the resistance, $V_R = R \frac{di}{dt}$; across the capacitor, $V_C = q/C$. Let our initial guess be the apedical equation:

$$L \frac{di^2/dt^2}{dt} + R \frac{di}{dt} - q/C = P(t) \quad (21.17)$$

Now we see that this equation is exactly the same as the second real equation (21.6), and it must also be solved in exactly the same manner. We suppose that $P(t)$ is oscillatory; we are dealing, then, with a generator with a pure sine wave oscillation. Then we can write our Eq. (21.7) as a complex \tilde{P} with the understanding that it must be ultimately multiplied by $e^{i\omega t}$, and the real part taken, in order to get the true P . Likewise, the charge q must be rearranged, and then integrated over the same amount of time T as in Eq. (21.8); we write the corresponding equation, integrated over the time of ωT (that is, T divided by ω). Thus Eq. (21.17) translates to

$$\left[L(\omega)^2 + R\omega \right] + \frac{1}{C} \tilde{q} = \tilde{P}$$

or

$$\tilde{q} = \frac{\tilde{P}}{L(\omega)^2 + R\omega \pm \frac{1}{C}}$$

which we can write in the form

$$q = P(t)(\omega^2 - x^2 + i\omega x) \quad (21.18)$$

where $x^2 = 1/C$ and $x = \omega/\sqrt{L}$ is exactly the same denominator as we had in the ω -circuit, now with exactly the same "constant product" $1/C$ corresponding cross between the ω -circuit and node \tilde{q} , the next entries in Table 21-1.

Table 21-1

General nomenclature	Mechanical analog	Electrical property
independent variable	time (t)	time (t)
dependent variable	position (x)	charge (q)
force	mass m	inductance (L)
displacement	displacement (x)	resistance ($R = \Omega$)
stress	surface (σ)	capacitance ($C = \text{farad}$)
constant frequency	$\omega = 2\pi f$	$\omega^2 = L/C$
period	$T = 2\pi\sqrt{m/L}$	$T = 2\pi\sqrt{C/R}$
degree of merit	$R = \text{const}$	$G = \text{const}$

We must use ω in small harmonic motion. In the electrical circuit, a different notation is used. (From one field to another, the analysis is not really very different, but the way of writing the equations is often different.) For d , it is customarily used instead of ω in electrical engineering, in rad/sec (rps). (After all, ω must be the angular f .) Also, the engineer would rather have a relationship between P and \dot{q} (the integral of \tilde{q} over time), just because they are more used to it that way. Thus, since $f = \omega/2\pi = \omega/2\pi$, we can just substitute f for ω and \dot{q} :

$$V = Q\omega^2 - R\dot{q} + 1/2\omega C\dot{q}^2 = 2f. \quad (21.19)$$

Another way is to modify Eq. (23.17), so that it accommodates one other sees it in this way:

$$A \sin(\omega t + \phi) = (1/\rho) \int J dt = 100 \quad (23.20)$$

At this rate, we have an alternating 100 ohm resistance if and current J which is just the same as (23.12) except divided by ρ , and that satisfies Eq. (23.14). The quantity $\omega + \phi = \psi$ is a complex number, and is used in quantum mechanical calculations that it has a name, was called the complex propagation ψ . Thus we can write $\psi = \phi$. The reason that engineers like to do this is that they learned something when they were young: "as far as business when they only have direct resistance and not. But they have become more educated and know no circuits, so they want them to be in book the same". Thus this writing $\psi = \phi$, the only difference being that the resistance is now given by a more complicated thing, imaginary, or $i\omega$. So that says that they cannot use what everyone else in the world uses for imaginary numbers, they have to use $i\omega$! But Cut, the ammeter said they did not invent so that rule long ago when ψ was ϕ ? (they may get into trouble when they talk about the real resistance for which they also use). The difficulties of science are so large exactly the different ways of doing the same, and all the other difficulties we shall have are due to man, not by nature.)

23-4 Resonance in nature

Although we have imagined the electrical case, in detail, we could also bring up some other cases in safety texts, and such exactly how the resonance argument is important. There are many of consequences in nature in which something is "resonantly" excited in which the resonance phenomenon occurs. We will look, in an earlier chapter, let us say demonstrate it. If we walk around our store, putting books off the shelves and simply shaking them up, there is an example of a grave consequence to us (Fig. 23-1) and comes from the same argument, what does. Only first we demonstrate, however, organized in taking the easiest possible example, it takes only five or six books arranged quite a series of shelves, and which are oscillating.

The first we can do from mechanics, the first one a large effect the atmosphere or the whole earth. In the atmosphere, which we suppose surrounds the earth evenly in all directions, is pulled to one side by the wind or, rather, squared pressure from a trouble here and there, and then if you let it go, it would probably go up and down; this is an oscillator. This oscillator is driven by the moon, which is a basically rotating galaxy in the earth, and the component of the force, say to every direction, has a cosine component, and as the response of the atmosphere to the tidal pull of the moon is that of an oscillator. The expected response of the atmosphere is shown in Fig. 23-6. A second feature is another, theoretical curve under discussion, is the peak from which rises take out of center. Now one might think, if we only have one point of this resonance curve, since we only have the one frequency, corresponding to the rotation of the earth under the moon, which occurs at a period of 12.42 hours = 12 hours for the earth, then the two don't fit together, plus a little, and to take the moon's lagging behind. But from the rest of the atmospheric tides, just from the papers themselves of today we can get ω_1 , ω_2 , and ω_3 . From these we can get ψ_1 , ψ_2 , and ψ_3 thus draw the entire curve! This is an example of very good science. From two nodes we obtain two numbers, and from these two numbers we draw a curve that agrees with ψ measured year. I made the very point that determines the curve. It is of the use unless we can measure everything else, and in the case of geostrophic, that is, Chen et al., result. But in this particular case there is another thing which we can give, however, namely that in the same tides as the tidal frequency ω_1 ; that is, if someone described the atmosphere as it would oscillate with the frequency ω_1 . Now there was just a large disturbance in 1851; the Krakatoa volcano exploded and had the island blown off, and it made such a large explosion, in fact so large that the period of oscillation of the atmosphere could be measured. It came out to $16\frac{1}{2}$ years. These obtained from

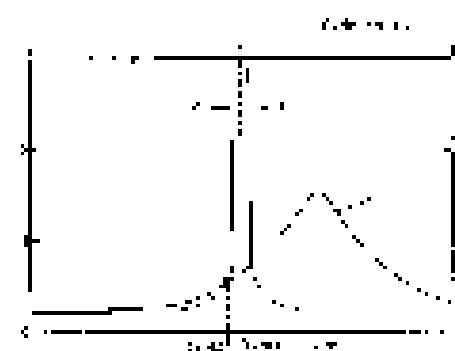


Fig. 23-6. Response of the atmosphere to a steady Earth-shaker. a is the required response of the atmospheric tide; a of gravitational origin: peak amplitude 130x1. It is derived from measured data (Hansen and others of Met Office, Work and MacDowell, "Rotation of the Earth," Cambridge University Press, 1940).

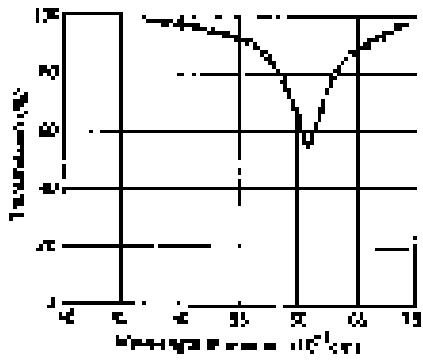


Fig. 23-7. Transmission of infrared radiation through a thin (0.17μ) sodium chloride film. [After R. S. Becker, J. Phys. B, 23, 723 (1922); Kittel, Introduction to Solid State Physics, Wiley, 1956.]

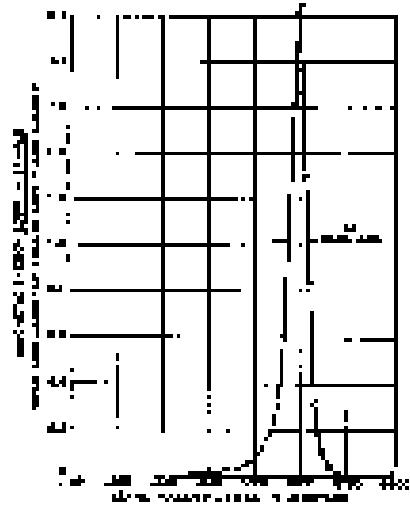


Fig. 23-8. Magnetic energy loss in para-xylylene organic compound as a function of applied magnetic field frequency. [Holden et al., Phys. Rev. 75, 161 (1949).]

Fig. 23-6 shows us 20 terms; and 20 minutes, so there we have at least one check on the validity of our understanding of the atmospheric waves.

Now we go to the next scale of mechanical oscillation. The time we make a sodium chloride crystal which has sodium ions and chlorine ions next to each other, as we described in an early chapter. These ions are constantly changing, alternately plus & minus. Now there is an interesting oscillation possible. Suppose the + would drift all the way to the right and all the negative charges to the left, and then go; they would then oscillate back and forth. As sodium, since against the calcium lattice. How can we now do such a thing? That is easy, for if we apply an electric field to the crystal, it will steal the plus sign one way and the minus charge the other way. So by using an external electric field we can produce net lattice oscillations. The frequency of the electric field needed is so high, however, that it corresponds to infrared radiation. So we try to find a resonance curve by measuring the absorption of infrared light by sodium chloride. Such a curve is shown in Fig. 23-7. The abscissa is not frequency, but is given in terms of wavelength, but that is just a technical matter, of course, since for a wave there is a definite relation between frequency and wavelength; or it is really a frequency scale, and a certain frequency corresponds to the length of a wave.

But what about the width? What determines the width? There are many cases in which the width, that is, how wide the curve is not really associated with ω_0 , that one would have thermally. There are two reasons why there can be a wider curve than the theoretical curve. If the atoms do not all have the same frequency, as might happen if the crystal were strained in certain regions, so that in these regions the oscillation frequency were slightly different than in other regions, then when we tune it many resonance curves may appear here as we apparently get a wider curve. The... for this of water is a very easy problem we cannot measure the frequency precisely enough. If we open the slit of the spectrometer fairly wide, so actually we don't see just one frequency, we actually had a certain range of frequencies, we may not have the resolving power needed to see a narrow curve. Otherwise, we cannot say whether the width in Fig. 23-7 is natural, or whether it is due to inhomogeneities in the crystal or to some width of the slit of the spectrometer.

Now we turn to a more esoteric example, and that is the swinging of a \leftrightarrow pendulum. If we take a magnet, with two south + north poles, in a constant magnetic field, the magnet will be pulled one way and the S end the other way, and this will in general be a torque on it, so it will vibrate about its equilibrium position, like a compass needle. However, the magnet is usually single-layered wires. These atoms here are in your memory, but the torque does not produce a simple motion in the direction of the field, but instead it has a, a precession. Now, look at it from the side, and one component is swinging, and we can detect or drive that swinging and measure the frequency. The curve in Fig. 23-8 represents a typical such resonance curve. What has been done here is slightly different, however. The frequency of the initial tick, that is used to drive this swinging, is always kept the same, while we would have expected that the increasing ω_0 would vary the amplitude of the curve. They could, as done, that way, but technically it was easier for them to keep the frequency ω_0 fixed and change the strength of the constant magnetic field, which corresponds to changing ω_0 in our example. They have plotted the resonance curve against ω_0 . Anyway, he is a typical resonance with a certain ω_0 and γ .

Now we go still further. Our next example has to do with a rotating atom. The motions of electrons and neutrons in nuclei are really in certain ways, and we can demonstrate this by the following experiment. We bombard a lithium atom with protons, and we discover that a certain rate for producing protons, occurring the at very low energy, has a sharp resonance. We note in Fig. 23-9, however, the following: from one reason, the horizontal scale is not a frequency, it is an energy. The reason is that in quantum mechanics, what we think of classically as the energy will turn out to be really related to a frequency of a wave on a string. When we calculate something which is simple, classical mechanics has to do with a frequency, we find that when we do quantum mechanics experiments with atomic

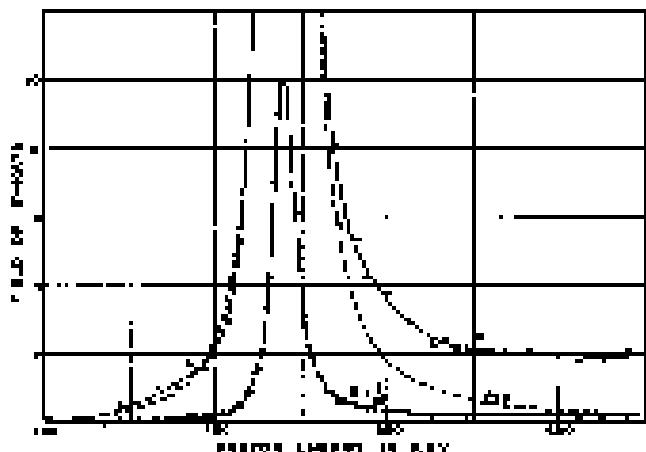


Fig. 23-9. The intensity of gamma radiation from lithium, in units of 10^{-3} , vs. energy of the bombarding proton. The dashed curve is a theoretical one calculated for a proton with no angular momentum $J = 0$. (Bonner and Evans, Phys. Rev. 73, 953 (1948))

more, we get the corresponding curve as a reaction scheme. In fact, this curve is a demonstration of this relationship, in a sense. It shows that if you plot and count how many days it costs to count, which of course they do.

Now we turn to another example which also involves a nuclear energy level, but with a much higher level here. The π^+ in Fig. 21-1 has a range of approximately $10,000$ cm when cold, while the π^+ $\rightarrow \pi^+ + \pi^-$ is approximately 10^{-2} cm, very cold; in other words, this has a β of 10^{11} . When the curve was measured it was the largest β of any reaction that had ever been measured. It was measured by Dr. M. W. Macmillan and G. C. Rieke of this subject group. The behavior of the curve is peculiar, because the technique for obtaining the slightly different intensities was to use the Doppler effect, by moving the source relative to the detector. One can see how delicate the experiment is when we realize that the speed involved is 10^8 centimeters per second! On the usual scale of the spectrum in frequency would correspond to a point about 10^{15} cm to the left—slightly off the paper!

Finally, if we look in an issue of the *Physical Review*, say that of January 1, 1962, will we find a resonance curve? Every book has a resonance curve, and Fig. 24-1 is the resonance curve for the one I am interested in now, which is very interesting. It is the resonance found in a certain reaction among strange particles, a reaction in which a K^- and a neutron interact. The reaction is denoted by saying how many of some kinds of particle is emitted, and depending on whether one how many come out, one gets different curves, but of the same shape and with the peak at the same energy. We thus believe that there is a resonance for a certain energy for each K^- meson. And presumably means that there is some kind of a cold, a condition corresponding to the resonance which can be obtained by scattering like a K^- and a proton. This is a new particle interaction. Today we do not know whether to call it a bump like this a "resonance" or simply a resonance. What there is is a very narrow resonance, which occurs at a very definite energy, just as though there were a particle of that energy present in nature. When the resonance is this, then we do not know whether to say there is a particle which does not live very long, or simply a condition of the reaction probability. In the second chapter, one point in particular the reader has when we start chapter two whether this resonance **was** not **absent**. So our next should now come still another particle in it!



Fig. 23-10. (Courtesy of Dr. R. M. Macmillan)



Fig. 23-11. Measurement representation of the cross section for the reaction (a) $K^- + p \rightarrow K^- + \pi^+ + \pi^-$ and (b) $K^- + p \rightarrow K^- + \pi^0 + \pi^0$. The lower curves (a) and (b) represent the measured resonance background, while the upper curves contain in addition the expected resonance. (Macmillan et al., Phys. Rev., Lett. 4, 28 (1962))

Transients

24-1 The energy of an oscillator

Although this chapter is entitled "transients," certain parts of it are, in a way, sort of the last chapter on forced oscillation. One of the features of forced oscillation which we have not yet discussed is the energy of the oscillation. Let us now consider total energy.

In a mechanical oscillator, how much kinetic energy is there? It is proportional to the square of the velocity. Now what about an important point? Consider an arbitrary quantity A , which may be the velocity or something else that we want to discuss. When we write $A = A_0 e^{i\omega t}$, A is complex - unless, the real and imaginary, A , the imaginary work is only the $\sin(\omega t)$ part; therefore if, for some reason, we want to take the square of A , it is not right to square the complex number and then take the real part because the real part of the square of a complex number is not just the square of the real part, but also involves the imaginary part. So when we write in that the energy we have to get away from the complex notation for A so it's better with the inner workings are:

Show this true physical A is the real part of $A_0 e^{i\omega t + \phi}$, that is, $A = A_0 \cos(\omega t + \phi)$, where A_0 , the complex number, is written as $A_0 e^{i\phi}$. Now the square of this real physical quantity is $A^2 = A_0^2 \cos^2(\omega t + \phi)$. The square of the quantity, then, goes up and down from a maximum to zero, like the square of the cosine. The square of the cosine has a maximum of 1 and a minimum of 0, and its average value is $1/2$.

In many circumstances we are not interested in the energy at any specific instant during the oscillation; for a single number of oscillations we usually want the average of A^2 , the mean of the square total over a period of time compared with the period of oscillation. In these circumstances, the average of the cosine signal may be real, so we have the following theorem: if A is represented by a complex number, then the mean of A^2 is equal to $|A|^2$. Now $|A|^2$ is the square of the magnitude of the complex A . (This can be written in many ways - some people like to write $|A|^2$ as $A \bar{A}$ with \bar{A} being A 's complex conjugate.) We shall prove this in several stages.

Now let us consider the energy in a forced oscillator. The equation for the forced oscillator is

$$m d^2x/dt^2 + 2m\alpha dx/dt + m\omega_0^2 x = F(t). \quad (24.1)$$

Now, provided, of course, $F(t)$ is a certain function of t . Now let us consider the situation: how much work is done by the outside force F ? The work done by the force per second, i.e., the power, is the force times the velocity. (We know that the differential work is $-F dx$ or $F dx$, and the power is dW/dt .) Thus

$$P = F \frac{dx}{dt} = m \left[\left(\frac{dx}{dt} \right) \left(\frac{d^2x}{dt^2} \right) + \omega_0^2 \left(\frac{dx}{dt} \right)^2 \right] = m \left(\frac{dx}{dt} \right)^2. \quad (24.2)$$

But the first two terms on the right can then be written as $m(d^2x/dt^2)(dx/dt)^2$; $m(dx/dt)^2$ is immediately verified by differentiating. That is to say, the term in brackets is a pure derivative of two terms that are easy to understand: one is the kinetic energy of motion, and the other is the potential energy of the spring. Let us call this quantity the stored energy; that is, the energy stored in the oscillator. Suppose that we want the average power over many cycles when the oscillator is being forced and has been running for a long time. In the long run, the stored

24-1 The energy of an oscillator

24-2 Damped oscillations

24-3 Mechanical transients

energy does not change—the derivative gives no average effect. In other words, if we average the power in the long run, all the energy disappears and ends up in the resistor as $\text{heat} = \text{power}^2/\text{resistance}^2$. There is some energy stored in the oscillation, but this does not change with time, if we average over many cycles. Therefore, the mean power ($\langle P \rangle$) is

$$\langle P \rangle = \langle \text{power} \rangle = \langle \dot{E} \rangle^2 \quad (24.1)$$

Using our method of writing complex numbers, and our choice, $\dot{E}_0 = (\text{A}^2)/\text{j}\omega L$, we see that this mean power, $\langle \cos(\omega t) \rangle = \text{j}\omega L$, then $d\dot{E}/dt = -\text{j}\omega^2 L^2$. Therefore, in these circumstances, the average power can also be written as

$$\langle P \rangle = \frac{1}{2} \text{Im}(\dot{E}^2) \quad (24.2)$$

In the notation for damped circuits, $\text{j}\omega L$ is replaced by the constant γ (if it is real), where γ corresponds to \dot{E}_0 , and ω corresponds to the resistance R . Thus the rate of the energy loss—the power divided by the damping function—is the resistance at the d.c. unit times the average square of the current:

$$\langle P \rangle = R \langle I^2 \rangle = R \cdot \langle \dot{E}^2 \rangle \quad (24.3)$$

This energy, of course, goes into heating the resistor; it is sometimes called the heating loss or the hunk heating.

Another interesting feature to observe is how much energy is stored. That is not the same as the power, however, although power was at first used to define it since energy, after all, has no sense excepting power, though it costs the only heating (radiative) losses. At any instant there is a certain amount of stored energy, so we would like to calculate the mean stored energy ($\langle E \rangle$) also. This we already calculated with the average of $(\dot{E} \cdot \dot{E})^2$, so we find

$$\begin{aligned} \langle E \rangle &= \text{Im}(\dot{E} \dot{E}^*) = \frac{1}{2} \text{Im}(\dot{E}^2) \\ &= \frac{1}{2} \text{Im}(\omega^2 - \omega_0^2) \text{Im} \dot{E}. \end{aligned} \quad (24.4)$$

Now, when an oscillator is very efficient, and if ω is near ω_0 , so that γ is large, the stored energy is very big. We can get a large stored energy from a relatively small force. But it requires a great deal of work in setting the oscillator going, but then to keep it steady, all it has to do is to fight the friction. The oscillator can have a great deal of energy if the friction is very low, and even though it is oscillating strongly, not much energy is being lost. The efficiency of an oscillator can be measured by how much energy is stored, compared with how much work the source does per oscillation.

How does the stored energy compare with the amount of work that is done in one cycle? This is called the Q of the system, and Q is defined as 2 π times the mean stored energy, divided by the work done per cycle. (If we were to say the work done per cycle instead of per cycle, then the 2 π disappears.)

$$Q = 2\pi \frac{\text{Im}(\dot{E}^2)}{\text{Im}(\dot{E}^2) \cdot \text{Im}(\dot{E}^*)} = \frac{\omega^2 + \omega_0^2}{2\gamma\omega} \quad (24.5)$$

Q is not a very useful number unless it is very large. When γ is relatively large, it gives a measure of how good the oscillator is. People have tried to define Q in the simplest and most useful way: various definitions came up from one another, but it is very large; all definitions are in agreement. The most generally accepted definition is Eq. (24.5), which depends on γ . For a good oscillator, close to resonance, we can simplify (24.5) a little by setting $\omega = \omega_0$, and we then have $Q = \omega_0/\gamma$, which is the definition of Q that we used before.

What is Q for an electrical circuit? To find out, we merely have to integrate \dot{E} over t , $\dot{E} = \text{A} \cos(\omega t)$, and $1/2$ for $\text{Im}(\dot{E}^2)$ (see Table 21-1). The Q of a resistor is $L\omega/R$, where ω is the resonant frequency. If we consider a circuit with a high Q , the reason is that the amount of energy stored in the oscillation is much larger compared with the amount of work done per cycle by the machinery that drives the oscillation.

24-3 Damped oscillations

We now turn to our main topic of discussion, transients. By "transient" is meant a solution of the differential equation when there is no force present, but when the system is not simple at rest. Of course, if x is standing still at the origin w.l.o.g. the force acting on it is zero (position "at rest" here!) suppose the spring has been stretched; say it was stretched by a distance a , while, and then released off the forces. What happens then? Let us first get a rough view of what will happen for a very high Ω system. As long as forces is acting, the stored energy stays the same, and there is a certain amount of work that is being done. Now suppose we turn off the force, and no more work is being done. Then the forces which are taking up the energy of the spring are no longer doing up its energy. There is no more driver. The losses ω_0^2 have to consume, so to speak, the energy that is stored. Let us suppose that $\Omega/\omega_0 = 1000$. Then the work done per cycle is $1/1000$ of the stored energy. Is it too reasonable that it is oscillating with no driving force, that in one cycle the system will still lose a tremendous of its energy? & which means it would have been supplied from the outside, and that is the constant condition, always losing $1/1000$ of its energy per cycle? So, in a sense, does a relatively high Ω system, we would expect that the following would be roughly right (we will later do it exactly, and it will turn out that it was right!)

$$dE/dt = -\omega E/\Omega. \quad (24.8)$$

This is rough because it is mainly for large Ω . In each cycle the system loses a fraction $1/\Omega$ of the stored energy E . Thus in a given amount of time all the energy will change by an amount $\omega E/\Omega$, since due to lack of inertia we can add with the time of $1/\omega$. What is the frequency? Let us suppose that the system moves as nicely, w/o losing any force, that it will go ω_0 will oscillate essentially the same freq. every all by itself. So we will guess that ω is the *resonant frequency* ω_0 . Then we deduce from Eq. (A) that the stored energy will vary as

$$E = E_0 e^{-\omega t/\Omega} = E_0 e^{-\omega t}. \quad (24.9)$$

This would be the minimum of the wave function, incident. What would the form be for amplitude of the oscillation x as a function of the time? The answer? No. The amount of energy in a spring, say, goes to the square of the displacement; the kinetic energy goes to the square of the velocity, so the total energy goes to the square of the displacement. Thus the displacement, the amplitude of oscillation, will decrease half as fast because of the square. In other words, we guess that the solution for the damped transient motion will be an oscillation of frequency close to the natural frequency ω_0 , in which the amplitude of the sine-wave motion will diminish as $e^{-\omega t}$:

$$x = A_0 e^{-\Omega t/\Omega} \cos(\omega_0 t). \quad (24.10)$$

This equation and Eq. 24-1 give us an idea of what we should expect: now let us try to understand the motion precisely by solving the differential equation of the motion i.e.:

so, starting with Eq. (24.1), with no outside force, how do we solve it? Doing physics, we didn't have to worry about the initial conditions w/o the slope when the solution is. And all with our previous experience, let us try an oscillation an exponential curve, $x = A e^{i\omega t}$ (Why do we say this? It is the easiest thing to differentiate!) We put this into Eq. (24.1) (with $m = 0$), using the rule that each time we differentiate x w.r.t. t we get $-i\omega x$ instead of $i\omega x$. So we really quite similar to substitution. This is no surprise looks like this.

$$(-\omega^2 + i\omega \cdot -i\omega^2) A e^{i\omega t} = 0. \quad (24.11)$$

The last factor must be zero for all t , which is impossible unless $(i)^2 = -1$ & A is not solution at all. It stands off, or (2)

$$\omega^2 + \alpha^2 = \omega_0^2 = 0 \quad (24.12)$$

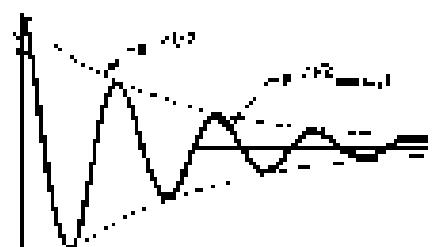


Fig. 24-1. A damped cosine oscillation.

If we can take this root and an m , then we will have a solution in which x need not be zero:

$$x = \text{Re}(2 - \sqrt{\omega_0^2 - \omega^2})e^{i\theta} \quad (24.15)$$

For a while we shall assume that ω is fairly small compared with ω_0 , so that $\omega_0^2 - \omega^2$ is definitely positive, and the i is entering the numerator to bring the square root. This only holds some time so that we get two solutions! Thus

$$\omega_1 = 2\sqrt{2} + \sqrt{\omega_0^2 - \omega^2}e^{i\theta} = \text{Re}(2 + \omega_0) \quad (24.16)$$

and

$$\omega_2 = 2\sqrt{2} - \sqrt{\omega_0^2 - \omega^2}e^{i\theta} = \text{Re}(2 - \omega_0) \quad (24.17)$$

Let us consider the first one, supposing that we had not the issue that the square root has two possible values. (But we know that a solution like $x = A e^{i\theta}$ is, when A is any constant of course.) Now, if we substitute ω_1 , because i is going to come up many times and it takes longer to write, we get all $x_1 = \sqrt{\omega_1^2 - \omega^2} e^{i\theta} = \omega_1 \cos(\theta) - i\omega_1 \sin(\theta) = -i\omega_1 = -\omega_0$, and the $x_2 = A e^{-i\theta} \omega_0 e^{i\theta} = \omega_0$, or what is the same, because of the wonderful properties of an exponential,

$$x_1 = \omega_0 e^{-i\theta} \omega_0 e^{i\theta} \quad (24.18)$$

First, we recognize ω_0 as an oscillation, an oscillation at frequency ω_0 , which is NOT exactly the frequency ω_0 , but is rather close to it if it is a good system... Second, the amplitude of the oscillation is increasing exponentially. If we take, for instance, the real part of (24.16), we get

$$x_1 = A e^{-i\theta} \omega_0 e^{i\theta} \cos \theta \quad (24.19)$$

This is very much like our general solution (24.14) except that the frequency really is ω_0 . This is the very exact result, the same thing—we have the right idea. But everything is not all right. When is x_1 right? Let's see what we have.

The other solution is ω_2 , and we see that the difference is only that the sign of ω_2 is reversed:

$$x_2 = A e^{i\theta} \omega_0 e^{-i\theta} \quad (24.20)$$

What does this mean? We shall soon prove that if x_1 and x_2 are each a possible solution of Eq. (24.3) with $P = 0$, then $x_1 + x_2$ is also a solution of the same equation. So the general solution x is of the undetermined form:

$$x = e^{-i\theta} (A_1 e^{i\theta} + A_2 e^{-i\theta}) \quad (24.21)$$

Now you may wonder why we bother to give this other solution, since we were only happy with the ω_1 in x all by itself. What is the extra use for this? Of course we know we should only take the real part! We know that we must take the real part but how do we mathematical know that we only wanted the real part? When we had a concrete driving force $F(t)$, we put in an x without force to go with it, and the imaginary part of the equation, as we speak, was driven in a definite way. But when we put $F(t) = 0$, our convention that x should be only the real part of whatever we write down is purely our own, and the mathematical equations do not know it yet. The physical world does not care, but the answer that we want or happy with x is not just its complex. The engineer does not know that we are definitely going to take the real part, so it has to present us, so to speak, with a complex conjugate type of solution, so that by putting them together we can always get a real solution; that is what x_2 is doing. That is, in order for x to be real, the x_1 will have to be the complex conjugate of x_2 , so that the imaginary parts disappear. So it turns out that A is the complex conjugate of A_1 , not our real solution is

$$A = e^{-i\theta} (A_1 e^{i\theta} + A_2 e^{-i\theta}). \quad (24.22)$$

So our real solution is an oscillation with a phase shift and a damping—just as advertised.

24-3 Damped oscillations

Now let us see if the above really works. We construct the electrical circuit shown in Fig. 24-2, in which we apply to an inductor the voltage across it is reduced & after an oscillatory form it vanishes by damped oscillations. It is an oscillatory circuit, and it generates a transient of some kind. It corresponds to a circumstance in which we suddenly apply a function to a system that is initially at 0. It is the electrical analog of a charged capacitor, oscillator, and we watch the oscillation on an oscilloscope where we can see the curves that we were talking to anyone. (The horizontal "motion" of the oscilloscope is driven at a uniform speed, while the vertical motion is the voltage across the inductor. The rest of the circuit is only a technical detail.) We would like to repeat the experiment many times, since the persistence of vision is not good enough to see only one trace on the screen. So we do the experiment over and over by closing the switch 60 times a second; each time we close the switch, we can start the oscilloscope triggered sweep, and it draws one curve, over and over, in Figs. 24-3 to 24-6 we see examples of damped oscillations normally photographed on an oscilloscope screen. Figure 24-3 shows a damped oscillation in a circuit which has a high Ω , a small r . In fact, r is so very low, it oscillates many times on the way down.

But let us say what happens if we increase Ω so that the oscillation dies out more rapidly. We can decrease Ω by increasing the resistance R in the circuit. When we increase the resistance in the circuit, it looks like this (Fig. 24-4). Then if we increase the resistance in the circuit still more, it dies out faster yet (Fig. 24-5). But when we just increase the resistance in the circuit, we cannot see any oscillation at all. The question is, is it because our eyes are not good enough? If we increase the resistance yet more, we get a curve like that of Fig. 24-6, which does not appear to have any oscillations, but it persists. Now, how can we explain that by the hand?

The resistance is, of course, proportional to the current in the inductor; therefore, specifically, $r \propto R/L$. Now if we increase the r in the solution (24-14) and (24-15) that we were so happy with earlier, it's + sign in when ω_0^2 cancels out; we must write it a different way, as

$$\omega_0^2 = \omega_0^2/4 - \omega_0^2 \quad \text{and} \quad \omega/2 = \sqrt{\omega_0^2/4 - \omega_0^2}.$$

These are now two real terms and, following the same line of the hand and reasoning as previously, we again find two solutions: $e^{i\omega t}$ and $e^{-i\omega t}$. If we now substitute for x_0 , we get

$$x = A e^{i\omega t} e^{i\omega_0 t} - B e^{-i\omega t} e^{-i\omega_0 t},$$

a non-exponential decay with no oscillations. Likewise, the other solution is

$$x = A e^{i\omega t} e^{-i\omega_0 t} - B e^{-i\omega t} e^{i\omega_0 t}.$$

Note that the square root cannot exceed $\omega_0/2$, because even if $\omega_0 = 0$, one term just equals the other. But ω_0^2 is like $\omega_0^2/4$ but $\omega_0^2/4$ is less than $\omega_0^2/2$, and the term in parentheses is therefore always a positive number. Good goodness! Why? Because if it were negative, we would find a ratio to a positive number, which would mean it was exploding! Increasing more and more resistance into the circuit we know it is not going to go back—quite the contrary. So now we have two solutions, each one by itself a damped exponential... but one is in a minus times plus, plus times minus the other. The general solution is of course a combination of the two, the constant is the combination depending upon how long the motion starts—what the initial conditions of the problem are. In Fig. 24-6, why this circuit appears to be starting, the A is negative and the B is positive, so we get the difference of two exponential curves.

Now let us discuss how we can find the two oscillations, A and B (A , A' and B'), if we know how the motion is started.

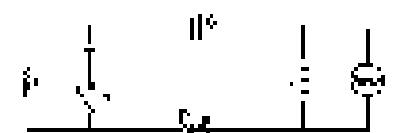


Fig. 24-2. An electrical circuit for damped oscillations.



Figure 24-3



Figure 24-4

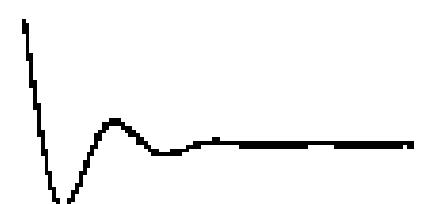


Figure 24-5



Figure 24-6

Suppose that $\dot{x}(t) = 0$ implying that $x = x_0$, and $d\dot{x}/dt = \dot{x}_0$. If we put $t = 0$, $x = x_0$, and $d\dot{x}/dt = \dot{x}_0$ in the equations

$$\begin{aligned} x &= e^{-\alpha t} (x_0 e^{\alpha t} + A e^{-\alpha t}), \\ \dot{x} &= e^{-\alpha t} (-\alpha x_0 + A\alpha e^{-\alpha t}) + (-\alpha/2 + \omega_0^2) A^2 e^{-2\alpha t}, \end{aligned}$$

we find, since $e^{2\alpha t} = e^{2\alpha t} - 1$,

$$\begin{aligned} x_0 &= A + A^2 = 2A, \\ \dot{x}_0 &= (-\alpha/2)(A - A^2) + (\omega_0^2 A - A^2) \\ &= 1A_0/2 - 3A_0(2/A), \end{aligned}$$

where $A = x_0 - M_0$, and $A^2 = A_B - M_0$. Thus we find

$$A_0 = x_0/2$$

and

$$G = \omega_0 + \omega_0 e^{2\alpha t}/2. \quad (24.21)$$

This completely determines A and A^2 , and therefore the complex curve of the Lissajous oscillation in terms of how it begins. Essentially, we can write the solution in either way if we note that

$$e^A + e^{-A} = 2\cos A \quad \text{and} \quad e^{2A} = e^{2\alpha t} = 2\sin A.$$

We may then write the complete solution as

$$x = e^{-\alpha t} \left[x_0 \cos \omega_0 t + \frac{\omega_0 + 2\omega_0^2/2}{\alpha} \sin \omega_0 t \right], \quad (24.22)$$

where $\omega_0 = +\sqrt{\omega_0^2 - \alpha^2}/4$. This is the mathematical expression for the very familiar damped cosine. We shan't make much use of it, but there are a number of points we should like to emphasize that are true in a more general sense.

First of all the "shape" of such a system as $\ddot{x} = -\omega^2 x$ (external force is expressed by zero) in representation of pure exponentials is *pure* (that is we write as e^{At}). This is a good solution to try in such circumstances. The values of α may be complex in general, the imaginary parts representing damping. Finally the extreme mathematical solution of the sinusoidal and exponential function discussed in Chapter 10 which appears physically as a change from oscillatory to exponential because *either* some physical parameter (in this case resistance, α) exceeds some critical value

Linear Systems and Review

25-1 Linear Differential Equations

In this chapter we shall discuss certain aspects of oscillating systems that are more generally true than the particular systems we have been discussing. For our particular system, the "harmonic" equation that we have been solving is

$$m \frac{d^2y}{dt^2} + c m \frac{dy}{dt} + my = f(t). \quad (25.1)$$

Now this particular combination of "operations" on the variable x has the interesting property that if we substitute $(x_1 + x_2)$ for x , then we get the sum of the outputs of x_1 and x_2 ; or, if we multiply x by a, then we get just a times the value combination. This is easy to prove. Just as a "function" function, because we are tired of writing $\text{Comb}(x)$ there is $(^2\text{C}_1)$, we shall use the symbol $L(x)$ instead. When we do this it means the left-hand side of (25.1), with x substituted in. Then the system of equations $\underline{x}(x_1 + x_2)$ would mean the following:

$$\underline{x}(x_1 + x_2) = m \frac{d^2(x_1 + x_2)}{dt^2} + c m \frac{d(x_1 + x_2)}{dt} + m(x_1 + x_2) \quad (25.2)$$

(We underline the \underline{x} since it is a linear composite and it is not an ordinary function.) It is sometimes true that an operator is linear but it makes no "sense" when we call it a function."

Our first statement was that

$$\underline{x}(x + y) = L(x) + L(y), \quad (25.3)$$

which follows from the fact that $y(x+y) = xy + yx$; $d(xy)/dx = dy/dx + x dy/dx$.

Our second statement was to statements,

$$\underline{x}(ax) = a\underline{x}(x). \quad (25.4)$$

[An $\underline{x}(y)$, $(^2\text{C}_1)$ and (25.4) are very closely related, because if we let $x = ax$ into (25.1), this is the same as setting $a = 1$ in (25.4), and so on.]

If there are linear problems, there may be more difficulties, and more terms in \underline{x} ; the question of interest is whether the two equations (25.3) and (25.4) are maintained or not. If they are, we call such a problem a linear problem. In this chapter we shall discuss some of the properties that exist because the system is linear. To appreciate the generality of some of the results, let us first return to our specific analysis of a special equation.

Now let's recall some of the properties of linear differential equations, having illustrated them already with the specific equation (25.1) that we have studied in detail. The first property of interest is this: suppose that we want to solve the differential equation for a transient, the free oscillation, with no driving term. That is, we want to solve

$$\underline{x}(x) = 0. \quad (25.5)$$

Suppose that, by some hook or crook, we have found a particular solution, which we shall call x_p . That is, we have $x = x_p$ for which $\underline{x}(x_p) = 0$. Now we notice that $x_p(t) = e^{rt}$ is a solution to the static equation, we can multiply this special solution by any constant whatever, and get a new solution. In other words, if we had a

- 25-1 Linear Differential Equations
- 25-2 Superposition of solutions
- 25-3 Oscillations in linear systems
- 25-4 Analogy in physics
- 25-5 Series and parallel impediment

motion of a certain "rigid" thin rod when one "end" is again a solution. Since $\underline{L}(x_1) = \underline{L}(x_2) = 0$, $\underline{L}(x_1 + x_2) = 0$.

Now, suppose that, by hand or by stock, we have already found several given x_1 , x_2 also \underline{x}_1 , \underline{x}_2 . (Remember this when we substituted $x = e^{\alpha t}$.) So, finding the function $\underline{x} = \underline{x}_1 + \underline{x}_2$ are values for x , that is, two solutions, x_1 and x_2 . Now let us show that the combination $(x_1 + x_2)$ is also a solution. In other words, if $\underline{L}(x_1) = x_1 - \underline{x}_1 = 0$, $\underline{L}(x_2) = x_2 - \underline{x}_2 = 0$, then $\underline{L}(x_1 + x_2) = \underline{L}(x_1) + \underline{L}(x_2) = 0$. So, if we have found a number of solutions for the motion of a linear system we can add them together.

Combining both two ideas, we see, of course, that we can also add this one and two of the other, if x_1 is a solution, x_2 is not. Thus, for any sum of these two solutions such as $x_1 + x_2$ is also a solution. If we happen to be able to find three solutions, then we find each any combination of the three will always be also a solution, and so on. It turns out that the number of ways we call **linearly independent solutions**⁴ but we have obtained for our oscillating problem is only one. The number of independent solutions that one finds in the general case depends upon what is called the number of degrees of freedom. We shall not discuss this in detail now, but if we have a second-order differential equation there may be many linearly independent solutions, and we have found some of them; so we have the most general solution.

Now let us go on to another proposition which applies to the situation in which the system is subjected to an external force. Suppose the equation is

$$\underline{L}(x) = F(t), \quad (25.5)$$

and suppose that we have found a specific, *particular* solution of it. Let us say that this solution is x_1 , and that $\underline{L}(x_1) = F(t)$. Suppose we want to find yet another solution; suppose we add x_1 to x_2 in one of the ways x_2 is a solution of the free equation (25.4), say x_2 . Then we see by (25.5) that

$$\underline{L}(x_1 + x_2) = \underline{L}(x_1) + \underline{L}(x_2) = F(t) + 0 = F(t). \quad (25.6)$$

Therefore for the "forced" equation we can add any "free" solution, and we get another solution. The free solution is called a **complementary solution**.

When we have no forces acting, and suddenly a new one on, we can not immediately get the steady motion that we solved for with the one *free* solution, but for a while there is a transient. While, since x_1 is free, then, if we wait long enough, the "forced" solution has a limit, and once it keeps on being caused by the force. Ultimately, for any purposes of time, the solution is unique, but initially the motions are distributed with interesting areas, depending on how the object was started.

14-2 Superposition of solutions

Now we come to another interesting proposition. Suppose that we have a certain particular driving force F_1 (let us say an oscillatory one with a certain $\omega_1 = \omega_1$, but any conclusion will be true for any form known form of F_1) and we have solved for the forced motion (with or without the transient; it makes no difference). Now suppose some other forces is acting, let us say F_2 , and we solve the same problem, but for a different force. Then another transient comes along and says, "I leave this problem for you to solve. I give the force $F_1 + F_2$." Can we do it? Of course we can do it, because the solution is the sum of the two solutions x_1 and x_2 for the forces taken separately—a most remarkable claim it is indeed. If we use (25.6), we see that

$$\underline{L}(x_1 + x_2) = \underline{L}(x_1) + \underline{L}(x_2) = F_1(t) + F_2(t) \quad (25.7)$$

⁴ Solutions which cannot be expressed as linear combinations of each other are called **independent**.

This is an example of what is called the principle of superposition for linear systems, and it is very important. It means the following: if we have a complicated function which can be broken up in any convenient manner into a sum of separate pieces, each of which is in some way simple, or else so that it can't easily give rise to which we have worked out how we can solve it separately, then "whatever is available for the whole must, however, we may simply add the pieces of the solution together, in the same manner as the total force is a mechanical sum of pieces." (Fig. 25-1).

I shall give another example of the principle of superposition. In Chapter 13 we saw that it was one of the basic facts of the laws of electricity that if we have a certain distribution of charges, and calculate the electric field E_1 arising from these charges in a certain place, and at the other hand, we have another set of charges, and we calculate the field E_2 due to this, at the same point in space; then if both charge distributions are present at the same time, the field E at that is the sum of E_1 due to one set plus E_2 due to the other. In other words, if we know the field due to a certain charge, then the field due to many charges is merely the vector sum of the fields of the charges taken individually. This is exactly analogous to the above proposition that if we know the result of two given forces taken at the same time, even if the forces are considered separately, then just again the sum of the corresponding individual responses.

The reason why this is true of electricity is that the most basic law of electricity, Maxwell's equations, which determine the electric field, turn out to be differential equations which are linear, i.e., which have the property (25.3). What corresponds to the force is the charge density, the electric field, and the equation which determines the electric field in terms of the charge density.

As another interesting example of the proposition, let us see how it is possible to "select" in a particular radio station at the same time as all the other stations are broadcasting. The radio station transmits, independently, an oscillating electric field of very high frequency which goes out over radio waves. Let's say that the amplitude of the oscillation of the field is unchanged, independent of the speed of the wave, v , which is very slow, and we are not going to worry about it. When one hears this station is broadcasting at a frequency of 780 kilocycles,⁷ this indicates that 780,000 oscillations per second is the frequency of the continuous flow of the driven antenna, and this drives the electrons up and down at that frequency in the antenna. Now at the same time we may have another radio station, in the same town, broadcasting at a different frequency, say 785 kilocycles per second, even the electrons in one antenna are also being driven by that frequency. Now the question is, how is it that we can separate the signal coming up into the antenna at 785 kilocycles from "the driving in at 780 kilocycles"? We certainly do not hear both stations at the same time.

By the principle of superposition, the response of the electric circuit is, in radio, the sum of ψ_1 which is a linear circuit in accordance with setting A to the electric field $E_1 = E_0 \cos(\omega_1 t + \phi)$, and ψ_2 which we will never drive, having been, in fact, every proposition of superposition, because in fact the two cannot exist having both of them in one system. But remember, for a resistive circuit, the response ψ_2 to the amplitude of ψ_1 per unit F , is a "frac-tion of the frequency," looks like Fig. 25-2. If it were a very high Q circuit, the response would have a very sharp maximum. Now suppose that the two stations are comparable in strength, that is, the two forces have the same amplitude. Then we note that we get the sum of ψ_1 and ψ_2 . But, in Fig. 25-2, ψ_2 remembers, while ψ_1 is small. So, in spite of the fact that the two signals are equal in strength, when they go through the radio circuit, a circuit of the form shown for ψ_2 , the frequency of the transmission of one station, then the response to this station is much greater than to the other. Therefore the complete response, with both signals acting, is almost all ψ_2 at ω_2 , and we have selected the station we want.

Now what about the tuning? How do we tune it? We change ω by changing the C or the L of the circuit, because the frequency of the circuit has to do with the combination of L and C . In particular, most radios are built so that one can change the capacitance. When we reverse the switch, we can make a new setting of

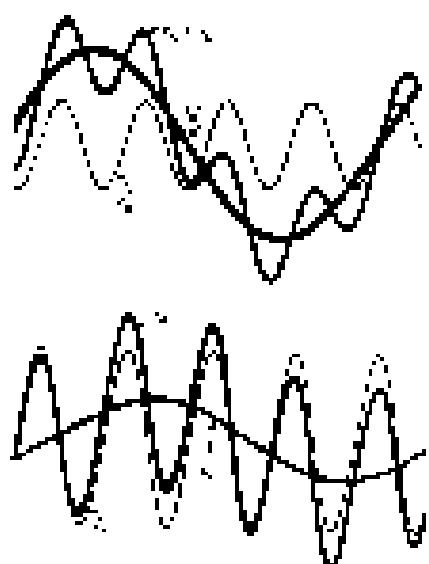


Fig. 25-1. An example of the principle of superposition for linear systems.



Fig. 25-2. The principle of superposition in electrodynamics.



Fig. 25-3. A sharply tuned resonance curve.

the coil, so that the natural frequency of the circuit is shifted, say, to ω_1 . In this circumstance we hear neither one station nor the other; we get silence, provided there is no other station at frequency ω_2 . If we keep on changing the capacitance until the resonant curve is at ω_2 , then of course we hear the other station. That is how tuning works. It is again the principle of superposition, now combined with a resonant response.*

To conclude this discussion, let us describe qualitatively what happens if we proceed further. It is always a linear problem with a given force, where the force is quite complicated. Out of the many possible procedures, there are two especially useful general ways that we can solve the problem. One is this: suppose that we can solve it for specific known forces, such as sine waves of different frequencies. We know it is child's play to solve it for sine waves. So we have the simplest "child's play" cases. Now the question is whether any very complicated force can be represented as the sum of two or more "child's play" forces. In Fig. 25-3 we already have a fairly complicated curve, and of course we can make it more complicated still if we add in more waves. So it is certainly possible to obtain very complicated curves. And, in fact, the reverse is also true: practically every curve can be obtained by adding together enough numbers of sine waves of different wavelengths (or frequencies) for each one of which we know the answer. We just have to know *how* each sine wave belongs to just to make the given F , and then our answer, x , is the sum (superposition sum) of the sine waves, each multiplied by its coefficient of A to F . This method of solution is called the *method of Fourier analysis* or *Fourier synthesis*. We are not going to actually carry out such an analysis just now; we only wish to indicate what is involved.

Another way in which our complicated problem can be solved is the following very interesting one. Suppose that, by some fortuitous mechanical effort, it were possible to solve our problem for a special force, namely an impulse. The force is quickly turned on, and then off. It is all over. Actually we need only care for an impulse of some unit strength; any other strength can be gotten by multiplication by an appropriate factor. We know that the response to an impulse is damped oscillation. Now what can we say about some other force, for instance a force like that of Fig. 25-4?

Such a force can be likened to a succession of blows with a hammer. First there is no force, and all of a sudden there is a steady force—impulse, impulse, impulse, impulse . . . and then it stops. In other words we imagine the external force to be a series of impulses, very close together. Now, we know the result for an impulse, so the result for a whole series of impulses will be a whole series of damped oscillations: it will be the curve for the first impulse, and then slightly later the curve for the second impulse, and the curve for the third impulse, and so on. Thus we can represent, mathematically, the complete situation for arbitrary functions if we know the answer for all impulses. We get the answer for any other force simply by integration. This method is called the *Gren's function method*. A Gran's function is a response to an impulse, and the method of calculating any force by putting together the response of impulses is called the *Gren's function method*.

The physical principles involved in both of these schemes are no simple, involving just one linear equation, that they can be readily understood, but the mathematical problems are, as indeed, the most subtle in science and so on, and so it is best advised for us to attack right now. For we most likely return to this some day when you have had more practice in mechanics. But the idea is very simple indeed.

Finally, we make some remarks on this linear solution that are important. The answer is simple, because we're in Euler form! So most of the time we can linear

* Of modern superheterodyne receivers the local oscillator is more complex. The amplitude is adjusted to a fixed frequency (called the *locoscene*); and an oscillator of variable variable frequency is combined with the input signal in a nonlinear circuit to produce a new frequency (the difference of signal and oscillator frequency), equal to the *IF frequency*, which is then amplified. This will be discussed in Chapter 32.

problems. But just think: it's important in physics that the fundamental laws of physics are often linear. The Maxwell equations for the laws of electricity and magnetism, for example. The general laws of quantum mechanics aren't, as far as we know, to be linear equations. That is why we spend so much time on linear oscillators: because if we understand linear oscillators, we are, really, in principle, in understand a lot of things.

We mention another situation where linear equations are found. When displacement is small, many functions can be approximated linearly. For example, if we have a simple pendulum, the second equation for its motion is

$$\frac{d^2\theta}{dt^2} = -g/l \sin \theta \quad (25.9)$$

This equation can be solved by direct integration, but the easier way to solve it is numerically, as was shown in Chapter 9 on Newton's Laws of Motion. A non-linear equation cannot be solved, numerically, any other way than numerically. Now for small θ , $\sin \theta$ is practically equal to θ , and we have, from eq. (25.9), after division by l , that these are many circumstances where small effects are linear! For the example here the swing of a pendulum through small angles. As another example, if we pull a table cloth, or a spring, the force is proportional to the extension. If we pull hard, we break the spring, and the force is a completely different function of the distance! Linear equations are important. In fact they are so important that perhaps 90 percent of the time we are solving linear equations in physics and in engineering.

25-3 Oscillations in Linear systems

Let us now review the things we have been talking about. In the past few chapters, it is very easy for the physics of oscillations to become obscured by the mathematics. The physics is actually very simple, and if we may forget the mathematics for a moment we will see that we can understand almost everything that happens in an oscillating system. First, if we have only the spring and the weight, it is easy to understand why the system oscillates: it is a consequence of friction. We pull the mass down and the force pulls it back up; it is zero when which is the place it likes to be. It cannot just stop entirely because of its momentum it keeps on going and swinging to the other side and back and forth. See, if there were no friction, we would simply continue an oscillatory motion, and indeed we get one. But if there is even a little bit of friction, then on the return cycle, the swing will not be quite as high as it was the first time.

Now what happens, cycle by cycle? That depends on the kind and amount of friction. Suppose that we could exert a kind of friction force that always remains in the same proportion to the other forces, of gravity and of the spring, as the amplitude of oscillation does. In other words, for smaller oscillations the friction would be weaker than for big oscillations. And in fact friction does not have this property, so a general law of friction must be carefully imagined for the easy purpose of creating a friction that is exactly proportional to the velocity, so that for big oscillations it is stronger and for small oscillations it is weaker. If we happen to have such a kind of friction, then all the end of each successive cycle the system is in the same position as it was at the start except a little bit smaller. All the forces are smaller in the same proportion: the spring force is reduced, the inertial effects are lower because the accelerations are now weaker, and the friction is less too, by our careful design. When we actually have that kind of friction, we find that each oscillation is in exactly the same place as the first one, except reduced in amplitude. If the first cycle dropped 10 percent, e.g., to 90 percent of what it was at the start, the next will drop it to 90 percent of 90 percent, and so on: the ends of the oscillations are reduced by the same factor of themselves in every cycle. An exponential function is a curve which does just that. It changes by the same factor in each equal interval of time. That is, every time the amplitude of one cycle ends, i.e. in the passing from one cycle to the next, the amplitude of the next is a^2 , and of the next, a^3 . So the amplitude is some constant raised to a power equal to the number of cycles times a :

$$A = A_0 a^n \quad (25.10)$$

But of course $\dot{\theta} \sim \omega$, so it is perfectly clear that the general solution will be some kind of oscillation, decaying toward time, in amplitude which goes with $e^{-\alpha t}$, or less. That can be written as $\theta = \theta_0 e^{-\alpha t}$, if α is positive and not less than 1. So this is not the solution itself like $\theta = A \sin(\omega t)$. It is very simple.

What happens if the friction is non-linear? Well, for example, consider a rubber band, so that the friction force is a certain constant amount, and is independent of the size of the oscillation that receives its direction each half-cycle. Then the equation is no longer linear, it becomes hard to solve, and must be solved by the numerical method given in Chapter 3, or by considering each half-cycle separately. The numerical method is the most powerful method of all, and can solve any situation. It is only when we have a simple problem that we can use mathematical analysis.

We have had analysis up to the point where it solves only the simplest possible equations. As soon as the equations get a little more complicated just a little, they can not be solved analytically. But the numerical method, which was advertised at the beginning of the course, can take care of any equation of physical interest.

Now, what about the resonance curve? Why is there a resonance? Just imagine for a moment that there is no friction, and we have one spring, which oscillates by $\theta_0 \cos(\omega t)$. If we let go, the pendulum just sits there, since it went by, of course we could make it go like mad. But if we close our eyes and do not touch it, and let it sit there, what happens, what is going on? Imagine! Since it has no friction, it continues, applying what is it trying to do wrong way. When we happen to have the tuning just right, of course, each tap is given at just the right time, and so it goes higher and higher and higher. So without friction we get a curve which looks like the solid curve in Fig. 22-7 for different frequencies. Our desire, to understand the resonance curve, in order to get the exact shape of the curve it is probably just as well to do the mathematics. The curve goes toward infinity as $\omega \rightarrow \omega_0$, where ω_0 is the natural frequency of the oscillator.

Now suppose there is a little bit of friction; then when the displacement of the oscillator is small, the friction does not stop it much. The resonance curve is the same, except when we are near resonance. The cost of having, infinite near resonance, is not only going to get us into trouble when we do by bumping each time is enough to compensate for the loss of energy by friction during the cycle. So the top of the curve remains off—it does not go to infinity. If there is more friction, the top of the curve is reduced off still more. Now someone might say, "I thought the width of the curve depended on the friction." This is because the curve is usually plotted so that the top of the curve is constant anyway. However, the mathematical expression is even simpler to understand. If we just plot on the curves on Fig. 22-7, we see that frequency is just the friction as it slows the top. If there is less friction, we can go farther up into that little bump before the friction cuts it off, so it's relatively narrow. Then the higher the peak of the curve, the narrower the width of the resonance region.

Finally, we take the case where $\omega > \omega_0$, which is called over-damped. It turns out that if there is too much friction, the system does not oscillate at all. The energy in the spring is barely able to move it against the frictional force, and so it slowly comes down to the equilibrium point.

22-4 Analogy in physics

The first lesson of this review is to note that masses and springs are not the only linear systems that there are others. In particular, there are electrical systems called linear circuits in which we find a complete analog to mechanical systems. We did not learn exactly how one of the objects in an electrical circuit works in the way that one or two in mechanics does in mechanics; we may assert, from an experimentally verifiable fact that they behave as stated.

For example, let us take the simplest possible circuit—just one. We have a piece of wire which is just a resistor, and we connect it to a difference in potential. If "difference" means that there are voltages along through the wire, then the current

another terminal. The work done is qV . The higher the voltage difference, the more work was done when we charge, or worse, "to be" from the high potential end of the terminal to the low-potential one. So charge release energy is going from one end to the other. Now the charges do nothing but fly from one end straight to the other end; the resistance being effectively negligible to the current, and the resistance unless the following are far almost all ordinary substances if there is a current I , that is, on the to many charges per second moving down, the number per second that comes available through the wire is proportional to the total current and there will be a linear proportionality to the total voltage there:

$$I = qE = R(q/dm) \quad (25.11)$$

The coefficient R is called the resistance, and the equation is called Ohm's law. The unit of resistance, because it is equal to one volt per ampere. In magnetized situations, to get such a frictional force in proportion to the voltage is difficult; in an electrical system, it is very easy, and this is why it is extremely convenient to meet resistors.

When often interested to know much work is done per second, the power loss of the energy liberated by the charges as they move over the wire. When we carry a charge (or voltage) V , the work required to move charge per second would be qV/qdm , which is the same as $W = qV = q^2R$. This is called the heating loss. This is how much heat is generated in the resistance per second, by the conversion of energy. It is in fact the resistive energy dissipation lost by a work.

Of course, there are other interesting properties of mechanical systems, such as the mass (potential), and it turns out that there is an electrical analog to inertia also. It is possible to make something to act as inductor, having a roughly called inductance, such that a current once started through the inductor, does not want to stop. It requires a voltage in order to change the current! If the current is constant, there is no voltage across an inductor. We already do not know anything about inductance, it is only that we always the current, but the effects of inductance show up. The equation is

$$V = L(dI/dt) = Ld^2q/dt^2, \quad (25.12)$$

and the unit of inductance, called the Henry, is just the one you applied to the inductance so we have here produced a change of charge per second in the current. Equation (25.12) is the law of Newton's law for electricity, if we will: $F = ma$ corresponds to $V = Ld^2q/dt^2$, and L corresponds to inertia. All of the consequent equations for the two kinds of systems will have the same calculations because, if we do the integrations, we can change any letter in the equations, changing letter and we get the same equation, the value variables will have a correspondence in the two systems.

Now what electrical thing corresponds to the mechanical spring, in which there was a force proportional to the stretch? If we start with $F = -kx$, and replace $F = E$ and $x = q$, we get $E = -qV$. It turns out that there is such a thing, in fact, it is the only one of the three circuit elements we can really realize and, because we did study a pair of parallel plates, and we found that if there were a charge of certain equal, opposite amounts on each plate, the field between them would be proportional to the size of these caps. So the work there in moving a test charge across the gap from one plate to the other is precisely proportional to the charge. The work is the product of the voltage difference, which is the time integral of the electric field from one place to another. It turns out for historical reasons, that the constant of proportionality is not called C , but $1/C$. We could have been called C , but it was not. So we have

$$V = q/C. \quad (25.13)$$

The unit of capacitance, C , is the factor a charge of one coulomb on each plate of capacitor upon a given a voltage difference of one volt.

There are two analogies, and the right one corresponding to the oscillating circuit becomes the following, by direct substitution of & 25.14 for α , etc.

$$v(t)/x_0(t) = v_0(x_0(0)) + \omega t - K, \quad (25.14)$$

$$2\omega^2 q(t)^2 + R^2 x_0(t)^2 + \omega^2 C = K. \quad (25.15)$$

Now everything we learned about (25.14) can be transformed to apply to (25.15). This is because it is linear, & much the same that this is a similar thing we learned.

Suppose we have a mechanical system which is quite complicated, but just consists of a spring, but several masses on several springs. i.e. think again for why do we do? Now it's. Perhaps, for best, we can make an equivalent circuit which will have the same equations as the thing we are trying to analyze. For instance, if we wanted to analyze a mass on a spring, why can we not make an equivalent circuit in which we see an inductance proportional to the mass, a resistance proportional to the constricting force, R^2 is proportional to K , all in the same way? Then, of course, this one final circuit will be the exact analog of our mechanical case, in the sense that whatever it does, in response to V (V also is more or less equivalent to the forces that are being), so the x would do in response to the V signal. So if we have a complicated thing with a whole lot of interconnected elements, we can incorporate a whole lot of resistances, inductances, & capacitors, etc to make the most wholly analogous system. What is the advantage here? One problem is just as hard (or as easy) as the other, because they are exactly equivalent. The advantage is not that it is any easier to solve the mathematical equations, but we discuss. That we have an electrical circuit, although this is the method used by electrical engineers!, but instead, the real reason for looking at the analog is that it is easier to make the electrical circuit out of charge + anything in the system.

Suppose we have designed an automobile, and want to know how much it is going to shake when it goes over a certain kind of bumpy road. We build an electrical circuit with inductances to represent the inertia of the wheels, spring constants as resistances to represent the springs of the wheels, and resistors to represent the shock absorbers, and we do for all other parts of the automobile. Then we see a bumpy road. All right, we apply it, where there's a generator, which represents wheels and such, a kind of source, and then look at how the left wheel rotates by measuring the charge on some capacitor. Having measured it (it is easy to do, we don't need to be applying too much). Do we need a shock absorber, or less shock absorber? With a computerized thing like this, we make the car actually change the shock absorber, and solve it all over again? Not, we simply turn a dial, and number two shock absorber number three, so we get a more shock absorber. The bumpers are better—All right, we try less. The bumpers are still worse; we change the stiffness of the spring (dial 17), and we do not all these things electronically, with merely the turn of a knob.

This is called an *analog computer*. It is a device where it isolates the problem that we want to solve by making another situation, which has the same equations, but in another circumstance of nature, and which is easier to build, to measure, to adjust, and to destroy.

25.6 Series and parallel impedances

Finally, there is an important item which is not quite at the center of review. This has to do with an electrical circuit in which there's more than one element. For example, when we have an inductor, a resistor, and a capacitor connected as in Fig. 24.3, we note that all the charges went through every one of the three, but the current is such a singly connected thing as the sum of all potentials along the wire. Since the current is the same at each node, the voltage across R is iR , the voltage across C is $Q_0/(iC)$, and so on. So, the overall voltage is the sum of these, and this leads to Eq. (25.15). Using complex numbers, we find that we could solve the equation for the steady state solution in response to v & q_0 .

oppositional force. We have found that $V = RI$. Now R is called the *impedance* of the particular circuit. This means that if we apply a sinusoidal voltage V , we get a current I .

Now suppose we have a more complicated circuit which has two pieces, where by definition these contain impedances, Z_1 and Z_2 , and we put them in series (Fig. 25-5a) and apply a voltage. What happens? It is now a little more complicated, but if I is the current through Z_1 , the voltage difference across Z_1 is $V_1 = IZ_1$. Likewise, the voltage across Z_2 is $V_2 = IZ_2$. The current I just goes through both. Therefore the total voltage is the sum of the voltages across the two sections and is equal to $V = V_1 + V_2 = (Z_1 + Z_2)I$. This means that the voltage on the complete circuit can be written $V = \bar{Z}I$, where the \bar{Z} of the combined section in series is the sum of the two Z 's of the separate pieces:

$$\bar{Z} = Z_1 + Z_2. \quad (25.6)$$

This is not the only way things may be connected. We may also connect them in another way, called *parallel connection* (Fig. 25-5b). Now we see that a given voltage across the terminals, if the connecting wires are perfect conductors, is effectively applied to both of the impedances, and will be the same across each independently. Therefore the current through Z_1 is equal to $I_1 = V/Z_1$. The current in Z_2 is $I_2 = V/Z_2$. It is the same voltage. Now the total current which is supplied to the terminals is the sum of the currents in the two sections: $I = I_1 + I_2 = V/\bar{Z}$. This can be written as

$$V = \frac{I}{(1/\bar{Z}_1) + (1/\bar{Z}_2)} = \bar{Z}I.$$

Thus

$$1/\bar{Z}_t = 1/\bar{Z}_1 + 1/\bar{Z}_2. \quad (25.7)$$

More complicated circuits can sometimes be simplified by taking pieces of them, working out the successive *impedances* of the pieces, and combining the circuit together step by step, using the above rules. If we have any kind of circuit with three impedances connected in π , since it ways, and if we include the voltage in the form of three generators having no impedances (where we pass charge through it, the generator adds a voltage V), then the following principles apply. (1) At any junction, the sum of the currents into a junction is zero. That is, I , the current which comes in must come back out. (2) Two carry a charge around my loop, and back as when it starts, the net work done is zero. These rules are called *Kirchhoff's laws* for electrical circuits. Their systematic application to complicated circuits often simplifies the analysis of such circuits. We mention them here in connection with Eqs. (25.6b) and (25.17). In fact you have already come across such things as you need to discuss in laboratory work. They will be discussed again in more detail next year.

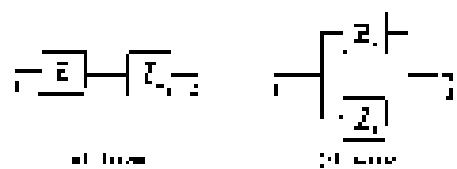


Fig. 25-5. Two impedances, connected in series and in parallel.

Appendix The Principle of Least Time

26-1 Light

This is the first of a number of chapters on the subject of electrodynamics. Light which we see is only a small part of the spectrum of the same kind of thing, the various parts of this spectrum being distinguished by different values of a certain quantity which varies. This variable quantity can be called the "wavelength," *A*. As *A* varies in the visible spectrum, the light apparently changes color from red to violet. If we explore the spectrum systematically, from long wavelength toward short ones, we shall begin with what are usually called *radio waves*. Radiations of relatively long *A* are in every range of wavelength, once even longer than those used in regular broadcasts. Next come waves whose wavelengths are approaching the limit 900 microns. Then there are the waves of higher energy, i.e., *radio waves*, *micrometers*, and so on. There are no actual boundaries between one range of wavelength and another, because not all *A* are present together with the *A*'s just before. The numbers associated with a given name for the waves are only approximate and, of course, so are the ranges we give in the different ranges.

From a long way down through the millimeter waves, we come to what we call the infrared, and thence to the visible spectrum. Then going in the other direction, we get into a region which is called the ultraviolet. When the wavelength *A* goes into the x-ray region, we cannot define precisely where this is; it is roughly at 10^{-3} m. or 10^{-4} μ . These are "soft" x rays, even there are extremely hard, and very hard x rays, too. In the x rays, and so on, for smaller and smaller values of *A*, dimension called the wavelength.

Within the last range of wavelengths, there are also some regions of approximation which are especially interesting. If one of these conditions exists so that the wavelengths involved are very small compared with the dimensions of the equipment available for their study, for instance the photon energies, and perhaps the energy density, in small comparison with the energy sensitivity of the equipment. Under these conditions we can make a rough but approximation by a method called *geometrical optics*. This is either *lens* or *ray* optics. In one circumstance, in the dimensions of the equipment, when it is difficult to manage with visible light but easier with x rays, the photon energies are still negligibly small, there is a very useful approximation to be made by studying the behavior of the waves without distinguishing the exact nature of the photons. This method is based on the classical theory of electromagnetic radiation, which will be discussed in a later chapter. Next, if we go to very small wavelengths where we can disregard the wave character but the photons have a very large energy compared with the sensitivity of the equipment, comes just a simple ray. This is the single photon you are which we will discuss very, very roughly. The complete theory, which handles the photons into account, will not be available to us for a long time.

In this chapter our discussion is limited to the geometrical optics region in which we forget about the wavelength and the photon character of the light, which will all be explained later on. We do not even feel it necessary to say what the light is, but just think. And it follows on a little while subsequently with the dimensions of apparatus. All this must be said in order to emphasize the *classical* part we are going to talk about, is only a very rough approximation; it is one of the subjects that we shall have to return to again and again, very quickly unless it becomes we shall almost immediately go to the more accurate method.

26-1 Light

26-2 Reflection and refraction

26-3 Fermat's principle of least time

26-4 Applications of Fermat's principle

26-5 A more precise statement of Fermat's principle

26-6 How it works

Although geometrical optics is just an approximation, it is of very great importance technically and it can also be historically. We shall present this subject more historically than some of the others in order to give some idea of the development of a physical theory or physical idea.

First, light is, of course, familiar to everybody, and has been familiar since long immemorial. Now our problem is, by what process do we see light? There may be many answers, but it usually settles down to one which is that there is scattering, which enters the eye—which becomes an object into the eye. We have heard that idea so long that we accept it, and it is easier, impressive for us, to realize that every intelligent machine must do something like scattering to come out of the second look for the object. For example, some other important idea is this: more than as light goes from one place to another, it goes in straight lines; if there is nothing in the way, and that the rays do not say to interfere with one another. That is, light is scattering in all directions at the room, but the light that is passing across our line of vision does not affect the light that comes to us from some object. This was once a very powerful argument against the corpuscular theory; it was used by Huygen. If light were like a lot of arrows shooting along, how could other arrows pass through them easily? Such philosophical arguments are not of much weight. One could always say that light is made up of arrows which go through each other.

24-2 Reflection and refraction

The discussion above gives enough of the basic view of geometrical optics—now we have to go a little further into the quantitative aspects. This is, we have light going only in straight lines between two points, now we want to study the behavior of light when it hits various materials. The simplest object is a mirror, and the basic idea is this: when the light hits the mirror, it does not continue in a straight line, but bounces off the mirror into a new set of light lines, which change when we change the inclination of the mirror. The question for consideration now, what is the relation between these angles? This has been a problem for a long time. The light striking a mirror bounces in such a way that the two angles, between each beam and the mirror, are equal. For some reason, it is customary to measure the angle from the normal to the mirror surface. This is the standard law of reflection is

$$n_r = n_i \quad (24.1)$$

This is a simple enough proposition, but a more difficult problem is this: when light goes from one medium into another, for example from air into water, where also we see that it does not go in a straight line. In the water the ray is bent, or bent to its path in the air; if we change the angle i , so that it comes even more nearly vertically, then the angle of "refraction" is not so great. But we tilt the beam of light at quite an angle, then the deviation angle is very large. The question is, what is the relation of one angle to the other? This also provided the excuse for a long time, and here they never found the answer! It is, however, one of the few places, in all of Greek physics that one may find a systematic and exact table. Almagest, written by Ptolemy, made a list of the angle in water, for each of a number of different angles in air. Table 24-1 shows the angle in the air, in degrees, and the corresponding angle as measured in the water. (Gibson says that the Greek scientist never did any experiments. But it would be impossible to explain this table of values without knowing the right law, except by experiment. It should be noted, however, that these do not represent independent measurements for each angle but only a few measurements taken from a few measurements, so they all fit perfectly in a paragraph.)

This, then, is one of the important steps in the development of physical science: we observe an effect, then we measure it and list it in a table; then we try to find the rule for which one thing can be connected with another. The classic one is that table was made in 140 A.D., but it was not until 1811 that someone finally found the rule connecting the two angles. The rule, found by Malus and

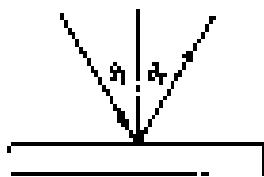


Fig. 24-1. The angle of incidence is equal to the angle of reflection

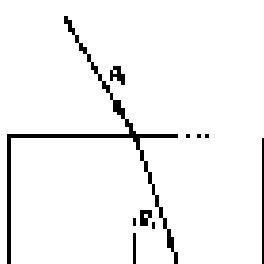


Fig. 24-2. A light ray is refracted when it passes from one medium into another.

Table 24-1

Angle in air	Angle in water
10°	3°
20°	13.1°
30°	22.1°
40°	30°
50°	35°
60°	40.1°
70°	44.1°
80°	48°

Snell's Dutch mathematician, is as follows: if θ is the angle in air and ϕ is the angle in the water, then it is known that the sine of θ is equal to some constant multiple of the sine of ϕ :

$$\sin \theta / \sin \phi = \text{const}$$

In water the number c is approximately 1.33. Equation (26.2) is called Snell's law; it permits us to predict how the light is going to bend when you go from air into water. Table 26-2 shows the angles in air and in water according to Snell's law. Note the remarkable agreement with Fizeau's law.

26-3 Fermat's principle of least time

Now let us further develop our discussion; we want to have the law of reflection. First we have an observation, then we have numbers that we measure, then we have a law which summarizes all the numbers. But the real object of science is that we can find a way of thinking with that the law is evident.

The last way of thinking that made the law about the behavior of light evident was discovered by Fermat in about 1660, and it is called the principle of least time, or Fermat's principle. The idea is this: "Without all possible paths it is only one in just from one point to another. Light takes the path which requires the shortest time."

Let us first show that this is true for the case of the mirror, i.e., this simple principle exists as both the law of straight line propagation and the law of the mirror. So we are going in a circle, shouldn't that be true? And the solution to the following problem: In Fig. 26-3 are shown two points A and B , and a plane mirror MN . What is the way to get from A to B ? In the easiest case? The answer is to go straight from A to B . But now we do the exercise that the light has to strike the mirror and come back to the shortest time. It is somewhat difficult. One way would be to go as quickly as possible to the mirror and then go to B from the point $A'AB$. Of course, we then have a long path $A'B$. If we move just a little to the right, to C , we slightly increase the travel distance, but we greatly decrease the speed, and so the total path length and therefore the travel time, is less. Thus can we find the point C for which the time is the shortest? We can find it very nicely by a geometrical trick.

We construct on the other side of MN an artificial point B' . What is the significance of the point B' ? As the point B is above the line. Then we draw the line BB' . Now because MN is a right angle and $AB \perp MN$, AB is equal to $B'B$. There is the second construction, $A'E \perp ED$, which is proportional to the time it will take if the light travels with constant velocity, is also the sum of the two angles $AED + EBD$. Before the problem becomes, when is the sum of these two lengths the least? The answer is easy: when the line goes through point C or a longer line from A to B . In other words, we have to find the point where we go toward the central point, and that will be the corner point. Now if $A'C'B$ is a straight line, then angle BCE is equal to angle $B'CE$ and hence to angle ACB . Thus the statement that the angle of incidence equals the angle of reflection is equivalent to the statement that the light goes to the center in such a way that it comes back to the point B in the least possible time. Originally, this statement was made by Hero of Alexandria that the light travels in such a way that it goes to the mirror and to the other point in the shortest possible distance, so it is not a modern theory. It was this the original Fermat imagined, himself! That we have reflection operated on a similar basis. (In refraction, light obviously does not use the path of shortest distance, so Fermat cited the law that it takes the shortest time.)

Before we go on to analyze refraction, we should make one more remark about the mirror. If we have a source of light at the point B and it sends light toward the mirror, then we see that the light which goes to A from the point B comes to A in exactly the same manner as it would come from A if there were an object at B' and no mirror. Now of course the wave fronts with the light which enters it physically can't be here in object B and a mirror which makes the light come

Table 26-2

Angle in air	Angle in water
0°	0°
30°	19°
45°	22°
60°	26°
75°	30°
90°	44.7°
70°	43°
60°	48°

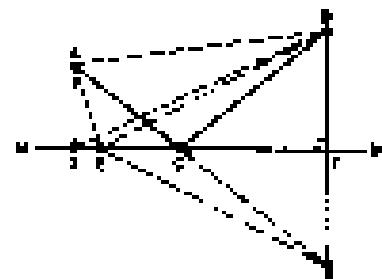


Fig. 26-3. Illustration of the principle of least time.

into the eye is exactly the same manner as it would have come into the eye if the object were at s' , then the eye-brain system interprets each, assuming it does not know how much, as being an object at s' . So the illusion is that there is an object behind the mirror is nearly due to this. See that the light which is entering the eye is entering it exactly the same manner, physically, as it would have entered had there been an object back there; except for the effect of the mirror, and our knowledge of the existence of the mirror, and so on, which is corrected in the brain.

Now let us demonstrate that the principle of least time will give Fermat's law of reflection. We must however, make an assumption about the speed of light in water. We shall assume that the speed of light in water is lower than the speed of light in air by a certain factor, say μ .

In Fig. 26-4, our position is again to go from A to B in the shortest time. To illustrate that the best thing to do is not just to go at a straight line, let us imagine that a beautiful girl has fallen out of a boat, and she is swimming back up to the shore at point X . The line material back, swimming, we are supposed to be land, and we set the distance and we can run and swim, say a . But we can run faster than we can swim. What should we do? Should we swim in a straight line? (You're doubtful.) However, by using a little more intelligence we would realize that we could be advantageously to travel a little greater distance in land in order to decrease the distance in the water, because we go or must slower in the water. Following the line of reasoning out, we would say the right thing to do is to run until we come fully into the water, and then the path becomes the curve case of all possible paths. If it is the shortest path, that means until we take any other, it will be longer. Since we were going the time it takes against the position of point X we want to get a curve something like that shown in Fig. 26-5, where point C corresponds to the shortest of all possible times. This means that if we move the point X to point X' , in the first approximation there is essentially no change in time because the slope is zero at the bottom of the curve. So our way of thinking the time will be unchanged. Then we move the place by a very small amount, and to showed that there be essentially no change in time. (Of course there is an infinitesimal, except of a second order, we ought to have a positive increase in displacement in either direction from C .) So we consider a nearby point X' and we calculate how long it would take to go from A to X' by the two paths, and compare the new path with the old path. It is very easy to do. We note the differences. Assume to be really very small, if the distance AC is short. First look at the path on land. If we drop a perpendicular to AC , we see that this path is supposed to be parallel to EC . Let us say we go t_1 using AC distance. On the other hand, in the water, by dropping a corresponding perpendicular, CF , we find that we have to go the extra distance CF , and that is equal to AC . Or, in view we split the time it would have taken to go the first part AC , but we have the time it would take to go the distance CF . Those times must be equal since, at the first approximation, there is no change in time. But supposing that in the water the speed is μ times as fast as it is air, then we must have

$$AC^2 = \mu \cdot CF \quad (26-3)$$

Therefore we see that when we have the light point Y in XYC — $c \cdot t_1$ on AC or, canceling the common c by c because lengths AC and CF are equal, that

$$XYC = CCF \quad \text{and} \quad YCF = XYC - t_1 \quad (26-4)$$

we have

$$\sin \theta_1 = \lambda \sin \theta_F \quad (26-5)$$

So we need to go from one point to another in the least time when the ratio of speeds is λ , the light should enter at such an angle that the ratio of the sines of the angles, λ , and λ , is the ratio of the speeds in the two media.

26-4 Applications of Fermat's principle

Now let me consider some of the interesting consequences of the principle of least time. First is the principle of reciprocity. If we go from A to B we have found the path of the least time; the light path in the aggregate, returning back again goes at the same speed in any direction; the return path will be the same path and therefore, if light can be sent one way, it can be sent the other way.

Another application is a glass block with plane parallel faces at an angle to a light beam. Light, in going through the block from a point A to a point B (Fig. 26-6) does not go through in a straight line, but instead it deviates. The law is the same by making the angle in each block the same, so though it does a little bit in the air, the beam is simply displaced parallel to itself because the angles it makes are the same.

A third interesting phenomenon is the fact that when we see the sun setting, it is actually below the horizon! It does not look as though it is below the horizon, but it is (Fig. 26-7). The earth's atmosphere is thin at the top and denser at the bottom. Light's wavelength changes with density of atmosphere, and so the light you see can get to points a beyond the horizon more easily. Instead of going in a straight line, it will follow a curve where it goes slowly by going through there at a steeper angle. When it appears to be below the horizon, it is actually already well below the horizon. Another example of the phenomenon is the rising sun, one lesson... while looking up at the sky. One sees "water" on the horizon when he gets there, it is as dry as the desert! The phenomenon is the following. What we are really seeing is the sky light "reflected" in the sunlight from the sky, heading P . The ray can end up in the eye, as shown in Fig. 26-8, when the air is very hot just above the ground but it is cooler up higher. Hence, air is more rarefied than cooler air is thinner, and this does reduce the speed of light here. That is to say, light goes faster in the hot region than in the cool region. Therefore, instead of the light proceeding to come in the straightforward way, it also has a lateral displacement, by which goes into the region where it is faster for awhile, in order to come in. See it is going in the eye.

An even more important example of the principle of least time, suppose that we could live in a strange situation where we have all the light. Let us suppose at point P , collected back together at another point P' (Fig. 26-9). But notice, of course, that the light can go in a straight line from P to P' . That is all right. But how can we arrange that naturally? i.e., i. given a light, let's say P , let the light starting out from the point P go to this spot P' ? We want to reflect all the light back to whatever collimator. How? If the light always takes the path of least time, then certainly it is much too hard to get over all those other paths. The only way then the path can be perfectly correct is take several separate paths and to make those times exactly equal! Otherwise, it would violate the law of least time. Hence here the problem of making a focusing system is merely bringing a desire + that it is to be the same time for the light to go on all the different paths!

This is easy to do. Suppose that we add a piece of glass in which light goes slower than it does in the air (Fig. 26-10). Now consider a ray which goes in P in the path PQP' . That is a longer path than from P directly to P' and no doubt looks a longer time. But if we were allowed to travel close to just the right thickness we shall be able figure in some thick iron ring, to compensate the extra time that it would take the light to go at a angle! In those circumstances we can arrange so that the iron ring will take the straight, smooth path from P to P' ; it is too long in the path PQ . Travelling since the ray QP' is not yet required, it is not quite as long as PQP' , and we do not have to compensate as much as for the straight-line, but we do have to compensate somehow... With an iron ring of a piece of glass (as shown in Fig. 26-10), PQ is this shape, so the light which comes from P will go to P' . This, of course, is well known to us, and we call such a device a lens and that lens. In terms of $\sin \theta$, we shall try to calculate what shape the lens has to have to make a perfect focus.

Take another example. Suppose we wish to change some distance so that the light from P always goes to P' (Fig. 26-11). For every part Q you'd come in mirror

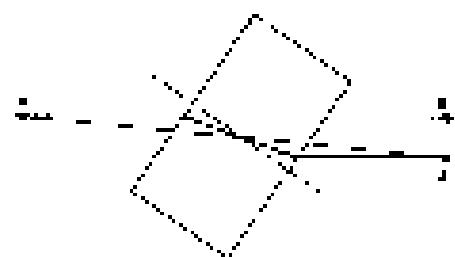


Fig. 26-6. A beam of light refracted as it passes through a transparent block.



Fig. 26-7. Near the horizon, the upper path is lighter than the true path by about $1/2$ degree.



Fig. 26-8. A prism.



Fig. 26-9. An optical "block box."

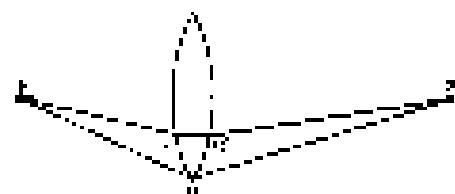


Fig. 26-10. A focusing optic, a lens.

and comes back, and r_1 -times must be equal. Here the light always travels in air, so the time and the distance are proportional. Therefore the distances d_1 , all the times are the same; it is the same as the statement that the total distance is the same. Thus the sum of the two distances r_1 and r_2 must be a constant. An ellipse is that curve which has the property that the sum of the distances from two points is a constant, for every point on the ellipse, thus we can be sure that the light from one focus will come to the other.

The second principle applies for us in finding the light of a star. The great 20th-century Poincaré principle is built on the following principle. Imagine a series billions of rays coming; we would like to cause all the light that comes in to come to a focus. Of course we cannot do the rays one by one all the way up to the star, but we still want to know whether the times are equal. Of course we know that when the various rays arrive at some place SF , perpendicular to the rays, all the times to this place are equal (Fig. 24-11). The rays must then come down to the mirror and proceed toward their equal times. That is, we find a curve which has the property that the sum of the distances $XF = SF + SF'$ is a constant, no matter where X is chosen. A ready way to find it is to extend the length of the line SF even to a plane CC' . Now if we arrange our curve so that $A'X' = AF$, $B'X' = BF$, $C'X' = CF$, and so on, we will have our curve, because then obviously $A'X' + A'X' = AF + BF = BC + CF$ will be constant. This curve is called a *parabola*; the name is made in the shape of a parabola.

The above examples illustrate the principles upon which such optical devices can be designed. The exact curves can be calculated using the principle and, to focus perfectly, the travel times must be exactly equal for all light rays, as well as being less than for any other energy path.

We have discussed these focusing optical devices, but now let me next discuss the further development of the theory. When we use the first principle of the principle of least time, our first inclination might be to say, "Well, I am not going to do it; it is delightful; but the question is, does it help us in understanding the subject?" Somebody may say, "Yes, look at how many things we can understand!" Another says, "Very well, but I can understand without that. I needn't know such things; every singer who makes equal angles with C is a singer." Then figure out a lens, too, because obviously the singer is in front through an angle given by Snell's law. Rightly he statement of least time and the sentence, "But angles are equal on reflection, and that the ratios of the angles are proportional on reflection, are the same." So is it merely a philosophical question, or one of utility? That can be argued on both sides.

However, the importance of a powerful principle is that it provides new things.

It is easy to show that there are a number of new things predicted by Fermat's principle. First, suppose that there are three media, water, and air, and we perform a refraction experiment, and measure the index of refraction between them. Let us call n_{12} the index of air against water, (2) , n_{23} the index of air against glass, (3) , if we measured water against glass, we would find another index, which we shall call n_{13} . But there is no reason why there should be any connection between n_{12} , n_{23} , and n_{13} . On the other hand, according to the idea of least time, there is a certain relationship. The index n_{13} is the ratio of two things, the speed in air to the speed in water, v_{air}/v_{water} , is the ratio of the speed in air to the speed in glass, v_{glass} , is the ratio of the speed in water to the speed in glass. Therefore we can tell that $n_{13} = n_{12} \cdot n_{23}$.

$$n_{13} = \frac{v_2}{v_3} = \frac{n_{12}}{v_{air}/v_{water}} = \frac{n_{12}}{n_{23}}. \quad (24.5)$$

In other words, we predict the refractive index for a new pair of materials can be obtained from the indices of the individual materials with respect to air or vacuum. So if we measure the speed of light in *n* materials, and have this get a single number for each material, namely its index relative to vacuum, called n , then



Fig. 24-11. An elliptical mirror.

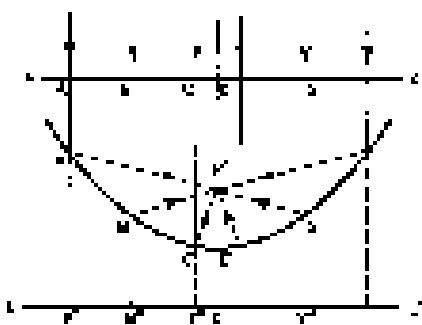


Fig. 24-12. A paraboloid mirror.

v_w is the speed in water, v_a is the speed in air, etc.). Then our formula is easy. The index for July 1st materials hand is

$$n_w = \frac{v_a}{v_w} = \frac{c_w}{c_a}. \quad (26-6)$$

Using only Snell's law, the basis for a principle of this kind,⁴ But of course this prediction works. The relation (26-6) was known very early, and was a very strong argument for the principle of least time.

Another argument for the principle of least time, another prediction, is that if we measure the speed of light in water, it will be lower than in air. This is a prediction which is completely different. Yet it is a brilliant prediction, because all we have so far measured is a single law which is a "familiar" part of which is quite different from the observations from which Descartes deduces the rule of least time. In fact, it is true that the speed in water is lower than the speed in air, by just the proportion that is needed to get the right index!

26-5 A more precise statement of Fermat's principle

Actually, we must make the statement of the principle of least time a little more carefully. It was not stated correctly above. It is now often called the principle of least time and we have gone along with the incorrect statement for so long that we don't even see what the error is! Let me illustrate. Suppose we had a mirror as in Fig. 26-1. When asked, "Is light sent in? Does it go to the mirror?" The path of least time is clearly AB. So some people might say, "Sometimes it is a question of time." It is a minimum in time, but as certainly as π is π , π could have a still longer value! The correct statement is the following: a ray going in a certain particular path has the property that if we make a small change (say a tiny horizontal shift) in the ray in any manner whatsoever, say in the location at which it comes to the mirror, or the shape of the mirror, or anything, the ray will have "second-order" change in the time. There will be only a second-order change in the time. In other words, the principle is that light takes a path such that there are many other paths nearby which take almost exactly the same time.

The difficulty is another difficulty with the principle of least time, and one which people who do not like this kind of a theory might never stomach. With Snell's theory we can "understand" light. Light goes along, it goes *in* surface, it bounces back, it goes *out* of the surface. The idea of causality, that it goes from one point to another and *arrives* at one point, is very comfortable. But the principle of least time is a completely cut-and-dried philosophical principle about the way nature works. Instead of saying "It is causal," it is causal *but* when we do something, something else happens, and this is always the worse up the situation, and nothing else would in the shortest time, or the extreme one, and choose that path. If we reverse it, it does not fit "calvin". That is itself the best he can find and check them against each other. He cannot say, "It is causal," in this way. This is the feature which, of course, not known to geometer, optics, and which is involved in consideration of investigating the meaning of it's to approximately how far away the light must "travel" the path in order to obtain the best performance. It's feasible to do this with light, because the wave-lengths are so terribly short. But with radio waves, say 10 m waves, the distances over which the radio-waves are checking the paths. If we have a source of radio-waves, a detector, are a slit, something (Fig. 26-13). The rays of light, we go from S to D because it is a straight line. And if we move down the slit a little bit—they still go—but now if we move the slit a little bit to D', the waves will not go through the wide slit from S to D', because they come several picoseconds apart, and very, very, very small. They all can expand to different times. On the other hand, if we prevent the radiation from spreading. To get the closing the slit down to a very narrow slit. Then there is but one path available, and the

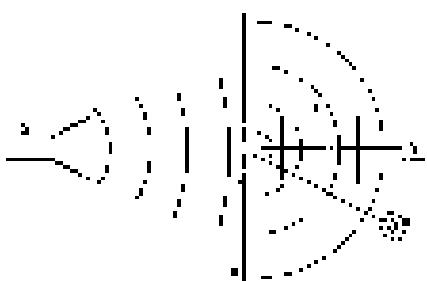


Fig. 26-13. The passage of radio-waves through a narrow slit.

⁴ Although a common feature of the standard assumptions is made, but adding a bit of common sense to the entire assumption makes it stronger: the actual angle of refraction is too little in size.

radiation passes it. With a narrow slit, more radiation reaches Dr. Jinn's needles, i.e. with a white slit.

One can do the same thing with light, but it is hard to synchronize all the parts. The best can be done with the following simple construction. Find a small, bright light, say an incandescent bulb in a street light far away in the direction of the sun in a curved automobile lamp. Then put two fingers in front of the lens, so as to break it much the most, and try to see the light source very clearly. You will see that the rays of the light, which were parallel before, becomes quite elongated, and even stretches into a long line. The reason is that the things are very close together, and the light which is supposed to come in a straight line is spread out; it is single, so rays which comes in at the eye it comes from several directions. Also you will notice, if you look very closely, side by side a lot of fringes along the edge line. Furthermore, the whole thing is colored. All of this we've explained in our time, but for the present I want to emphasize that light does not always go in straight lines, and it is not true is very easily demonstrated.

26-4 How it works

Finally, we give a very crude view of what actually happens. *How* the whole thing really works. Let us assume we now believe in an older, quantum, experimentally accurate viewpoint, but of course only qualitatively described. In following the light from A to B in Fig. 26-3, we find that the light does not seem to travel the form of waves of light. Instead, the rays seem to be made up of photons, and they actually produce clicks in a photon counter if we are using one. The total power of the light is proportional to the average number of photons that come in per second, and what we calculate is the chance that a photon goes from A to B, as by hitting the counter. The law for this chance is the following: *law of a wavelet*. Take any point here for the time t_0 that puts there make a complex number, or wave-like complex vector, $e^{i\theta}$, where angle θ is proportional to the time. The number of counts you would get is the frequency of the light. Now, the counter will not have, say, just at a different time, on the same spot it is armed through a calculating logic. The angle θ is always proportional to the time. Take all the wavelets along the path, and to you for each one, then the answer is that the chance of arrival of a photon is proportional to the *square* of the length of the light vector, from the beginning to the end.

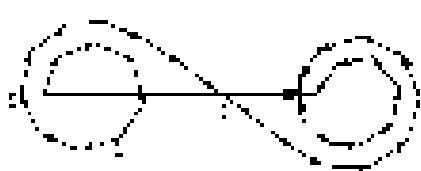


Fig. 26-14. The summation of probability amplitudes for many neighboring paths.

Now let us show how this follows the principle of least time for a mirror. We consider all rays, i.e., possible paths, A-B-C-D-E, etc., in Fig. 26-3. The path A-B-C has a certain *a priori* contribution, but the next path, A-E-B, has a quite different final angle θ quite different, but to say the point C corresponds to minimum time, where if we change the θ to be times of set change. So take while the time derivative, and then they begin to change less and less as we go near point C (Fig. 26-14). So the paths which are having added to contribute to it, the same angle for while near C and then gradually the time lag is increasing again, and the paths go around the other way, and so on. Incidentally, the time lag is light like. The total probability is the sum of them out and to the other, squared. Almost all of these successive probabilities occurs in the same order and the amplitudes in the same direction (as in the same phase). All the contributions from the paths which have very different times we change the point, cancel themselves out by going in different directions. That is why it works. Furthermore, parts of the mirror itself reflects almost exactly. This is, because all we did was to to put a piece of the diaphragm inside the spiral ends, and that makes only a very small change in the light, so this is the situation between the aluminum plates. The paths will a probability, which depending on an accumulation of errors, and the principle of least time.

Geometrical Optics

27-1 Introduction

In this chapter we shall discuss some elementary applications of the ideas of the previous chapter to a number of practical devices, using the approximation called geometrical optics. This is a most useful approximation in the practical design of many optical systems and instruments. Geometrical optics is often very complicated, so we are content to consider it either when it is perfectly good, or when it is good enough so that we can design instruments roughly, using rules that we can easily meet that with failure not at all, since they are practically of high school level. Of course, if we want to know about the small errors of lenses and similar devices, the subject goes on considerably; that is, we also must go beyond geometrical optics. There has been a great deal written in the last few years about the subject of aberrations, and it is advised to read about the subject in the simple way that the rays travel through various surfaces. Let us begin with the book tells how to do, using the law of refraction from one surface to another, and so find out where they come out, say, "After four successive lenses." People have said that this is too tedious, but today, with computing machines, it is the right way to do it. One can set up the problem and make the machine do the work, after which it is very easy. So the subject is really ultimately quite simple, and involves no new principles. Furthermore, in addition to the rules of thin lenses, there is a general rule for lenses that is characteristic of other fields, and that is the reason given to follow the subject very far, with me in constant agreement.

The most advanced and the best theory of geometrical optics was worked out by Huygen, and it turns out that this has very important applications in mechanics. It is extremely important, however, to remember that it is in optics that we use Huygen's theory for the subject of diffraction and diffraction gratings, which is studied in the second year or at graduate school. So, representing that geometrical optics contributes very little toward its own sake, we now go on directly to the geometric properties of simple optical systems on the basis of the principles outlined in the last chapter.

In order to go on, we must have some geometrical formulas, which are the following: (1) the real image will be smaller than the object, (2) if it is real, then the magnification will be negative, (3) the difference in size between two different images is proportional to the distance d (Fig. 27-1). These conclusions? The difference $s' - s = d'$ can be found in a number of ways. One way is this. We see that $s'^2 = d'^2 + \Delta^2$, or $(s' - d)^2 = d'^2$. But $s' - d = d' + d - d$. Thus

$$\Delta \propto d^{3/2} \quad (27-1)$$

This is all the geometry we need to compute the magnification of images by curved surfaces!

27-2 The focal length of a spherical surface

The first and simplest situation to discuss is a single refraction in glass, separation, two media with refractive index n_1 and n_2 (Fig. 27-2). We can use the case of arbitrary indices of refraction to the system, because when we always the most important thing is the specific situation, and the problem is very simple to do in this case. Suppose, suppose the n_1 on the left, the n_2 is 1 and on the right it is $1/n$, where n is the index of refraction. The light travel's more slowly in the glass by a factor n .

27-3 Introduction

- 27-2 The focal length of a spherical surface
- 27-3 The focal length of a lens
- 27-4 Magnification
- 27-5 Compound lenses
- 27-6 Aberrations
- 27-7 Working power

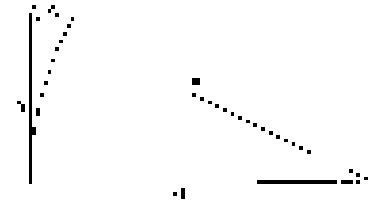


Figure 27-1

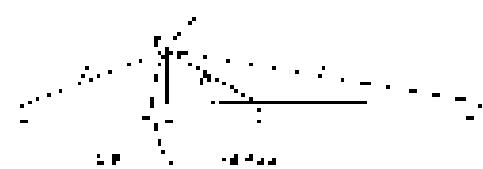


Figure 27-2. Refraction by a single refracting surface.

Now suppose that we have a point at O , at a distance a from the front surface of the glass, and another point O' at a distance b inside the glass, and we desire to design the curved surface in such a manner that, even ray from O which hits the surface, at any point P , will be bent so as to pass through the point O' . For that to be true, we have to design the surface in such a way that the time taken for the ray to go from O to O' , that is, the distance $OP + O'P$ divided by the speed of light (the speed being c units), plus $c \cdot OP$, which is the time it takes to go from P to O' , is equal to a constant independent of the point P . This condition applies however only for determining the surface. The answer is that the surface is a very complicated function of x and y , and that a student may entertain himself by trying to calculate it by any other geometry. It is simpler to try a graphical construction, namely to $c = \infty$, because then the curve is a two-dimensional circle and is more recognizable. It is interesting to compare this curve with the parabolic curve we found for a focusing mirror when the light is coming from infinity.

By the proper surface cannot easily be made to focus the light from one point in another, requires a rather complicated surface. It turns out in practice that we do not try to make such complicated surfaces actually, but instead we make a compromise. Instead of trying to get all the rays to come to a focus, we suppose that only the rays fairly close to the axis OO' come to focus. The farther ones may deviate if they want to, under any circumstances the surface is complicated, and we use instead a spherical surface with the light curvature in the axis. It is much easier to calculate a sphere than other surfaces, than it is preferable for us to find out what happens to rays striking a spherical surface, supposing that only the ones near the axis are going to be focused perfectly. These rays of distance near the axis are sometimes called *convergent rays*, and what we are analyzing are the conditions for the focusing of parallel rays. We shall discuss after the analysis for this interested by the fact that the rays are not always close to the axis.

Thus, supposing P lies on OO' and we drop a perpendicular PQ such that the height PQ is a . For a moment, we imagine that the surface is a plane passing through O . In that case, the time needed to go from O to O' would exceed the time from Q to O' , and also, the time from P to O' would exceed the time from Q to O' . But that is why the glass must be curved, because the total excess time must be compensated by the delay in passing from P to Q . Now the excess time along route QQ' is a^2/c^2 , and the excess time on the other route is (b^2/c^2) . This excess time, which must be covered by the delay in going from PQ , differs from what it would have been in a vacuum, because there is a "medium" present. In other words, the time to go from P to Q is less as if it were straight in the air, but it is slower by the factor n , so the excess delay is n times greater than a^2/c^2 . And how large is n ? If the point O' is the center of the sphere and its radius is R , we see by the same formula that the distance PQ is equal to $b^2/2R$. Therefore we discover that the law $(n - 1)ab^2/2R = a^2/c^2$, and that gives us the radius of curvature R of the surface that we need.

$$(n^2/2) + (nv^2/c^2) = (n - 1)b^2/2R \quad (27.2)$$

or

$$C/n = C/v = (v - 1)c/b. \quad (27.3)$$

If we have a position O and another position O' , and want to focus light from O to O' , then we can calculate the required radius of curvature R of the surface by this formula.

Now it turns out interestingly that the same lens with the same curvature R , will focus for other distances, namely, for any pair of distances such that the sum of the two distances, multiplied by n , is a constant. This is a given fact, will you keep it we limit ourselves to assume rays from infinity from $O \approx O'$, but between an infinite number of other pairs of points, so long as those pairs of points have the characteristic $b_1/n + b_2/n$ is a constant, characteristic of the lens.

In particular, an interesting case is that in which $b \rightarrow a$. We can see from the formula that as one a increases, the lens decreases. In other words, if point O is

goes on to point O' on medium 2 and vice versa. As point O goes toward infinity, point O' keeps moving to the left until it reaches a certain distance, called the focal length, f , from the medium. If parallel rays come in, they will meet at a distance f . Likewise, we could imagine it the other way, if one takes the reciprocity rule ("plus will go from O to O'), of course now the point O' is to O . Therefore, if we have a light source inside the glass, we might want to know where the focus is. In particular, if the light in the glass were an infinity (some problem which would become a focus outside). This distance is called f . Of course, we can do it the other way. If we had a ray source outside the glass, went through the surface, does it break or not as a parallel beam? We can easily find out what f is:

$$n_1 n_2 = (n - 1)/R \quad \text{or} \quad n^2 + R(n - 1) = 1. \quad (27-2)$$

$$1/f = (n - 1)/R \quad \text{or} \quad f = R/(n - 1). \quad (27-3)$$

We see an interesting thing: we divide focal length by the corresponding index of refraction, we get the same result. The question, in fact, is general. It is true of any system of lenses, no matter how complicated, that it is inversing its reciprocal. We could prove here that this general—(we merely noted it for a single surface, but it is general)—is true, in general that the two focal lengths of a system are related in this way. Sometimes Eq. (27-3) is written in the form

$$1/n + 1/n' = 1/f. \quad (27-4)$$

This is more useful than (27-3) because we can move n / n' around so that we can measure the successive indices of refraction of the lens. If we are not interested in inverting a lens... in this is a law, it is the case, but, simply, like all of a field, the interesting quantity is f , not the n 's, the n 's, the R 's, and the d 's.

Now let's take a student exam if it becomes less confusing. What happens there? If $n < 1$, then $n_2 < n_1$, and therefore n' is negative, our equation says that the light will travel only with a negative value of n . Whether this is real? Does mean something very interesting and very difficult. It is still useful to recall, in other words, even when the numbers are negative. What it means is shown in Fig. 27-1. If we draw the rays which are diverging from O , they will be here, if it is true, at the surface, in G . They will not come to a focus, because it is so close at that they are "beyond parallel". However, they diverge as if they had come from a point O' outside the glass. This is an apparent image, sometimes called a virtual image. The image O' in Fig. 27-1 is called a real image. If the light really exists to a point, it is a real image. But if the light appears to be coming from a point, a light can pass different from the actual point, it is a virtual image. So when $n < 1$, we can, negative it, use (27-3) on the other side of the surface, and every thing is $> n$.

Now consider the interesting case where it is equal to infinity, then we have $(1/n) - (1/n') = 0$. In other words, $f = \infty$, which means that if we look from a given medium into a new medium and as a point is in the new medium, it appears to be deeper by a factor n . Likewise, we can use the same equation backwards, so that if we look into a given medium at a distance, that is, d , at the distance outside the glass medium, it will appear as though the light is coming from $n - 1$ to n back (Fig. 27-4). When we look at the bottom of a swimming pool from above, it does not look as deep as really is, by a factor 3/4, which is the reciprocal of the index of refraction of water.

What would you do? Assume to use the spherical mirror. But if one expects me to do it, I would be asked to work it out for himself. Therefore we leave it to the student to work out the formula for the spherical mirror, but we mention that it is well to adopt certain conventions concerning the distances involved:

- (1) The object distance s is positive if the point O is to the left of the surface;
- (2) the image distance s' is positive if the point O' is to the right of the surface;
- (3) The value of curvature R of the surface is positive if the surface is to the left of the surface.

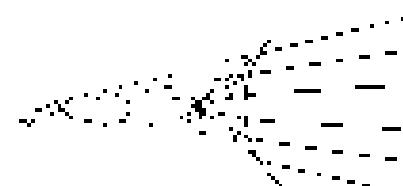


Fig. 27-1. A virtual image

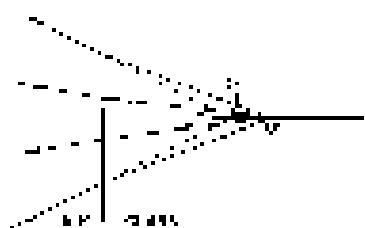


Fig. 27-4. A plane surface image. The light from O to O' .

In Fig. 27-2, for example, n_1 , n_2 , and R are all positive; in Fig. 27-3, n_1 and R are positive, but n_2 is negative. If we had used a concave surface, our formula (27.1) would still give the same result if we merely make it a negative quantity.

To get the right sign for the corresponding formula for a mirror, using the above conventions, you will find that from $p = \frac{R}{n_1 - n_2}$ — I thought the formula (27.1) (as though n_1 were in front of the mirror, had an index -1) the right sign is also obtained.

Although the derivation of Eqs. 27.1 to 27.3 is simple and elegant, using first light rays, you can also work out the static law, along the Snell's law, remembering that the angles are so small that the sines of angles can be replaced by the angles themselves.

27-3 The focal length of a lens.

Now we go on to consider another situation, a very practical one. Most of the lenses that we use have two surfaces and just one. How does this affect matters? Suppose that we have two surfaces of the sort shown in Fig. 27-4, with air filling the space between them (Fig. 27-5). We want to study the problem of focusing from a point O on one side to another point O' . However we do this? The easiest, as this. First, use Formula (27.1) for the first surface, forgetting about the second surface. This will tell us that the light which was emerging from O will appear to be converging or diverging, depending on the sign, from some other point than O' . Now we consider a new problem. We have a second surface, between O' and O , in which rays are converging toward a certain point O' . When will they actually come to O' ? We are in again in trouble again. We find that they converge at O'' . Then, if necessary, we can get through $75 \times$ lenses by just using the same formula in succession, but there is no need!

There are some other high-class techniques that would solving problems like this in the few cases in our lives that we might have to chase the light through ten surfaces, but it is better just to think it through the surfaces when the problem arises, or it is better to use a lot of common sense because it may be we will never have to pass through any surfaces at all!

In any case, the principle is this: when we go through one surface, we find a new position, a new focal point, and then take this point as the starting point for the next surface, and so on. In order to solve problems, since on the second surface we are going from n_1 to n_2 in air rather than from 1 to n_1 , and since in many systems there is more than one kind of glass, so that there are indices n_1 , n_2 , . . . , we usually need a generalization of equation (27.3) for a case where there are two different indices, n_1 and n_2 , rather than only n . Then it is not difficult to prove that the general form of (27.3) is

$$(n_1/n_2) = (n_2/n_1) = (R_2 - R_1)/R. \quad (27.7)$$

The really simple is the special case in which the two surfaces are very close together—so close that we may ignore errors due to the thickness. If we draw the lens as it is in Fig. 27-6, we may ask a question: How fast the lens to be built so as to focus light from O at O' ? Suppose the light comes exactly to the edge of the lens, at point P . Then the excess time in going from O to O' is $(n_1 - n_2) d / c$, if d is the thickness of the lens. Ignoring the momentary presence of the thickness d of glass in front of O , how we take the time for the direct path equal to that for the path $O P O'$? We have to let a piece of glass whose thickness $= d$ be thicker so that the light interested in going through this piece of glass is enough to compensate for the excess time above. Then from the thickness of the lens at the center must be given by the relationship

$$(n_1 d / c) = (n_2 d / c) = (n_2 - n_1) R. \quad (27.8)$$

We can then express the radius of the lens R , and R_2 of the two surfaces. Paying attention to one convention (by taking n_1 to be $n_2 < n_1$ for convex lens),

$$R = (n_1 / 2n_2) + (n_2^2 / 2n_1). \quad (27.9)$$

The above, we finally get

$$(n_1/v) - (n_2/v) = (n_1/n_2)R = \gamma/R_{12} \quad (27.16)$$

Now we can again say that if one of the points is at infinity, the other will be at a point which we will call "the focal length". The focal length is given by

$$1/f = (n_1 - n_2)/R_{12} = \gamma/R_{12} \quad (27.17)$$

where $\gamma = n_2/n_1$.

Note, if we take the *vacuum* case, where n_2 goes to infinity, we see that f is the "focal length" γ^{-1} . This time the focal lengths are equal. (This is another special case of the general rule that the ratio of the two focal lengths is the ratio of the indices of refraction in the two media in which the rays travel.) In the particular optical system, the initial and final media are the same, so the two focal lengths are equal.)

To get the focal length of the second medium, if we bring x_2 over from the *vacuum* region with its radius of curvature and a central index, we could measure the focal length f_2 , by seeing where a point at infinity forms. However, if the focal length f is to be here, we'd have to write our equation in terms of the focal length directly, and this demands the \rightarrow

$$(1/v) - (1/v') = 1/f \quad (27.18)$$

Now let's examine the formulae we've had up to now in different situations. Here it implies that v' is infinite (infinite distance from the lens). This means that $v_2 < R$. Light comes to a distance, and this is *not* defocused. Another interesting thing it says is that light passing a lens in the same direction it came from is defocused. However, if the lens is reversed, And in this I mean that $v_2 & v'$ are equal (they are both equal to R). In other words, if we want a *symmetric* situation, we find that they will both focus at a distance $2R$.

27.4 Magnification

So far we have discussed the forming of real images of points on the axis. Now let's discuss also the forming of *virtual* images away from the axis, but a little at all, so that we can understand the properties of magnification. When we sit right back so as to see a bright filament, e.g., filament under a "pocket" lamp or a screen, we know that on the screen we get a "picture" of the same filament, because of a larger α and it is brighter than the filament. We must remember that the light comes to a focus from each point of the filament. It makes no difference that a lamp is brighter, for us and yet the other less intense radiation is symmetrically. In Fig. 27.7, we come up the following facts:

- (1) A ray that comes in parallel from a point parallel toward a convergent lens will meet at the focus F on the other side of a "ray" from the lens.
- (2) A ray that passes F before it from the lens on one side comes out parallel to the axis on the other side.

This is all we need to establish from us (27.19) by symmetry, is (27.20). Suppose we have an object, a same distance a from the lens; let b be height of the object, α , i.e., then we know that one of the rays, in fact PQ , can be bent so as to pass through the focus F on the other side. Now if a lens will focus just 1.5 or all, we can find out where it will find out where just one converging ray, because the new rays will be when the lens focuses, α , i.e., b . We could try use our ingenuity to find the exact direction of one other ray. If we remember that a parallel ray goes through the lens and the lens does not change through the "axis of" convergent lenses. So we draw ray AB through F . It's true that the actual rays which are being defocused may be much more limited than the two we have drawn, but they are random in figure, so we make believe that we can assume this ray goes on it would come out straight, we draw it by parallel to AB . The intersection is the point we want. This will determine the exact place and the correct height. Let

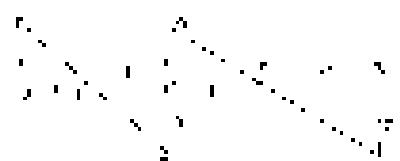


Fig. 27.7. The geometry of focusing by a thin lens.

us call the "focal length" and the distance from the lens, x . Now we may derive a lens formula. Using the similar triangles $\triangle PVY$ and $\triangle XZY$, we find

$$\frac{v}{f} = \frac{x}{x} \quad (27.13)$$

Similarly, from triangles $\triangle WXR$ and $\triangle QYL$, we get

$$\frac{v'}{f'} = \frac{x'}{x} \quad (27.14)$$

Solving each for v and v' , we find that

$$\frac{v}{f} = \frac{v'}{f'} = \frac{x}{x} \quad (27.15)$$

Equation (27.15) is the **focal length formula**. In it is everything we need to know about lenses. It tells us the magnification, in Eq. (27.15), in terms of the object size and the focal lengths. It also indicates the two distances x and x' which

$$xx' = ff' \quad (27.16)$$

which is a much neater form of work over than Eq. (27.13). We leave it to the student to demonstrate that if we let $v = x + f$ and $v' = x' + f'$, Eq. (27.15) is the same as Eq. (27.16).

27-4 Compound lenses

Without actually deriving it, we shall briefly describe the general result when we have a number of lenses. If we have a system of several lenses, how can we possibly analyze it? That is easy. We start with some source and calculate where the first lens, using formula (27.16) or (27.15) or any other calculation formula, or by drawing diagrams. So we find an image. Then we call this image as the source for the next lens, and use the second lens with whatever its focal length; so to again find an image. We simply repeat the foregoing for the successive lenses. That is all there is to it. It is nothing new in principle, we just multiply it out. However, there is a very interesting net result of the effects of any sequence of lenses on light that starts and ends up in the same medium, say air. Any optical instrument—a telescope or a microscope with lots and lots of lenses and mirrors—has the following property: There exist two planes, called the **principal planes** of the system (these do not have to be thin lenses), which have the following properties: (1) If light comes in ... to a system parallel from the left side, it comes out at a certain distance x in front of the second principal plane equal to the focal length; just as though the system were a thin lens located at this plane. (2) If $x > 0$ light comes in the other way, it exits to a focus at the same distance x from the first principal plane, again as though it were there (see Fig. 27.8).

Of course, if we measure the distances x and x' , and y and y' as before, the formula (27.16) has to be modified to take into account the fact that this lens is **assolutely** *thin*—provided that we measure the focal length from the principal planes and not from the center of the lens. It so happens that for a thin lens the principal planes are coincident. It is just as though we could take a thin lens, slice it down the middle, and separate it into two parts. Let it *not* separate. But say that one is *pop-out* immediately on the right side of the second plane from the same point as . . . see also the first plane! The principal planes and the focal lengths must be found either by experiment or by calculation, and then the whole set of properties of the optical system are determined. It is very interesting that this is not too complicated when we deal directly with such a thin, uncomplicated optical system.

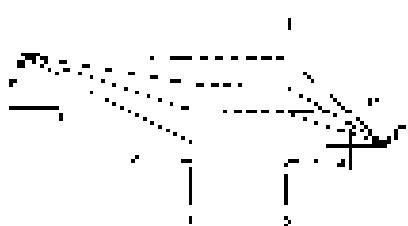


Fig. 27-8. Illustration of the principal planes of an optical system.

27-6 Aberrations

Before we get into what could now marvelous lenses are, we must return to how the different lenses interact with light. Because of the fact that we focus lenses ourselves, gravity causing a point's light rays to converge near the lens. A lens, thus having a finite size will, in general, exhibit aberrations. For example, if the lens is on the side, of course, given enough time, a ray that is parallel to the axis will still converge to the focus very well. But if the lens is inclined, the ray begins to deviate from the focus, perhaps by hitting itself, and a ray striking near the top edge comes down and misses the focus by a few millimeters. So instead of getting a point image, we get a smear. This effect is called *spherical aberration*, because it is a property of the spherical surfaces we used placed in the right shape. This could be remedied, for any specific object distance, by tailoring the shape of lenses to focus on a single point by using several lenses arranged so that the aberrations of the individual lenses tend to cancel each other.

Lenses have another fun type of difference, where two different speeds or different indices of refraction, in the glass, and therefore the focal length of a given lens is different. This is refraction. So if we move away its focal, the images will have colors, because after we focus for one color, the others will be out of focus or vice versa. This is *chromatic aberration*.

There is yet another fun type of aberration that occurs if the focus is off the axis. You'll perfect any more, unless it gets far enough off the axis. The easiest way to verify this is to take a lens and then to see that the rays are bending in at a large angle from the axis. Then the image that is formed will usually be quite blurry and there may be no place where it focuses well. There are thus several kinds of aberrations other than the optical design can't handle by us up more focus to compensate each other's effects.

Now, what if we have to be *geometrically* aberrant? Is it possible to make an absolutely perfect optical system? Suppose we had a multi-lens optical system that is supposed to bring light exactly to a point. Now, say my friend has paid all this effort and can get this lens. Given all how perfect this system has to be, the system can have some kind of an inherent tendency to the light. If we take the first ray from the lens that can escape to the side if the system is perfect (if you like), the first four rays are exactly equal. But nothing is perfect, so the question is, how wrong can the light be for the ray and not be worth calling any "other"? That depends on how perfect we want to make the system. Your supplier can work to make the rays as perfect as is possibly can be made. Then, of course, our impression is that we have to compromise, but many rays follow exactly the same time as possible. But it turns out that this is not true, that beyond a certain point we are trying to do something that is terrible because the theory of geometrical optics does not work!

Remember: that the principle of least time is not an *exact* formulation, under the principle of conservation of energy or the principle of conservation of momentum. The principle of least time is only an approximation, and it is exacting... but there must be a limit that is allowed and will not affect any apparent difference. The answer is that if we ever arranged them later in the system, or, worse yet, if one that is further out, and the central ray, the difference in time is less than about the period of time it spends in the oscillation of the light. Otherwise there is no use arranging it any further. Light is an oscillatory thing with a definite frequency that is related to the wave length, and if we have arranged that the time difference for different rays is less than about a period, there is no use arranging any farther.

27-7 Resolution power

Another interesting question is how important technical question with all cameras, instruments—how much we're going to see, how sharp. If we build a microscope, we want to see the objects that we are looking at. That means, for instance, that if we are looking at a bacterium with a spot on each end, we want to know just

Fig. 12-2 illustrates what we imagined the ... Our singer thus, that all we hope to do is to get enough magnification—well, we always add another lens, and we can always magnify again and again, and with the increases of magnification, all the spherical aberrations and chromatic aberrations can be decreased in ... and there is no reason why we can't keep on magnifying the image. So the limitations of a microscope are not ... it is impossible to build a lens that magnifies every colorchromatic. We can build a system of lenses that magnifies by 100 diameters, but we still could not have one point that are two places beyond because of the limitations of geometrical optics, because of the fact that lens time is not precise.

So suppose we take that telescope now, and that two points have to be on either side of the focal length, so separate points can be seen in a very beautiful way, and we know that the time is taken for each to be on focus. Suppose that we imagined the aberrations now, and I take the faraway object point P (Fig. 12-9); ... the rays from here to focus T_1 ... really are stuck to ... it is not T_1 , because it is not a perfect system, but that is another problem. Now take another nearby point, P' , and ask whether, coming up with the aberration from T_1 . In other words, whether we can tell the difference between them. Of course, according to geometrical optics, there should be two point images, but we can see more than the increased and we may not be able to make out that there are two points. But consider that the second point is different in a distance different from from the first one is that the two times for the two end rays $P-T_1$ and $P'-T_1$ on each side of the lens focusing on the lens to go from one end to the other, must not be equal, from the two possible object points to a given image point. Why? Because, if the times were equal, if we were born Δt off focus at the same point, so the times would going to be equal. But he has much more they were to take so that we can say that is Δt the time to be a constant form, so that we can distinguish the two images, perhaps? The general rule for the resolution of any optical instrument is this: two different point sources can be resolved only if one source is beyond or just a point than the time for the ray from the other source to reach that point, as compared with the time for the ray from the other source to reach that point. It is necessary that the difference in time between the ray from the left-hand ray to the ray from the right exceed a certain minimum number of periods of oscillation of the light.

$$t_2 - t_1 \geq \frac{1}{c} \Delta \phi \quad (12-1)$$

where c is the frequency of the light (number of oscillations per second), expressed in cycles per second; $t_2 - t_1$ is the difference of times of travel of 100,000 points, is called D , and if the ray-angle of the rays is called θ , then one can demonstrate that $(D/c)\theta$ is exactly equivalent to the statement that D must exceed $\lambda/\Delta n$, where Δn is the index of refraction of air and λ is the wavelength. The simplest thing that we can see is that λ is approximately the wavelength of light. A corresponding formula exists for telescopes, which $\Delta \phi$ is the smallest difference in angle between two stars that can just be distinguished.

* The angle is called the "angle D in radians between conjugate rays."

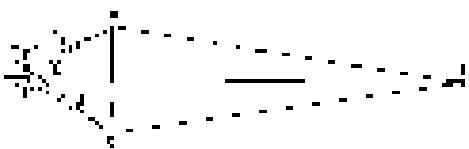


Fig. 12-9. The resolving power of an optical system.

Electromagnetic Radiation

24-1 Electromagnetism

The most Grammatical mistake in the development of physics can occur in which two synonymous words, where phenomena would previously and accustomed to be different are suddenly discovered to be the same. To put it simply, the same thing. That's what Grammatical mistake is. The history of such synthesis, and the basis of the science of physical science is mainly the same as above by synthesis.

Besides the importance of symmetry in the development of basic science, the EMR must also occur. In J. C. Maxwell's time may not be known about the mutual and inverse laws of electricity and magnetism with the laws of the behavior of light. As a result, the properties of light were poorly understood. But still all of the stuff that is so important and interesting was that it was felt necessary to invent a special creation for it in other writing. Grammatical mistake was that he was forced to do this because "Light is electricity and magnetism, and there is light."

For this eliminating moment, he wrote a long paperwork in his private laboratory, containing various laws of electricity and magnetism. This story we shall reserve for detailed study next year. However, the story is, suffice to say, the gradually disappeared of preexisting electricity and magnetism, or electric forces of attraction and repulsion, and of magnetic forces showed that, although these forces were rather complex, they all fell on symmetry as the signs of their existence. We came, for example, but I assume easily, how a stationary charge gives us the electric force field which is inversely proportional to the square of distance. As a consequence, at sufficiently great distances there is very little influence. Thus system of charges is consistent. Maxwell noted that the equations of the laws that had been discovered up to this time were mutually inconsistent. When he tried to put all of them together and in order for the whole system to be consistent, he had to add another term in his equations. With this term, the system became an analog credibility, which was that if the electric and magnetic fields would fall off much more slowly with the distance than the forces do, more rapidly, apparently as the first power of the distance. And so he realized that electric force is to do that can affect other charges far away, also be present of the back effects with which we are familiar, like induction, polarization, and so on.

In summary, it turned that a charge will fly in Europe even, with zero acceleration we'll be heard thousands of miles away in New York. That's it's possible to have a field that not only is not a source, but only precisely as the last center of the source. Finally, that's our light that was recognized to be electric and magnetic in human habitation over vast distances guaranteed by an almost in reality a full visibility of the electrons in the world. All these phenomena we summarize by the word *radiation*, more specifically *electromagnetic radiation*. There being one or two other kinds of radiation also. A man always radiates electromagnetic waves to other

A. G. Thus is the universe *spun together* — the atomic nucleus of a star, that will have sufficient influence at the great distance to set the electrons in motion in myself, and never know about the stars. If this law had not exist, we would not be literally in the dark about the outer world. And the electron comes in a galaxy two billion light years away, which is the farthest object we can see from so far — can still influence in a significant and detectable way the current in the great "Giant" in front of a radio telescope. And so now they work. Listen and the problem

24-1 Electromagnetism

24-2 Radiation

24-3 The Radio Telescope

24-4 Interference

This remarkable phenomenon is what we shall discuss in the present chapter. At the beginning of this course in physics we outlined a broad picture of the world, but we are now better prepared to understand some aspects of it, and so we shall now go over some parts of it again in greater detail. We begin by describing the position of physics at the end of the 19th century. All that was then known about the fundamental laws can be summarized as follows.

First there were laws of forces: one force was the law of gravitation, which we have written down several times: the force on an object of mass m due to another of mass M is given by

$$\mathbf{F} = G m M \frac{\mathbf{r}}{r^3}, \quad (25.1)$$

where \mathbf{r} is a unit vector directed from mass M , and r is the distance between them.

Next, the laws of elasticity and magnetism, as known at the end of the last century, are that: (i) electrical forces acting on a charge q can be described by two fields, called \mathbf{E} and \mathbf{B} , and the velocity v of the charge q by the equation

$$q v \times \mathbf{B} = q(\mathbf{E} + v \times \mathbf{B}), \quad (25.2)$$

To complete this law, we have to say what the formulas for \mathbf{E} and \mathbf{B} are in a given circumstance: if a number of charges are fixed, \mathbf{E} and \mathbf{B} are each the sum of contributions from each individual charge. So if we can find the \mathbf{E} and \mathbf{B} produced by a single charge, we must simply add all these contributions! If no charges in the universe except the field \mathbf{E} and \mathbf{B} . This is the principle of superposition.

What is the formula for the electric and magnetic field produced by one individual charge? It turns out that this is very complicated, and it takes great ingenuity and simplification to appreciate it. But that is not the point. We write down below now only the formula, the reader will be happy to note, and say: "A-ha, that is possibly to summarize all the fundamental knowledge in one page, with no details that he is now familiar with." This law for the field of an individual charge is incomplete and inaccurate; we know (except for quantum mechanics) but it looks rather complicated. We shall not quote all the pieces now; we only wish it shown to give an impression, to show that it can be written, and so that we can see instead of the roughly what's known now. As a matter of fact, the most useful way to write these two laws of electricity and magnetism is not the way we shall now write them, but involves what we call 'field operators', which we shall learn about next year. But the mathematical problems for these are different now, and so we write below in a convenient form for calculation, but in approximations that we now know.

The electric field, \mathbf{E} , is given by

$$\mathbf{E} = \frac{-q}{4\pi\epsilon_0} \left[\frac{\mathbf{r}_P}{r^3} + \frac{v}{c} \frac{d}{dt} \left(\frac{\mathbf{r}_P}{r^3} \right) - \frac{1}{c^2} \frac{d^2}{dt^2} \frac{\mathbf{r}_P}{r^3} \right]. \quad (25.3)$$

What do the various terms mean? Take the first term, $T_1 = -q\epsilon_0/c^2 r^3 \mathbf{r}_P$. That, of course, is Coulomb's law, which we already know: q is the charge that is producing the field, \mathbf{r}_P is the unit vector in the direction from the point P where \mathbf{E} is measured, r is the distance from P to q . But Coulomb's law is wrong. The discoveries of the 19th century showed that influences cannot travel fast. There is a non-instantaneous speed c , which we now call the speed of light. T_1 is incorrect just like Earth is Coulomb's law, not only because it is not possible to know where the charge is now, and so what distance it is now, but also because it is only thing that can affect the field at a given place and time is the behavior of the charge in the past. How far in the past? The time delay, or retardation, associated with the time it takes, at speed c , to get from the charge to the field point P . The delay is r/c .

In to allow for this time delay, we put $+1/c$ in front of T_1 , meaning how far away it was when the information now arriving at P left q . But to remember suppose that the charge carries a light and that the light can only move to P at the speed c . Then when we look at q , we wouldn't see where it is now, of course, but where it was at some earlier time. What appears in our T_1 must be the retarded

direction to the direction it used to be—the sense of retarded direction—and you'll understand distance r . That would be easy enough to understand, yes, but it is also wrong. The whole thing is much more complicated.

There are several more terms. The next term is ω , though we are were trying to allow for retardation. But the effect is retarded, if we neglect ω , it is very crude. In addition there should be added the dipole moment D and add a correction term, which is to take el. charge times the time delay that we use. Nature seems to be stamping in to pass what the field is, the present time is going to be, by taking the rate of change and multiplying by the time that is ω delayed. But we are not yet through. There is a third term—the second derivative, with respect ω , of the unit vector in the direction of the dipole. Now the dipole is finite, and that is all there is to the electric field from an arbitrary moving charge.

The magnetic field is given by

$$\mathbf{B} = \mathbf{B}_0 \propto \mathbf{E}/c. \quad (19.3)$$

We have written these down only for the purpose of showing the beauty of nature even in a way that goes beyond mathematics. We do not pretend to understand why it is possible to write so much in such a small space. In (18.3) and (19.3) contain the machinery by which electric generators work, how light carries all the phenomena of electricity and magnetism. Of course, to complete the story we do need to know something about the behavior of the materials involved, the properties of matter—which we did described briefly by (18.3).

To finish with our description of the world of the 19th century we must mention one other great synthesis which occurred in that century, one with which Maxwell was a great deal involved and that was the synthesis of the phenomena of light and mechanics. We shall study this subject soon.

What had to be added at the 20th century was that the dynamical laws of Newton were found to be not enough, and quantum mechanics had to be introduced to correct them. Newton's laws are approximately valid when the scale of things is sufficiently large. The quantum mechanics laws, combined with the laws of electricity, have only recently been combined to form a set of laws called quantum electrodynamics. In addition, there was discovered a number of new phenomena, of which the first was radioactivity, discovered by Becquerel in 1896—but just accepted it in 1900—the last century. This phenomenon of radioactivity was followed up to produce our knowledge of the atomic nuclei of atoms (i.e., the nucleus), and not electrons, but new particles were different interactions. Doubtless others have still to be discovered.

For those people who know more than you do, who happen to be reading this, we should add that when we say that (19.3) is a complete description of the knowledge of electrodynamics, we are not being entirely accurate. There was a problem that was not quite solved at the end of the 19th century. When we try to calculate the field from all the charges existing in space (say that we want the field to exist at), we get zero from it by trying to find the distance, for example, of a charge from itself, and looking something by that criterion which is zero. The problem of how to recall the part of the field which is generated by one very charge q which we want the field to exist but yet actual exists. So we leave it here; we do not have a complete solution to this puzzle yet, and so we shall avoid the puzzle for as long as we can.

19-2 Radiation

That last is a summary of the world picture. Now let us go on to discuss the phenomena called radiation. To discuss these phenomena, we must extract from eq. (18.3) only that part which is inversely as the distance and goes to the source of the charges. It turns out that when we find the field of a piece, it is completely sufficient that it's legitimate to study forces and electromagnetism in an elementary way by taking it as "by law" of the form the field produced by a moving charge is \mathbf{E} . We shall take it especially as a given law which we will need to use in this next year.

Of the terms appearing in (28) it is the last one evidently goes inversely as the square of the distance, and the second is only a correction for delay, so it is easy to see that, but of them vary inversely as the size of the charge. All of the terms which are important in the first and third term, which is not very important, after all. What can this say to us at the stage and note the direction of the motion? If we can project the end of \vec{r} on the surface of a unit sphere, as the energy waves proceed, the end \vec{r} never wiggles, and the acceleration of the unit vector \vec{r} is also not changing for. That is all. Thus

$$\vec{E} = \frac{q}{4\pi r^2} \frac{\vec{A}_0}{r^2} \quad (28.5)$$

is a statement of the laws of radiation, because that is the only important law in which we get far enough away that the terms the varying inversely as the distance, (the case that the charges have fallen off or more than we are not interested in here).

Now we can go on with the further at studying (28) by to see what it means. Suppose a charge is moving at the moment whatsoever. And we are observing it from a distance. We imagine it is moving that in a straight line. Let's call that true light that we are trying to explain; we imagine it is a little while, that, then we would see this while doing something. But we don't see exactly how it is moving because right now because of the delay that we have been trying to do what counts is the moving source. The unit vector \vec{r} is pointed toward the apparent position of the charge. Otherwise, the end \vec{r} is, goes on a slight curve, because the acceleration has two components. One is the transverse part, because the end \vec{r} is always going along, and the other is radial, please because it stays on a sphere. It is easy to demonstrate that the latter is much smaller than varies as the inverse square of r when r is very great. This is easy to see, for when we imagine that we move a given source farther and farther away, then the wiggles of \vec{r} , look smaller and smaller, inversely as the distance. But the radial component of acceleration is varying more rapidly than inversely as the distance. So for practical purposes \vec{r} we have to do is project the motion on a plane at unit distance. Therefore we get the following rule. Imagine that we look at the moving charge and that we try to see it delayed, like a projector lamp, up to a screen at a distance r at unit distance. A real projector, of course, does not do two account the fact that light is going at a finite speed, but let us do the world as he sees it. We want to see what the picture would look like. So we see a dot, representing the charge, moving about in the screen. The acceleration of this dot is proportional to the electric field. That is $\vec{F} \rightarrow \vec{r}$ we need.

Thus Eq. (28.5) is the complete and correct formula to remember; even velocity effects are all contained in it. However, we often want to apply it to a still situation where \vec{r} either the charges are moving only a bit. Because it is relatively very rare. Since they are moving slowly, they do not move an appreciable distance from where they start, so that the delay time is practically constant. Then the law is still simple, but not the delay time is "well". Thus we imagine the charge is moving a very tiny motion after effectively constant velocity. The delay at the distance r is r/c . The motion becomes the following: if the charged object is moving it is very small motion and it is linearly displaced by the distance r/c , then through the unit vector \vec{r} , is displaced \vec{r}/c , and since it is practically constant, the exponent of r^2 , r^2/c^2 is simply. A cancellation of a factor c^2 is left, in order time, and so finally we get the law as was, which is

$$E(r) = \frac{-q}{4\pi r^2 c^2} \cdot \left(1 - \frac{r}{c}\right). \quad (28.6)$$

Only the exponent of r , perpendicular to the line of sight is important. Let us see why this is. Essentially, if charge is moving in and out straight to us, the direction of that direction does not wobble at all, and it has no acceleration. So it is only the sideways motion which is important, only the motion in that we see projected on the screen.

28-2 The dipole radiator

As part of fundamental "law" of electromagnetism, we said, we are going to assume that (24.5) is true, i.e., that the electric field produced by an accelerating charge which is moving non-relativistically at a very large distance r , appears as $E = q/r^2$. The electric field varies inversely as r^2 and is zero at $r = \infty$; the acceleration of charge q is passed onto the "light" of eight "fingers" or "fingers" which follow today's acceleration, but the idea, which that it did at an earlier time, does not yet delay being a time, etc. In the remainder of this chapter we do a series of tests so that we can understand it better; maybe you know we are going to use it to understand all of the phenomena of light and wave propagation, such as reflection, refraction, interference, diffraction, and scattering. It is to be noted here, and I'll say now, that if you take (24.5) as you're told only up to $\propto r^{-2}$ the $\propto r^{-2}$ is that we could conjecture were (25.6) fits and how well it does.

We shall discuss (28.3) further now, etc. In the meantime, we shall suppose $\propto r^{-2}$ but not the other two terms, here. We may suppose a number of experiments which illustrate the character of this law. At once, to do so, we need to consider a charge. Let this q be a single charge, but if we can make a pair many charges may be used, all in some way, for example, that the field $\propto 1/r^2$ be the sum of the effects of higher order individual charges; we just add the "fingers". And, again, consider two cases of what happened in a generator, as shown in Fig. 28-1. The idea is this: the generator makes a potential difference, $V = E \cdot d$, which pulls electrons away from wire A and pushes them into wire B, in a current, and then, on discharge, in due time, the electrons repel each other, the electrons out of B and jump back to A. If both these two cases except for the two wires are exactly the same, we can disprove in $\propto r^{-2}$ is the current and the momentary energy and waveform dependent. In Fig. 4 and shown out in wire A. The fact that we find two wires in a generator is merely that there is a bus which splits it. For example, that we actually have one potential difference of V because there are two wires in wire A, single wire. And, if the V is very small compared with the distance light travels in one oscillation period \approx about 10^{-15} seconds, etc. Then, etc. Thus we find the voltage, etc. But, we need to apply our law, which is to this case, get the same thing. That is, to do something in instrument to detect the electric field, and the instrument we use is the static thing, a pair of wires like A and B. If no field is field \propto applied to such a device, it will work just as when $V = 0$ will pull the electrons up on both wires or down on both wires. This should be detected by means of a voltmeter, connected between A and B, and a tiny fine microphone for the information into an amplifier where it is amplified so we can hear the positive frequency note with which the radiofrequency is modulated. When this probe feels the electric field, there will be a cold noise, one might call it, because the sound other than the in-phase field driving it. There will be no noise.

Now consider a room in which the wires we are measuring conductive objects in. An electric field will take effect, even in these other objects. And this field makes the other charge-groups there and in going \propto in proportion, these also push or attract our probe. This last successful experiment was just field \propto field \propto field \propto field. So, let the influence from the walls, floor, windows, etc. in different rooms—since they're small, or the phonometer will not turn until appear to be precisely and perfectly to accord with Eq. (28.5), he will believe enough that we can't believe or appreciate the law.

Now we run the generator and one hears the noise of signal. We find a strong test when the detector, A , is parallel to the generator. That point (Fig. 28-2). We find a certain amount of field also at any other point in the loop about A , a lot of noise, and so on, but no signal at A . On the other hand, when the generator is at B the field is zero. That is, right, because we found with that the field \propto rank be the acceleration of the q 's, or projected perpendicular to the line of sight. Therefore when we took them on C , the charge is moving toward and away from D , and there is no effect. So that's when the C and D in (28.5) there's no effect—the charge is not really moving or, secondly, the forces perpendicular to the q 's should be perpendicular to A and at the point of C and D , on (28.5) we put D at C but rotate it 90° , we would see no signal. And this is just what we find. The

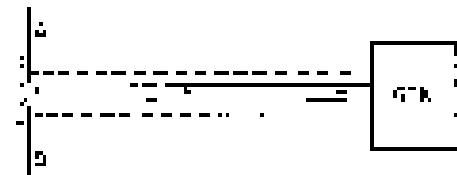


Fig. 28-1. A high-frequency signal source, or other charge up and down in two wires.

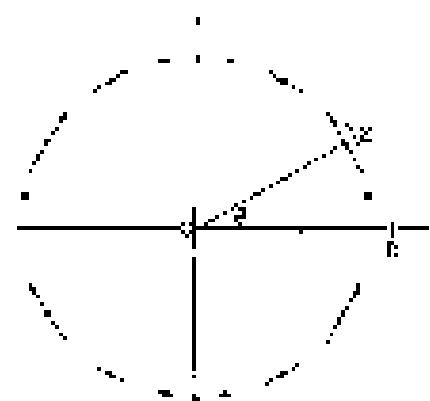


Fig. 28-2. The instantaneous electric field in a closed circuit of a localized, linearly accelerating charge.

electric field is added vertically, and so he is out of phase. When we move S_2 to some intermediate angle, we see the interference again because it is oriented as shown, because although E_2 is vertical, it does not produce a field that is simply parallel to E_1 . It is the projection of the wavefunction perpendicular to the direction of light that counts. This graph is very bad than Fig. 1, because of the projection effect.

28-4 Interference

Now we may ask what happens when we have two sources side by side a few centimeters apart (Fig. 28-3). The law is that the two sources should add their effects in phase. I mean both of the sources are connected to the same generator and are both moving up and down; the source S_1 has, so that the total electric field is the sum of the two fields twice as big as it was before.

Now comes another interesting possibility. Suppose we make the charges in S_1 and S_2 both oscillate up and down, but with the timing of S_2 so that they are 180° out of phase. Then the field produced by S_2 will be in the opposite direction and the field produced by S_1 will be in the opposite direction, but they cancel each other out so we should get no effect at point P . The phase of oscillation is easily adjustable by means of a plug which is carrying the signal to S_2 . By changing the length of the pipe we change the time it takes the signal to arrive at S_2 , and thus we change the phase of that oscillation. By adjusting it, it is easy to cancel indeed the place where there is no more signal left, in spite of the fact that both S_1 and S_2 are moving. The fact that they are both moving can be checked because, "over there out, we can see the motion of the orbit." So the two of them together can produce zero if everything is adjusted correctly.

Now, it is very interesting to show this. We add on all the new fields in the vector addition. We have just checked it for up and down motion, but let's check something else different. First, we restore S_1 to C ; S_2 is in the atmosphere; that is, they are again moving together. But now we turn S_2 through 90°, as shown in Fig. 28-4. Now we should know at point P the sum of two oscillations, one of which is vertical and the other horizontal. The electric field is the vector sum of these two in-phase signals. They are both moving at the same time and go up and zero together, the total field should then be equal to 90°. If you turn S_2 by 90°, maximum value, it should be at about 15°, and not vertical. And if we turn it at right angles to that direction, we should get zero, which is easy to measure. Indeed we observe just such behavior!

Now, how about the "numerical"? How can we determine that? But the signal is constant! We could, with a great deal of equipment, measure the time at which it arrives, but there is another, very simple way. Referring again to Fig. 28-1, suppose that S_1 and S_2 are in phase. They are both moving together, and they produce equal electric fields at point P . But suppose we go to a certain phase θ when S_2 is about 90° and far from S_1 . Then, in accordance with the principle that the acceleration should be balanced by a uniform velocity along θ , the coordinates are not equal; the signal cannot keep in phase. Thus it should be possible to find a position θ at which the distances of P from S_1 and S_2 differ by some amount Δ , in such a manner that there is no net signal. That is, the distance d is to be by distance light goes at one-half the speed of the generator. The time per oscillation, and find a point where the difference is greater by a whole cycle, that is, say, the signal from the last antenna reaches point P with a delay in time that is greater than each of the several periods by just the length of time it takes for the wave to travel to point S_2 once, and we have the condition. This produces at P a signal in phase again. At point S the signal is strong again.

This completes our discussion of the experimental verification of one of the fundamental features of Eq. (30.6). Of course we have not really checked the law according to the electric field strength, or the fact that there is also a magnetic field that goes along with the electric field. To do so would require rather sophisticated techniques that would hardly add much understanding to this point. In any case, we have checked these features that are the two greatest importance for our laser applications, and we shall come back to study some of the subtle properties of electric charges in waves next year.

Interference

29-1 Electromagnetic waves

In this chapter we shall discuss the subject of the passing of waves in electromagnetic fields. We have qualitatively demonstrated that there are transverse and longitudinal oscillation fields in wave motion, and our discussion is based on the field in mathematics, that is, not just qualitatively.

We have already physically analyzed the meaning of formula (28.8) more or less fully, so there are no difficulties to be made about it mathematically. In the first place, if a charge is moving along a straight line, the motion of varying amplitude, the field at some angle θ from the axis of the motion is in a direction normal to the line of sight and in the plane containing both the acceleration and the line of sight (Fig. 29-1). If the motion is uniform, then again the electric field can be computed:

$$\mathbf{E}(t) = -\frac{\omega a(t)}{4\pi\epsilon_0 c^2} \hat{x}(\theta) \sin \theta, \quad (29.1)$$

where $a(t) = a/\omega$ is the constant a at the time $t = \omega/2\pi$, called the retarded value of $a(t)$.

Now it would be interesting to know the nature of the field under different conditions. For this other is interesting, of course, is the behavior of $a(t)$, and to understand it we can take the simplest case, a positive point charge moving uniformly. At the first instant, let's suppose that we start in one position and see how the field there changes with time. The nature of the wave must again be described by field lines. Let us call each describes in space at a given instant. So what we want is a "snapshot" picture which tells us what the field is in different places. Of course it depends upon the acceleration of the charge. Suppose the charge is at rest. And at one particular instant, it was actually standing still, and it suddenly accelerated in some manner, as shown in Fig. 29-2, and then stopped. Then, a little bit later, we measure the field at a distance r ; see that we may expect the field will change, as shown in Fig. 29-3. At each point the field is determined by the acceleration of the charge at an earlier time, the antecedent time, using the delayed a . The field at t , therefore, depends on the retarded acceleration, $a(t)$, at an earlier time. So, in general, Fig. 29-3 is really, in effect, a "reversed" version of the wave equation, in that of course the field has to travel forward in time by a constant wave factor c , which we often take to be unity. This is really just symmetrizing the mathematical behavior of $a(t) = a/\omega$. Evidently if we add a little time to a , we get the same sort of $a(t) = a/\omega$ as we would have if we had only added a little distance, $a = a/2\pi$.

So what do we say if we add a little a to $a/2\pi$, or the retardation $a/\sqrt{1 + a^2/2\pi^2}$? Larger value by exactly a units? $a = a/2\pi$? That is, six times greater? The answer is no, because we are not concerned from the source. That is the lesson why we say only light is passing; it is not true. This is equivalent to saying that the field is delayed, or to saying that the electric field is moving in space as time goes on.

An interesting special case is that where the charge q is moving to and from in an oscillatory manner. The case which we will consider initially is the z -oscillate, we cut in when the displacement and velocity were equal to zero at time $t = \pi/2\omega$, in magnitude. The oscillation time is π/ω . Then the acceleration is

$$a = \omega^2 q \cos \omega t = \omega q \sin \omega t. \quad (29.2)$$

29-1 Electromagnetic waves

29-2 Energy of radiation

29-3 Standing waves

29-4 Two dipole radiation

29-5 The mathematics of interference

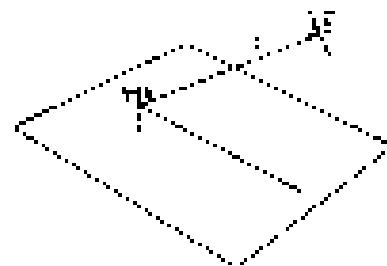


Fig. 29-1. The electric field \mathbf{E} due to a positive charge whose retarded acceleration is a .

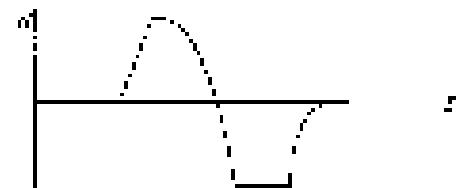


Fig. 29-2. The acceleration of a point charge as a function of time.

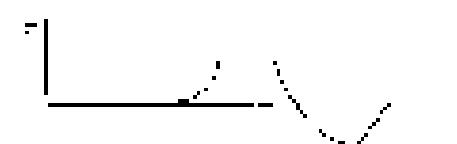


Fig. 29-3. The electric field E as a function of position x at a fixed time. (The y coordinate is ignored.)

where a_0 is the maximum acceleration, $-a^2/k_0$. Putting this formula into (29-1) we find

$$E = -\frac{q^2 a_0}{2 k_0} \sin^2(\theta/2) \cos(\theta/2) = E_0 \sin^2(\theta/2) \quad (29-2)$$

Now, ignoring the angle θ and the constant factors, let us see what just happens. This is a function of position or as a function of time.

29-2 Energy of radiation

First of all, at any particular moment or at one particular place, the strength of the field varies inversely as the distance r , as we have found previously. Now we must point out that the representation of a wave, or the energy effect due to such an effect is "field" can't be *any* proportional to the square of the field, because if it were so, we have some kind of a charge or a cavity or in the steering field, then as we let the field go on, the walls of it makes it move. In this is a linear oscillation. The acceleration, velocity, and displacement, produced by the electric field acting on the charge has all proportionality to the field. So the kinetic energy which is developed in the charge is proportional to the square of the field. So we shall take it that the charge that a field can induce is a quantity proportional somehow to the square of the field.

This means that the energy that the source can deliver decreases as we get further away; it goes down as the square of the distance. But that has a very important preconception if we want to pick up the energy we could from the wave at a certain time at a distance r (Fig. 29-4), and we do the same at another distance r' , we find that the amount of energy we can pick up is proportional to the square inversely as the source of it, i.e., the area of the surface intersected by the cone goes like this as the square of r . So the energy that we can take up while we are within a given spherical angle is the same, no matter how far away we are. In particular, the total energy that we could take out of the whole wave by picking up absorbing oscillators all around is a certain fixed amount. So the fact that the amplitude of E varies as $1/r$ is the same as saying that there is an energy flux which is now $-J_0$, an average value over an area, spreading over a greater and greater effective area. This is not exactly why a charge has oscillated. It has lost some energy which it can never recover; the energy keeps going, but has and has energy within a dimension. So if we are far enough away than our basic approximation is good enough, the charge is still carrying the energy which has been, say, radiated away. Of course the energy still exists somewhere, and is available to be picked up by other systems. We shall study this energy more together in Chapter 32.

Let us now consider more carefully how the wave (29-3) varies as a function of time at a given place, and as a function of position at a given time. Again, we ignore the $1/r$ variation and the constants.

29-3 Sinusoidal waves

First let us fix the position x , and watch the field as a function of time. It is noteworthy at the angular frequency ω . The angular frequency ω can be defined as the rate of change of phase (radians per second). We have already called such a thing, so it should be quite familiar to us by now. The period is the time needed for one oscillation, one complete cycle, and we have worked that out too; it is $2\pi/\omega$. Because a full one period is one cycle of the oscillation.

Now we introduce a new quantity which is used largely in the physics. This has to do with the opposite situation, in which we instead look at the wave as a function of distance x . If you are outside the field, and look inside the wave, (29-3) is the best story. That is, since from Eq. (29-3) we are ignoring the $1/r$ effect, if we change the position x , the intensity with x , we can define a quantity called the phase, represented by kx . This is defined as the rate of change of phase and distance (radians per meter). That is, as we move in space at a fixed time, the phase changes.

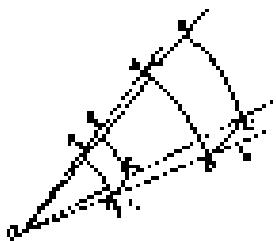


Fig. 29-4. The energy flowing within the zone $OABCD$ is independent of the distance r at which it is measured.

There is another way to state the meaning of the period, and we might say it in the present in space. But it is usually easier to remember, synthesized as it is, to associate both to the distance occupied by one complete cycle. It is easy to see that, for the wavelength λ , the time t is λ/v , because k times t would be the number of waves at that location, multiplying the speed, being the product of the rate of change of the waves per meter, gives the number of cycles, and $k\lambda = v$, and $v = \lambda f$, so $t = \lambda/v = \lambda/f$. It is equally easy to see that $\lambda = v/t$.

Now we turn to a truly wave picture of a radio relationship between the frequency and the wavelength, but the above definition of wave length is really quite general. Let us assume enough and the frequency may now be related to the same wave at other physical circumstances. However, in our circumstance, the rate of change of phase with distance is easily determined, because if we call $\phi = \omega t + \delta\phi$ the phase, and differentiate (physically) with respect to distance x , the rate of change is $d\phi/dx = k$:

$$\left| \frac{d\phi}{dx} \right| = k = \frac{\omega}{v}. \quad (20.4)$$

This is a more wave-like representation than in time, such as

$\nu = 6.7 \times 10^{14}$ Hz	$\lambda = 4.5 \times 10^{-14}$ m
$\nu = 10^9$ Hz	$\lambda = 3 \times 10^{-9}$ m

Why is λ decreasing so much? It is only long, of course, because we still did not wait for one period but took the waves travelling at the speed v . At $\nu = 10^9$ Hz, it takes about 30 cm of distance before there has been just one wavelength.

In a physics sense the entire concept of light ν is not necessarily related to time in simple way. If we call the distance along the direction of the wave front moving from one point to another λ , and if we consider the frequency ν will be given in general as $\nu = c/\lambda$.

Now that we have introduced the idea of wavelength, let me say something you may already know consciousness in which ν is a key factor. In fact, we recall that the field is made up of several waves, one of which carries universally as a carrier wave which is inversely proportional to the wavelength. This would be one such wave, and it dominates the field near the source, but the two important ones, and the other parts are also dropped away. Naturally, the answer is "if we do the averaging away," because some which carry inversely to the square ultimately become negligible compared with one. However, this being "far enough," it is almost as qualitatively, but the other factors are of order λ/c smaller than the first term. Thus, so long as we are beyond a few wavelengths, ν is a good approximation to the field. Sometimes one might say that a few wavelengths is called the "width of the field."

20-4. Two-dipole radiation

Now let us discuss the most basic method for calculating the effects of two dipoles in space. We find the net field at a given point. This is very easy in the two cases that we considered at the previous chapter. We shall first consider the effects qualitatively, and then quantitatively. To start off the simplest case, where the two dipoles are situated with their centers in the same horizontal plane as the center of the line of the dipole a vertical.

Figure 20-5(a) represents the situation of two such dipoles, a , and b , perpendicular to my page and with a wavelength apart in a 90° position, and are oscillating together at the same rate, which we will call ω . Now we would like to know the intensity of the radiation in various directions. This is relatively easy, since the amount of energy radiated need comes positive per second, which is proportional to the square of the field, according to $E = B/c$. So far I look at, when we want to know how bright the light is, is the square of the electric field, not the electric field itself. (The electric field tells the strength of the field but does

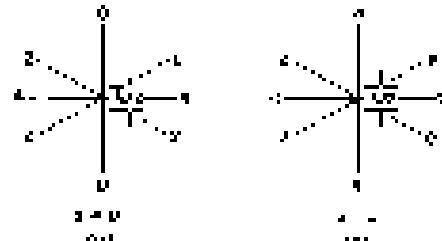


Fig. 20-5. (Left) Intensity in surface directions from two dipole oscillators, one-half wavelength apart. Left: a dipole $A = A_0$. Right: two off-period out-of-phase $B = B_0$.

conductivity change, but the "mean" or energy that is going past, in watts per square meter, is proportional to the square of the field. We shall derive the constant of proportionality in the next chapter.) If we look at the array from the N-S side, both oscillators are phase-shifted $\pi/2$ radians, so the electric field is zero as strong as it would be from a single one-tube. Therefore the intensity is zero there, or twice as it would be if there were only one oscillator. (The numbers in Fig. 29-5 represent how strong the intensity would be in this case, compared with what it would be if there were only one dipole oscillating of unit strength.) Now, if we turn the N-S direction along the line of the oscillators, since they are half a wavelength apart, the effect of one oscillator turns out to be just of phase by exactly half an oscillation from the other, and the total has again odd or zero. At a certain particular intermediate angle (in fact, at 45°) the intensity is 2, and it goes up to 4, 5, 6, and so on. We have to learn how to find it now, when the other angles, 1, 3, & 4 represent of setting two oscillators with different phases.

Let us quickly look at some other cases of interest. Suppose the oscillators are again made $\lambda/2$ wavelength apart, but the phase of one is set back behind the center of its oscillation (Fig. 29-6). In the W direction the intensity is now zero, because one oscillator is "pushing" when the other one is "pulling." But in the S direction the signal from the near antenna lags a certain time, and that of the other comes half a period later. But the latter has already had $\lambda/2$ a prior setting, so it is strong, and therefore it is now exactly in step with the first one, and the intensity in this direction is 4 units. The intensity in the direction at 90° is still 2, as we can prove later.

But we come to an interesting case which shows up a possibly useful feature. Let us remark that one of the reasons the phase relations of oscillators are interesting is for learning radio transmitters. For instance, if we build an antenna system and want to send a radio signal, say to Hawaii, where the antennas are as in Fig. 29-5(a) and we broadcast with our two antennas in phase, Hawaii Hawaii is to the west of us. Then we decide that when you are so far away to broadcast toward Alaska's Canada. Since Canada is to the west, all we have to do is to reverse the phase of one of our antennas, and we can broadcast to the north so we can build an entire system with minimum interference. This is one of the simplest situations; we can send the beams in all directions and send most of the power in the direction in which we wish to transmit, without ever moving the antenna. In lots of the practical cases, however, while we are broadcasting toward Alaska we are moving a lot of power to Texas, Texas, and it would be interesting to ask whether it is possible to send it in only one direction. At least right we might think that, with a pair of antennas of this nature the result is always going to be symmetric. So let us consider a case that is more or less symmetrical, to show the possible variety.

If the antennas are separated by one-quarter wavelength, and if the N one is reinforced behind the $\lambda/2$ one at time, then what happens (Fig. 29-6)? In the W direction was $\pi/2$, so we will see "zero." In the S direction we get zero here for the signal from S because it is "out of time"; but the $\lambda/4$ comes 90° later in, and, being $\lambda/2$, adds 90° behind it. Its beat in phase, therefore it arrives, although $\lambda/2$ out of phase, and there is no effect. On the other hand, in the E direction, the N signal arrives earlier than the S signal by $\lambda/2$ total, so because it is a quarter-wavelength close. But a $\lambda/2$ is set so that it is really in 90° behind in time, that is, one-quarter-wave delay difference, and therefore the two signals begin negative in phase, adding the field strengths twice as large, and the energy four times as great.

Thus, by using some cleverness in spacing and phasing one antenna, we can send the power all in one direction. This is a directional over a great range of angles. But we arrange it so that it is broadest in more sharply in a particular direction. Let us consider the case of the two antennas which we are sending to Texas, but it is spread over quite a wide beam. In other words, if we are still getting half the intensity, we are wasting the power. Can we do better? Can this? Let us take a situation in which the separation is $\lambda/4$ - $\lambda/2$ in Fig. 29-6.

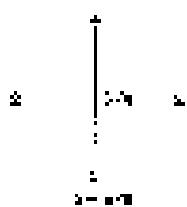


Fig. 29-6. A paired dipole antenna during maximum power in one direction.

29.7), which is more or less comparable to the situation in which we expect that in the $\theta = 90^\circ$ direction, with separation of several wavelengths, there is "almost no effect of a wavelength". Here the picture is quite different.

If the two oscillators are wave lengths apart, we take the important case to have $\theta = 90^\circ$, we see that in the $E \parallel \theta$ direction, they are in phase, and we get a strong intensity, four times what we would get if one of them were there alone. On the other hand, at a very small angle away, the intensities are different by 10% and the intensity is zero. To be accurate, if we draw a line from one oscillator to a distant point and the difference of the two distances is $\lambda/2$, the far oscillation, but they are out of phase, so the distance does not matter. (The figure is not drawn to scale; it is only a rough sketch.) This means that we do indeed have a very good reason in the question we asked, because if we put more than a fifth of a wavelength off the center of the $E \parallel \theta$ direction, because if we put more than a fifth of a wavelength off the center of the $E \parallel \theta$ direction, the intensity would drop to zero. This is why it is better to have the two oscillators as far apart as possible. We will see in the next chapter that this is what happens in the $E \parallel \theta$ direction, and the intensity is zero. This is what we learned in Chapter 28.

Now how can we arrange to get rid of these extra maxima in "phase loss" as you called? We could get rid of the unwanted $E \parallel \theta$ direction in a fairly clever way. Suppose we want to produce a certain ratio of strengths between the two that we already have. That is, the outside ones are at 100, and that between them, say, 10, so we have produced an antenna, and we have them all in phase. These are now six antennas, and if we add up all the intensity of the $E \parallel \theta$ direction, it would not be much higher with six antennas than with one. The field would be six times and the intensity thirty-six times stronger (the square of the field). We get a unity of intensity in that direction. Now if we look at neighboring pairs, we find a wave length, say λ , but if we put it there, we have to move to get a "in-phase pair" we get a much smaller "bump" now—let us try to see why.

The reason is that although we might expect to get a "bump" when the distance is a wave length, it is not then dipole 1 and dipole 2 in phase, and they are in phase, and so operating in trying to get extra strength in that direction. But numbers 3 and 4 are roughly $1/\sqrt{2}$ wavelength apart with 1 and 2, and 5 and 6 roughly $1/\sqrt{2}$ pushed together, and 6 just has the tail, but is opposite phases. Therefore there is very little in every in this direction—but there is something. It does not balance exactly. The kind of thing keeps on increasing, we get very nice humps, and we have a broad beam in the direction where we expect it. Then this pattern is not the same as something else we chosen; namely, since the distance between successive dipoles is 2λ , it is possible to find a range where the distance is between two successive dipoles, say, one wave length, so that the effects from all of them are in phase again. Each one is delayed relative to the next one by $2\pi/2\lambda$, so that all come back in phase, and we have another strong beam in that direction. It is very evident, this pattern is not—it is possible to put the dipoles closer than one wave length apart. If we put it more antennae, closer than one wavelength apart than the same happens. So the lesson is this can happen at certain angles, if the spacing is big enough and we change it, so very interesting are these phenomena in other applications—not radio broadcasting, but in diffraction gratings.

29.5 The mathematics of interference

Now we have finished our analysis of the problem of dipole sources out-of-phase, and we must learn how to analyze them quantitatively. To find the effect of a source's being at a certain angle in the most general case, where the two oscillators have some arbitrary relative phase ϕ in one module and the strengths A_1 and A_2 are not equal, we find that we have two beams having the same frequency, but with different phases. It is very easy to find this phase difference, it is once again $2\pi/2\lambda$ due to the difference in distance, and the difference in the phase of the two oscillators. Mathematically, we have to find the sum of two waves: $A_1 + A_2 \cos(\omega t + \phi_1) + A_2 \cos(\omega t + \phi_2)$. Now let us do it:

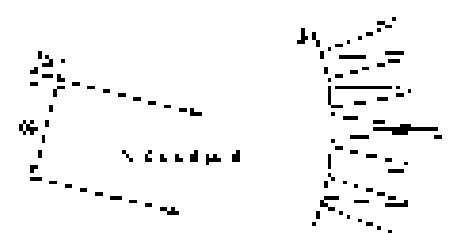


Fig. 29-7. The intensity pattern for two dipoles separated by 10λ .

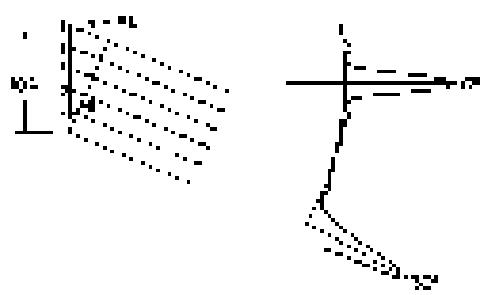


Fig. 29-8. A single dipole antenna in out-of-phase and in-phase intensity patterns.

It is really very easy, and we assume that we already know how to do it. However, we shall outline the procedure in some detail. First, we say if we are given two oscillators and know enough about ceses and laws, it would work it out. If we know two vectors \mathbf{A}_1 and \mathbf{A}_2 , amplitude let us say may be A_1 and A_2 , the phase ϕ_1 and ϕ_2 , let us say may be $\omega t + \phi_1$ and $\omega t + \phi_2$. In such circumstances, for example we could use this the geometrical method of solving the problem, we have

$$R = A_1 \cos(\omega t + \phi_1) + A_2 \cos(\omega t + \phi_2). \quad (29.1)$$

Once, in our trigonometry class, we may have learned the rule that

$$\cos A - \cos B = 2 \cos((A + B)/2) \cos((A - B)/2). \quad (29.2)$$

If we know that, then we can immediately write R as

$$R = A_1 \cos(\omega t + \phi_1) + A_2 \cos(\omega t + \phi_2) = R_0 \cos(\omega t + \phi_0). \quad (29.3)$$

Now first note we have an oscillatory wave with a new phase and a new amplitude. In general, the result will be two oscillators wave with a new amplitude A_0 , which we may call the resultant amplitude, oscillating at the same frequency but with a phase difference ϕ_0 , called the resultant phase. In view of this, our particular case has the following results that the resultant amplitude is

$$A_0 = 2A_1 \cos(\phi_1 - \phi_2), \quad (29.4)$$

and the resultant phase is the average of the two phases, and we have completely solved our problem.

Now suppose that we cannot remember that the sum of 180 degrees is twice the sum of half the sum times the cosine of all the differences. Then we may use another method of analysis which is more geometrical. Any cosine function or sine can be visualized as the harmonic projection of a rotating vector. Suppose there was a vector \mathbf{A}_1 of length A_1 rotating with time, at that angle with the horizontal axis $\omega t = \phi_1$. We shall leave out the unit of time minute, and see that it makes no difference. Suppose that we take a snapshot at the time $t = 0$, although, in fact, the picture is moving at angular velocity ω (Fig. 29-9). The projection of \mathbf{A}_1 along the horizontal axis is precisely $A_1 \cos(\omega t + \phi_1)$. Now at $t = 0$, the second wave would be represented by another vector, \mathbf{A}_2 , of length A_2 and at an angle ϕ_2 , also rotating. They are both rotating with the same angular velocity ω , and therefore the relative positions of the two are fixed. The two just rotate like a rigid body. The horizontal projection of \mathbf{A}_2 is $A_2 \cos(\omega t + \phi_2)$. But we know from the theory of waves that if we add the two vectors in the ordinary way, by the parallelogram rule, and draw the resultant vector \mathbf{A}_0 , the resultant phase of the resultant is the sum of the separate phases of the other two waves. This solves our problem. It is easy to check that the gives the correct result for the special case we treated above, where $A_1 = A_2 = A$. In this case, we see from Fig. 29-9 that the angle between \mathbf{A}_1 and \mathbf{A}_2 and makes an angle $\phi_1 - \phi_2$ with each. The vector we see that $A_0 = 2A_1 \cos(\phi_1 - \phi_2)$, as before. Also, as we see from the a angle, the phase of \mathbf{A}_0 , as it goes around, is the average angle of \mathbf{A}_1 and \mathbf{A}_2 when the two amplitudes are equal. Clearly, we can do the same for the case where the amplitudes are different, just as easily. We can still that the geometrical way of solving the problem.

There is another way of solving the problem, and that is the analytical way. That is, instead of having actually to draw a picture like Fig. 29-8, we can make something done which is to draw this vector in pictures instead of drawing the vectors, we write a complex number to represent each of the vectors. The real parts of the complex numbers are the actual physical quantities. So if $\mathbf{A}_1 = A_1 \cos(\omega t + \phi_1)$ we can write it in the form $A_1 e^{i(\omega t + \phi_1)}$ (just as well as $A_1 \cos(\omega t + \phi_1)$) and $A_2 e^{i(\omega t + \phi_2)}$. Now we can add the two:

$$R = A_1 e^{i(\omega t + \phi_1)} + A_2 e^{i(\omega t + \phi_2)} = (A_1 \cos \theta_1 + A_2 \cos \theta_2) + i(A_1 \sin \theta_1 + A_2 \sin \theta_2). \quad (29.5)$$

or

$$R = A_0 e^{i(\omega t + \phi_0)} = A_0 \cos \theta_0 + iA_0 \sin \theta_0. \quad (29.6)$$

The solves the problem that we wanted to solve, because it represents a resultant as a complex number of magnitude A_2 and phase φ_2 .

To see how this worked out, let me find the amplitude A_2 which is the "length" of \vec{A} . To get the "length" of a complex quantity, we always multiply its quantity by its conjugate (using \ast , which gives the length squared). The complex conjugate is the same expression, but with the sign of one i 's reversed. Thus we have

$$A_2^2 = (A_1 e^{i\varphi_1} + A_2 e^{i\varphi_2}) (A_1 e^{-i\varphi_1} + A_2 e^{-i\varphi_2}). \quad (29.15)$$

In multiplying this out, we get $A_1^2 + A_2^2$ (here the i 's cancel), and for the cross terms we have

$$A_1 A_2 (e^{i\varphi_1 - i\varphi_2} + e^{i\varphi_2 - i\varphi_1}).$$

Now

$$e^{i\varphi} + e^{-i\varphi} = \cos \varphi + i \sin \varphi - \cos \varphi - i \sin \varphi = 0.$$

That is to say, $e^{i\varphi} + e^{-i\varphi} = 2 \cos \varphi$. Our final result is therefore

$$A_2^2 = A_1^2 + A_2^2 + 2 A_1 A_2 \cos(\varphi_2 - \varphi_1). \quad (29.16)$$

As we saw, this agrees with the length of \vec{A}_2 in Eq. 29-9, using the rules of trigonometry.

This the sum of the two effects has the intensity A_2^2 we would get with one of them alone, plus the intensity A_1^2 we would get with the other one alone, plus a correction. The correction we call the interference effect. It is really only the difference we were given directly by adding the intensities, and who's usually interested. We can't interference whether it is positive or negative. Interference in ordinary language usually suggests opposition or hindrance, but in physics we often do not consider going the way it was originally designed! If the interference term is positive, we call that case constructive interference, because enough intensity seems to anybody other than a physicist! The opposite case is called destructive interference.

Now let us see how to apply our general formula (29.16) for the case of two oscillators to the specific situations which we have discussed qualitatively. To apply the general formula, it is only necessary to find the phase difference $\varphi_2 - \varphi_1$ exists between the signals arriving at a given point. (It depends only on the source difference, of course, and not on the phase itself.) So let us consider the case where the two oscillators, of equal amplitude, are separated by some distance d and have an increasing velocity phase ωt . (When ωt is at phase zero, the phase of the other is $\pi/2$). Then we ask what the phase will be in some time with distance d from the E-W line. [Note: the ωt is not the same θ as appears in (29.1). We are doing *vector* using the rectangular symbol, here θ ; or the semi-circular symbol ϑ (Fig. 29-10);? The phase relation is $\theta = \omega t$ and we know that the difference in distance from it to the two oscillators is d on θ so that the phase difference is $\omega d / \lambda$. The wave velocity is c on θ so that the phase difference is $\omega d / c$. The number of wavelengths is d/λ , so that the phase difference is $2\pi(d/\lambda)$. Those who are more sophisticated might want to multiply the wave number k , which is the rate of change of phase with distance, by d/λ since k is nearly the same.] The phase difference due to the distance difference is thus $2\pi(d/\lambda)\sin(\omega t)$, but, due to the timing of the oscillators, there is an additional phase α . So the phase difference at any instant will be

$$\varphi_2 = \varphi_1 - \alpha + 2\pi(d/\lambda)\sin(\omega t). \quad (29.17)$$

Let's look at some cases. Thus if, we have to do, we substitute this expression into (29.16) for the case $A_1 = A_2$, and we can calculate all the various cases of two sources of equal intensity.

Now let us see what happens in our various cases. For lesson #6, for example, the intensity is 2 at 20° in Fig. 29-5 is the following. The two vectors are at $\omega t = 10^\circ$, down $\sqrt{3}/2$, down $1/2$. Their sum $= A_1 = \sqrt{3}/2$ at 20° , and so the interference term is zero. (We are adding two vectors at 90° .) The result is the hypotenuse of a 45° right-angle triangle, where $1/\sqrt{2}$ is the unit amplitude A_1 going in, we get with the intensity of one just along there. All the other cases can be worked out in this same way.

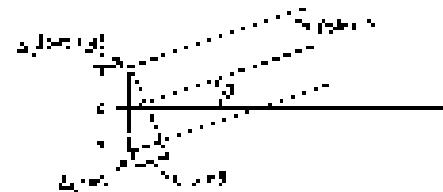


Fig. 29-10. Two oscillators of equal amplitude, with a phase difference α between them.

Diffraction

30-1 The resultant amplitude due to *n* equal oscillators

This chapter is a direct continuation of the previous one, although the name has been changed from *Interference* to *Diffraction*. No one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage, and there is no specific, important physical difference between them. The fact we can do, roughly speaking, is to say that when there are only a few sources, say two, interfering, then the result is usually called interference, but if there is a large number of them, it seems that the word diffraction is more often used. So, we shall not worry about whether it is interference or diffraction, but continue directly from where we left off in the middle of the subject in the last chapter.

Thus we shall now discuss the situation where there are *n* equally spaced oscillators, all of equal amplitude but different from one another in phase, either because they are driven differently in phase, or because we are looking at them at an angle such that there is a difference in time delay. For one reason or another, we have to add something like this:

$$R = A[\cos(\omega t + \phi_0) + \cos(\omega t - \phi_1) + \cos(\omega t - 2\phi_2) + \cdots + \cos(\omega t + (\pi - 1)\phi_n)], \quad (30.1)$$

where ϕ is the phase difference between our oscillator and the last one as seen in a $\phi/2$ later direction. Specifically, $\phi = \alpha - 2\pi d \sin \theta/\lambda$. Now we must add all the terms together. We shall do this geometrically. The first one is of length A , and it has zero phase. The next is also of length A and it has a phase equal to ϕ . The last one is again of length A and it has a phase equal to 2ϕ , and so on. So we are evidently going around an equiangular polygon with n sides (Fig. 30-1).

Now the vertices, of course, all lie on a circle, and we can find the net amplitude most easily if we find the radius of that circle. Suppose that Q is the center of the circle. Then we know that the angle OQS is just a phase angle ϕ . (This is because the radius QS bears the same geometrical relation to A_2 as QO bears to A , so they form an angle ϕ between them.) Therefore the radius r must be such that $A = 2r \sin \phi/2$, which fixes r . But the large angle OQT is equal to $\pi\phi$, and we find that $A_N = 2r \sin \pi\phi/2$. Combining these two results to eliminate r , we get

$$A_N = A \frac{\sin \pi\phi/2}{\sin \phi/2}. \quad (30.2)$$

The resultant intensity is thus

$$I = I_0 \frac{\sin^2 \pi\phi/2}{\sin^2 \phi/2}. \quad (30.3)$$

Now let us examine this expression and study some of its consequences. In the first place, we can check it for $n = 1$. It checks: $I = I_0$. Next, we check it for $n = 2$: writing $\sin \phi = 2 \sin \pi/2 \cos \phi/2$, we find that $A_N = 2A \cos \phi/2$, which agrees with (20.12).

Now the idea that led us to consider the addition of several sources was that we might get a much stronger intensity in one direction than in another; that the nearby maxima which would have been present if there were only two sources will have gone down in strength. In order to see this effect, we plot the curve that results from (30.3), taking π to be extremely large and plotting the cosine term

30-1 The resultant amplitude due to *n* equal oscillators

30-2 The diffraction grating

30-3 Resolving power of a grating

30-4 The parabolic antenna

30-5 Colored films; crystals

30-6 Diffraction by opaque screens

30-7 The field of a plane of oscillating charges

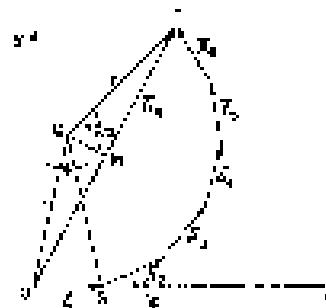


Fig. 30-1. The resultant amplitude of *n* equally spaced sources with net successive phase differences ϕ .

$\phi = 0$. In the first place, if ϕ is exactly 0, we have 0/0, but if ϕ is infinitesimal, the ratio of the two sines squared is simply n^2 , since the sine and the angle are approximately equal. Thus the intensity of the maximum of the curve is equal to n^2 times the intensity of one oscillator. That is easy to see, because if they are all in phase, then the little vectors have no relative angle and all of them adding so the amplitude is n times, and the intensity n^2 times stronger.

As the phase ϕ increases, the size of the two sines begins to fall off, and the first time it reaches zero is when $n\phi/2 = \pi$, because $\sin \pi = 0$. In other words, $\phi = 2\pi/n$ corresponds to the first minimum in the curve (Fig. 20-2). In view of what is happening with the arrows in Fig. 20-1, the first minimum occurs when all the arrows come back to the starting point; that means that the total sum of vector angle in all the arrows, the total phase difference between the first and last oscillator, must be 2π to go up the loop.

Now we go to the next maximum, and we want to see that it's really much smaller than the first one, as we had hoped. We shall not go precisely to the maximum position, because both the numerator and the denominator of (20.3) are zero, and $\sin n\phi/2$ varies quite slowly compared with $\sin n\phi/2$ when ϕ is large, so $\sin n\phi/2 = 1$ we are very close to the maximum. The next maximum of $\sin^2 n\phi/2$ comes at $n\phi/2 = 3\pi/2$, or $\phi = 3\pi/2n$. This corresponds to the arrows having traversed the circle one and $1/2$ full turns. On putting $\phi = 3\pi/2n$ into the formula to find the size of the maximum, we find that $n^2/5\pi/2 = 1$ in the numerator (which is why we picked this angle), and in the denominator we have $\sin^2 3\pi/2n$. Now if n is sufficiently large, then this angle is very small and the sine is equal to the angle; so for all practical purposes, we can put $\sin 3\pi/2n = 3\pi/2n$. Thus we find that the intensity at this maximum is $I = I_0(4\pi^2/9\pi^2)$. But $4\pi^2/9\pi^2$ was the maximum intensity, and so we have $4/9\pi^2$ times the maximum intensity, which is about 0.017, less than 5 percent of the maximum intensity. Of course, but we're decreasing intensities farther out. So we have a very sharp central maximum with very weak subsidiary maxima on the sides.

It is possible to prove that the area of the whole curve, including all the little bumps, is equal to $2\pi n I_0$, or twice the area of the dotted rectangle in Fig. 20-2.

Now let us consider further how we may apply Eq. (20.3) in different circumstances, and try to understand what's happening. Let us consider our sources to be all on a line, as drawn in Fig. 20-3. These are n of them, all spaced by a distance a , and we shall suppose that the intrinsic relative phase, due to the next, is θ . Then if we are observing in a given direction θ from the normal, there is an additional phase $2\pi d \sin \theta/a$ because of the time delay between each successive row, which we talked about before. Thus

$$\begin{aligned}\phi &= \phi + 2\pi d \sin \theta/a \\ &= \phi + 2\pi d \sin \theta\end{aligned}\quad (20.4)$$

First, we shall take the case $\theta = 0$. That is, all oscillators are in phase, and we want to know what the intensity is as a function of the angle θ . In order to find out, we merely have to put $\phi = 2\pi d \sin \theta$ into formula (20.3) and see what happens. In the first place, there is a maximum at $\theta = \phi = 0$. That means that when all the oscillators are in phase there is a strong intensity in the direction $\theta = 0$. On the other hand, an interesting question is, where is the first minimum? That occurs when $\theta = 2\pi/n$. In other words, when $2\pi d \sin \theta/a = 2\pi/n$, we get the first minimum of the curve. If we get rid of the 2π 's so we can look at it a little better, it says that

$$nd \sin \theta = \lambda. \quad (20.5)$$

Now let us understand physically why we get a minimum at that position. nd is the total length l of the array. Referring to Fig. 20-3, we see that $nd \sin \theta = l \sin \theta = l - \theta$. What (20.5) says is that when θ is equal to our wavelength, we get a minimum. Now why do we get a minimum when $\theta = \lambda$? Because the contributions of the various oscillators are then uniformly distributed in phase from 20-2

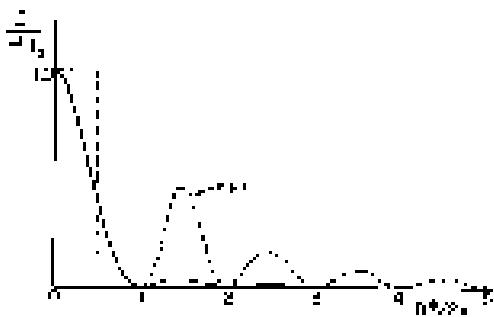


Fig. 20-2. The intensity as a function of phase angle for a large number of oscillators of equal strength.



Fig. 20-3. A linear array of n equal oscillators, driven with phases $\omega_t = \omega$.

AMF. The arrows (Fig. 30-1) are going around a whole circle—we are adding equal vectors in all directions, and each π sum is zero. So when we have an angle such that $A = \lambda$, we get a minimum. That is the first minimum.

There is another important feature about formula (30.3), which is that if the angle ϕ is increased by any multiple of 2π , it makes no difference to the formula. So we will get other strong minima at $\phi = 3\pi, 4\pi, 6\pi$, and so forth. Near such of these great maxima the pattern of Fig. 30-2 is repeated. We may ask ourselves, what is the geometrical circumstance that leads to these other great maxima? The condition is that $\phi = 2m\pi$, where m is any integer. That is, $2\pi d \sin \theta/2 = 2m\pi$. Dividing by 2π , we see that

$$d \sin \theta = m\lambda. \quad (30.6)$$

This looks like the other formula, (30.5). No, that formula was $d \sin \theta = \lambda$. The difference is that here we have to look at the individual sources, just when we say $d \sin \theta = m\lambda$, that means that we have an angle θ such that $\delta = m\lambda$. In other words, each source is now contributing a certain number, and successive rays are out of phase by a whole multiple of 360° , and therefore are contributing in phase, because out of phase by 360° is the same as being in phase. So they all contribute in phase and produce just as good a maximum as the one for $m = 0$ that we discussed before. The subsidiary beams, the whole shape of the pattern, is just like the one near $\phi = 0$, with exactly the same radiation envelope, etc. Thus such an array will send beams in various directions—each beam having a strong central maximum and a certain number of weak “side lobes.” The various strong beams are referred to as the zero-order beam, the first-order beam, etc., according to the value of m . m is called the order of the beam.

We call attention to the fact that if d is less than λ , Eq. (30.6) can have no solution except $m = 0$, so that if the spacing is too small there is only one possible beam, the zero-order one centered at $\theta = 0$. (Of course, there is also a beam in the opposite direction.) In order to get subsidiary exact maxima, we must have the spacing d of the array greater than one wavelength.

30-2. The diffraction grating

In technical work with antennas and wires it is possible to arrange that all the phases of the little oscillators, or antennas, are equal. The question is whether and how we can do a similar thing with light. We cannot at the present time literally make little optical-frequency radio stations and hook them up with infinitesimal wires and drive them all with a given phase. But there is a very easy way to do what amounts to the same thing.

Suppose that we had a lot of parallel wires, equally spaced at a spacing d , and a radio-frequency source very far away, practically at infinity, which is generating an electric field which arrives at each one of the wires at the same phase (it is so far away that the time delay is the same for all of the wires). (One can work out cases with curved arrays, but let us take a plane one.) Then the external electric field will drive the electrons up and down in each wire. That is, the field which is coming from the original source will shake the electrons up and down, and in moving, these represent new generators. This phenomenon is called scattering; a light wave from some source can induce a motion of the electrons in a piece of material, and these motions generate their own waves. Therefore all that is necessary is to set up a lot of wires, equally spaced, drive them with a radio-frequency source far away, and we have the situation (let us wish, without a whole lot of special wiring). If the incidence is normal, the phases will be equal, and we will get exactly the circumstances we have been discussing. Therefore, if the wire spacing is greater than the wavelength, we will get a strong intensity of scattering in the normal direction, and in certain other directions given by (30.6).

This can also be done with light! Instead of wires, we use a flat piece of glass and make notches in it such that each of the notches scatters a little differently than the rest of the glass. If we then shine light on the glass, each one of the notches

will represent a source, and if we space the lines very finely, but not closer than a wavelength (which is technically almost impossible anyway), then we would expect a rather curious phenomenon: the light will only pass straight through, but there will also be a strong beam at a finite angle, depending on the spacing of the notches! Such objects have actually been made and are in common use—they are called *diffractive gratings*.

In one of its forms, a diffractive grating consists of notching but a plane glass sheet, transparent and colorless, with scratches on it. There are often several hundred scratches to the millimeter, very carefully arranged so as to be equally spaced. The effect of such a grating can be seen by arranging a projector so as to throw a narrow, vertical line of light (the image of a slit) onto a screen. When we put the grating close to the beam, with its scratches vertical, we see that the line is still there but, in addition, on each side we have another strong patch of light which is colored. This, of course, is the slit image spread out over a wide angular range, because the angle θ_{m} (30.6°) depends upon λ , and lights of different colors, as we know, correspond to different frequencies, and therefore different wavelengths. The longest visible wavelength is red, and since $d \sin \theta = \lambda$, that requires a larger θ . And we do, in fact, find that red is at a greater angle from the central image. There should also be a beam on the other side, and indeed we see one on the screen. There, though, be another solution of (30.6) when $m = -1$. We do see that there is something vaguely there—very weak—and there are even other beams beyond.

We have just argued that all these beams ought to be of the same strength, but we see that they actually are not and, in fact, not even the first ones on the right and left are equal! The reason is that the grating has been carefully built up to do just this. How? If the grating consists of very fine notches, uniformly wide, spaced evenly, then all the intensities would indeed be equal. But, as a matter of fact, although we have taken the simplest case, we could also have considered an array of pairs of antennas, in which each member of the pair has a certain strength and some relative phase. In this case, it is possible to get intensities which are different in the different orders. A grating is often made with little "sawtooth" cuts instead of little symmetrical notches. By carefully arranging the "sawtooths," more light may be sent into one particular order of spectrum than into the others. In a practical grating, we would like to have as much light as possible in one of the orders. This may seem a complicated point to bring in, but it is a very clever thing to do, because it makes the grating more useful.

So far, we have taken the case where all the phases of the sources are equal. But we may have a formula for ϕ when the phases differ from one to the next by an angle α . That requires adding up our intensities with a slight phase shift between each one. Can we do that with light? Yes, we can do it very easily, for suppose that there were a source of light at infinity, at an angle such that the light is coming in at an angle θ_0 , and let us say that we wish to discuss the scattered beam, which is scattered at an angle θ_{sc} . The θ_{sc} is the same as we have had before, but the θ_0 is merely a means for arranging that the phase of each source is different: the light coming from the distant driving sources first hits one scratch, then the next, then the next, and so on, with a phase shift from one to the other, which, as we see, is $\alpha = d \sin \theta_0 / \lambda$. Therefore we have the formula for a grating in which light both comes in and goes out at an angle:

$$\phi = 2\pi d \sin \theta_{\text{out}} / \lambda - 2\pi d \sin \theta_0 / \lambda. \quad (30.7)$$

Let us try to find out where we get strong intensity in these circumstances. The condition for strong intensities is, of course, that ϕ should be a multiple of 2π . There are several interesting points to be noted.

One case of rather great interest is that which corresponds to $m = 0$, where d is less than λ ; in fact, this is the only solution. In this case we see that $\sin \theta_{\text{sc}} = \sin \theta_0$, which means that the light comes out in the same direction as the light which was entering the grating. We might think that the light "goes right through." No, it is *different* light that we are talking about. The light that goes right through is from the original source; what we are talking about is the new light which is

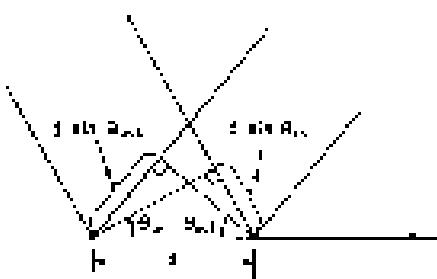


Fig. 30-4. The path difference for rays scattered from adjacent rulings of a grating is $d \sin \theta_{\text{sc}} - d \sin \theta_0$.

generated by scattering. It turns out that the scattered light is going in the same direction as the original light, in fact it can interfere with it—a feature which we will study later.

There is another solution for this same case. For a given $k_{\text{sc}} R_{\text{out}}$ (say) as the supplement of θ_{sc} . So not only do we get a beam in the same direction as the incoming beam but also one in a vector direction, which, if we consider it carefully, is such that the angle of incidence is equal to the angle of separation. This we call the reflected beam.

So we begin to understand the basic geometry of reflection: the light that comes in generates motions of the atoms in the reflector, and the reflector then regenerates a new wave, and one of the solutions for the description of scattering, the anti-scatterism if the spacing of the scatterers is small compared with one wavelength, is that the angle at which the light comes out is equal to the angle at which it comes in.

Next, we discuss the special case when $d \rightarrow 0$. That is, we have just a solid piece of material, so to speak, but of finite length. In addition, we want the phase shift from one scatterer to the next to go to zero. In other words, we put more and more oscillators between the other ends, so that each of the phase differences is varying smaller, but the number of scatterers is increasing in such a way that the total phase difference, between one end of the bar and the other, is constant. Let us see what happens to (10.3) if we keep the difference in phase $\pi\phi$ from one end to the other constant (say $\phi = 45^\circ$), letting the number go to infinity and the phase shift α of each one go to zero. But now α is so small that $\sin \phi = \phi$, and if we also recognize $\pi^2 J_0$ as I_0 , the maximum intensity at the center of the beam, we find

$$I = 4I_0 \sin^2 \frac{\pi d}{\lambda} / \pi^2. \quad (10.8)$$

This limiting case is what is shown in Fig. 30-2.

In such circumstances we find the same general kind of a picture as for finite spacing with $d > \lambda$; all the side lobes are practically the same as before, but there are no higher-order maxima. If the scatterers are all in phase, we get a maximum in the direction $\theta_{\text{out}} = 0$, and a minimum when the distance λ is equal to π , just as for finite d and n . So we can even analyze a continuous distribution of scatterers or oscillators, by using integrals instead of summing.

As an example, suppose there were a long line of oscillators, with the charge oscillating along the direction of the line (Fig. 30-3). From such an array the greatest intensity is perpendicular to the line. There is a little bit of intensity up and down from the equatorial plane, but it is very slight. With this result, we can handle a more complicated situation. Suppose we have a set of such lines, each producing a beam only in a plane perpendicular to the line. To find the intensity in various directions form a series of long wires instead of infinitesimal wires, is the same problem as it was for infinitesimal wires, so long as we are in the central plane perpendicular to the wires; we just add the contribution from each of the long wires. That is why, although we actually analyzed only tiny antennas, we might as well have used a grating with long, narrow slots. Each of the long slots produces an effect only in its own direction, not up and down, but they are all aligned to give a other horizontally, so they produce interference that way.

Thus we can build up more complicated situations by having various distributions of scatterers in lines, planes, or in space. The first thing we did was to consider scatterers in a line, and we have just extended the analysis to strips: we can work it out by just doing the necessary summations, adding the contributions from the individual scatterers. The principle is always the same.

30-3 Resolving power of a grating

We are now in a position to understand a number of interesting phenomena. For example, consider the use of a grating for separating wavelengths. We noticed that the whole spectrum was spread out on the screen, so a grating can be used as an instrument for separating light into its different wavelengths. One of the

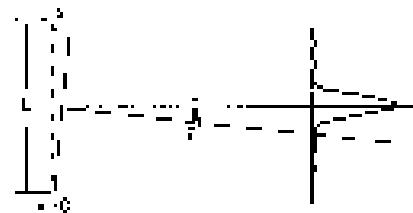


Fig. 30-5. The intensity pattern of a continuous line of oscillators has a single strong maximum and many weak "side lobes."

interesting question is: supposing that there were two sources of slightly different frequency, or slightly different wavelength, how close together in wavelength could they be such that the grating would be unable to tell that there were really two different wavelengths? The red and the blue were clearly separated. But when one wave is red and the other is slightly redder, very close, how close can they be? This is called the *resolving power** of the grating, and one way of analyzing the problem is as follows. Suppose that for light of a certain color we happen to have the maximum of the diffracted beam occurring at a certain angle. If we vary the wavelength the phase $2\pi d \sin \theta / \lambda$ is different, so of course the maximum occurs at a different angle. That is why the red and blue are spread out. How different in angle must it be in order for us to be able to see it? If the two maxima are exactly on top of each other, of course we cannot see them. If the maximum of one is far enough away from the other, then we can see that there is a double hump in the distribution of light. In order to be able to just make out the double hump, the following simple criterion, called Rayleigh's criterion, is usually used (Fig. 30-6). It is that the first minimum from one bump should sit at the maximum of the other. Now it is very easy to calculate, when one minimum sits on the other maximum, how much the difference in wavelength is. The best way to do it is geometrically.

In order to have a maximum for wavelength λ' , the distance d (Fig. 30-3) must be λ'/m , and if we're looking at the n th-order beam, it is $m\lambda'/d$. In other words, $2\pi d \sin \theta / \lambda' = 2\pi m, \text{or } \sin \theta, \text{ which is } \Delta, \text{ is } N \text{ times } n, \text{ or } m\lambda/d$. For the other beam, of wavelength λ , we want to have a minimum at this angle. That is, we want Δ to be exactly one wavelength λ more than even. That is, $\Delta = m\lambda - \lambda = m\lambda/d$. Thus if $\lambda' = \lambda + 1/\Delta$, we find

$$m\lambda/d = 1/\Delta; \quad (30.9)$$

The ratio λ/Δ is called the *resolving power* of a grating; we see that it is equal to the total number of lines in the grating, times the order. It is not hard to prove that this formula is equivalent to the formula that the error in frequency is equal to the reciprocal times the difference between extreme path differences (lower to take account



Fig. 30-6. Illustration of the Rayleigh criterion. The maximum of one pattern falls on the first minimum of the other.

In fact, that is the best way to remember it, because the general formula works not only for gratings, but for any other instrument whatsoever, while the special formula (30.9) depends on the fact that we are using a grating.

30-4 The parabolic antenna

Now let us consider another problem in resolving power. This has to do with the antenna of a radio telescope, used for determining the position of radio sources in the sky, i.e., how large they are in angle. Of course if we use any old antenna and find signals, we would not know from what direction they came. We are very interested to know whether the source is in one place or another. One way we can find out is to lay out a whole series of equally spaced dipole wires on the Australian landscape. Then we take all the wires from these antennas and feed them into the same receiver, in such a way that all the delays in the feed lines are equal. Thus the receiver receives signals from all of the dipoles in phase. Then it adds all the waves from every one of the dipoles in the same phase. Now what happens? If the source is directly above the array, at infinity or nearly so, then its radio waves will excite all the antennas in the same phase, so they all feed the receiver together.

Now suppose that the radio source is at a slight angle θ from the vertical. Then the various antennae are receiving signals a little out of phase. The receiver adds all these out-of-phase signals together, and so we get nothing, at the angle

* In our case $T = \lambda/c = \omega/c^2$, where c is the speed of light. The frequency $\nu = c/T$, so $\Delta\nu = c\Delta\lambda/\lambda^2$.

θ is too big. How big may the angle be? Answer: we get zero if the angle $\theta/L \approx 0$ (Fig. 20-4) corresponds to a 180° phase shift, that is, if λ is the wavelength λ . This is because the vector contributions form together a complete polygon with zero resultant. The smallest angle that can be resolved by an antenna array of length L is $\theta = \lambda/L$. Notice that the receiving pattern of an antenna such as this is exactly the same as the intensity distribution we would get if we turned the receiver around and made it into a transmitter. This is an example of what is called a *reciprocity principle*. It turns out, in fact, to be generally true for any arrangement of antennas, a radio, and so on, that if we first work out what the relative intensities would be in various directions if the receiver were a transmitter instead, then the relative directional sensitivity of a receiver with the same collection of dipoles, the same array of antennas, is the same as the relative intensity of emission would be if it were a transmitter.

Some radio antennae are made in a different way. Instead of having a whole lot of dipoles in a long line, with a lot of feed wires, we may arrange them not in a line but in a curve, and put the receiver at a certain point where it can detect the scattered waves. This curve is cleverly designed so that if the radiowaves are coming down from above, and the wires scatter, making a new wave, the wires are so arranged that the scattered waves reach the receiver all at the same time (Fig. 26-12). In other words, the curve is a parabola, and when the source is exactly on its axis, we get a very strong intensity at the focus. In this case we understand very clearly what the resolving power of such an instrument is. The arranging of the antennae on a parabolic curve is not an essential point. It is only a convenient way to get all the signals to the same point with no relative delay and without feed wires. The angle such an instrument can resolve is still $\theta = \lambda/s$, where s is the separation of the first and last antennae. It does not depend on the spacing of the antennae and they may be very close together or in fact be all one piece of metal. Now we are describing a telescope in, of course. We have found the resolving power of a telescope! (Sometimes the resolving power is written $\theta = 1.22\lambda/D$, where D is the diameter of the telescope. The reason that it is not exactly λ/D is this: when we worked out that $\theta = \lambda/D$, we assumed that all the lines of dipoles were equal in strength, but when we have a circular telescope, which is the way we usually imagine a telescope, not as much signal comes from the outside edges, because it is cut like a square, where we get the same intensity all along a side. We get somewhat less because we are using only part of the telescope there; thus we can appreciate that the effective diameter is a little shorter than the true diameter, and that is what the 1.22 factor tells us. In any case, it adds a little pedantic to put such precision into the resolving power formula.)

20-5 Colored films; crystals

The above, then, are some of the effects of interference obtained by adding the various waves. But there are a number of other examples, and even though we do not understand the fundamental mechanism yet, we will see day, and we can understand even now how the interference occurs. For example, when a light wave hits a surface of a material with an index n , let us say at normal incidence, some of the light is reflected. The reason for the reflection we are not in a position to understand right now; we shall discuss it later. But suppose we know that some of the light is reflected both on entering and leaving a reflecting medium. Then, if we look at the reflection of a light source in a thin film, we see the sum of two waves; if the thicknesses are small enough, these two waves will produce an interference, either constructive or destructive, depending on the signs of the phases. It might be, for instance, that for red light, we get an enhanced reflection, but for

* This is because Rayleigh's criterion is a rough idea in the first place. It tells you where it begins to get very hard to tell whether one image was made by one or by two stars. Actually, if sufficiently careful measurements of the exact intensity distribution over the diffracted image (that can be made), the fact that two stars can make the spot can be proved even if $\lambda < k$, that is, even if λ/L .

blue light which has a different wavelength), perhaps we get a destructively interfering reflection, so that we see a bright red reflection. If we change the thickness, i.e., if we look at another place where the film is thicker, it may be reversed; the red interfering and the blue not, so it is bright blue, or green, or yellow, or whatever. So we see colors when we look at thin films and the colors change if we look at different angles, because we can appreciate that the thicknesses are different at different angles. Thus we suddenly appreciate another hundred thousand situations involving the colors that we see on oil films, soap bubbles, etc., at different angles. But the principle is all the same: we are only adding waves at different phases.

As another important application of diffraction, we may mention the following. We used a grating and we saw the diffracted images on the screen. If we had used monochromatic light, it would have been at a certain specific place. Then there were various higher-order images also. From the positions of the images, we never tell how far apart the lines on the grating were, if we knew the wavelength of the light. From the difference in intensity of the various images, we could find out the shape of the grating scratches, whether the grating was made of wires, scratches, matches, or whatever, without being able to see them. This principle is used to discover the positions of the atoms in a crystal. The only complication is that a crystal is three-dimensional; it is a repeating three-dimensional array of atoms. We cannot use ordinary light, because we must use something whose wavelength is less than the space between the atoms or we get no effect; so we must use radiation of very short wavelength, i.e., x-rays. So, by shining x-rays into a crystal, and observing how intense is the reflection in the various orders, we can determine the arrangement of the atoms inside without ever being able to see them with the eye. It is in this way that we know the arrangement of the atoms in various substances, which permitted us to draw those pictures in the first chapter, showing the arrangement of atoms in salt, and so on. We shall later come back to this subject and discuss it in more detail, and therefore we say no more about this most remarkable idea at present.

30-6 Diffraction by opaque screens

Now we come to a very interesting situation. Suppose that we have an opaque sheet with holes in it, and a light on one side of it. We want to know what the intensity is on the other side. What most people say is that the light shines through the holes, and produces no effect on the other side. It will turn out that are gets the right answer, in an excellent approximation, if one assumes that there are sources distributed with uniform density across the open holes, and that the phases of these sources are the same as they would have been if the opaque material were absent. Of course, actually there are no sources at the holes, in fact that is the only place that there are certainly no sources. Nevertheless, we get the correct diffraction pattern by assuming the holes to be the only places that there are sources; that is a rather peculiar fact. We shall explain later why this is true, but for now let us just suppose that it is.

In the theory of diffraction there is another kind of a situation that we should briefly discuss. It is usually not discussed in an elementary book so early as this, only because the mathematical formulas involved in deriving these results are rather elaborate. Otherwise it is exactly the same as we have been doing all along. All the interference phenomena are the same; there is nothing very much more or less involved, only the circumstances are more complicated and it is harder to add the waves together. That is all.

Suppose that we have light coming in from infinity, casting a shadow of an object. Figure 30-7 shows a screen on which the shadow of an object AB is made by a light source very far away compared with our wavelength. Now we would expect that outside the shadow, the intensity is all bright, and inside it, it is all dark. As a matter of fact, if we plot the intensity as a function of position near the shadow edge, the intensity rises and then overshoots, and wobbles, and oscillates about in a very peculiar manner near the edge (Fig. 30-8). We now shall discuss the reason for this. If we use the theorem that we have not yet proved, then we can

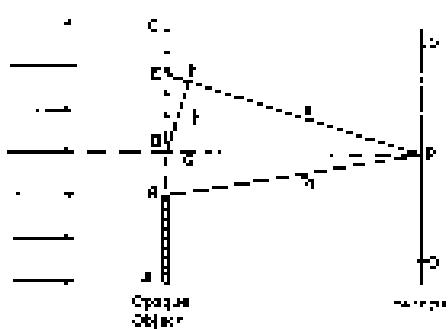


Fig. 30-7. A distant light source casts a shadow of an opaque object on a screen.

replace the actual problem by a set of effective sources uniformly distributed over the open space beyond the object.

We imagine a large number of very closely spaced antennas, and we want the intensity at some point P . That looks just like what we have been doing—very quiet; because our screen is at ∞ . Yet quiet; because our screen is at ∞ . We do not want the intensity at infinity, but, at a finite point. To calculate the intensity at some particular place, we have to add the contributions from all the antennas. First there is an antenna at D_1 , exactly opposite P ; if we go up a little bit in angle, let us say a height h , then there is an increase in delay (there is also a change in amplitude because of the change in distance, but this is a very small effect if we are at ∞ , far away, and is much less important than the difference in the phase). Now the path difference $\Delta P = DP$ is $h^2/2c$, so that the phase difference is proportional to the square of how far we go from D , while in our previous work it was infinite, and the phase difference was linearly proportional to h . When the phases are linearly proportional each vector adds at a constant angle to the next vector. What we now need is a curve which is made by adding a lot of infinitesimal vectors with the requirement that the angle they make shall increase, not linearly, but as the square of the length of the curve. To construct that curve involves slightly advanced mathematics, but we can always construct it by actually drawing the arrows and measuring the angles. In any case, we get the marvelous curve (called Cornu's spiral) shown in Fig. 30-8. Now how do we use this curve?

If we want the intensity, let us say, at point P , we add a lot of contributions of different phases from point D on up to infinity, and from D down only to point D_2 . So we start at D_2 in Fig. 30-8, and draw a series of arrows of ever-increasing angle. Therefore the total contribution above point D_2 all goes along the spiraling curve. If we were to stop integrating at some place, then the total amplitude would be a vector from D to that point; in this particular problem we are going to infinity, so the right answer is the vector B_{ext} . Now the position on the curve which corresponds to point D_2 or the object depends upon where point P is located, since point D , the reference point, always corresponds to the position of point P . Thus, depending upon where P is located above D , the beginning point will fall at various positions on the lower left part of the curve, and the resultant vector B_{ext} will have many maxima and minima (Fig. 30-9).

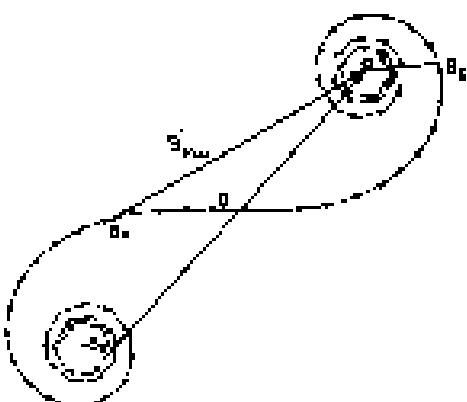


Fig. 30-8. The addition of amplitudes for many in-phase oscillators whose phase delays vary as the square of the distance from point D of the previous figure.

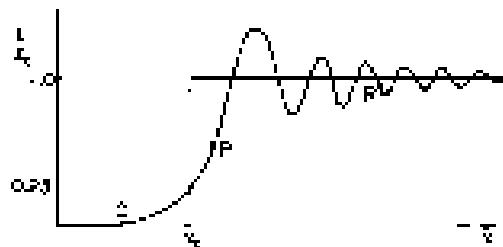


Fig. 30-9. The intensity near the edge of a shadow. The geometrical shadow edge is at x_0 .

On the other hand, if we are at Q , on the other side of P , then we are using only one end of the spiral curve, and not the other end. In other words, we do not even start at D , but at D_2 , so on this side we get an intensity which continuously falls off as Q goes further into the shadow.

One point that we can immediately calculate with ease, to show that we really understand it, is the intensity exactly opposite the edge. The intensity here is $1/4$ that of the incident light. Reason: Exactly at the edge (at the endpoint D of the arrow in Fig. 30-8) we have half the curve that we would have had if we were far into the bright region. If our point R is far into the light we go from one end of the curve to the other, that is, one full unit vector; but if we are at the edge of the shadow, we have only half the amplitude $\cdot 1/4$ the intensity.

In this chapter we have been finding the intensity produced in various directions from various arrangements of sources. As a final example we shall derive a formula which we shall need for the next chapter on the theory of the index of refraction. Up to this point relative intensities have been sufficient for our purpose, but this time we shall find the complete formula for the field in the following situation.

30-7 The field of a plane of oscillating charges

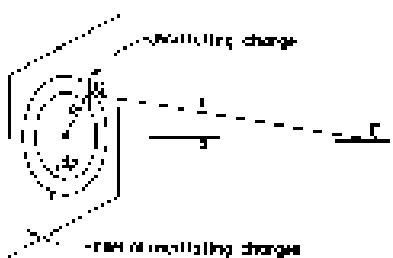


Fig. 30-10 Radiation field of a sheet of oscillating charges.

Suppose that we have a plane full of sources, all oscillating together, with their motion in the plane and all having the same amplitude and phase. What is the field at a finite, but very large, distance away from the plane? (We cannot get very close, of course, because we do not have the right formulae for the field close to the sources.) If we let the plane of the charges be the XY-plane, then we want to find the field at the point P far out on the Z -axis (Fig. 30-10). We suppose that there are n charges per unit area of the plane, and that each one of them has a charge q . All of the charges move with simple harmonic motion, with the same direction, amplitude, and phase. We let the motion of each charge, with respect to its own average position, be $x_1 e^{i\omega t}$. Or, using the complex notation and remembering that the real part represents the actual motion, the motion can be described by $x_1 e^{i\omega t}$.

Now we find the field at the point P from all of the charges by finding the field there from each charge q , and then adding the contributions from all the charges. We know that the radiation field is proportional to the acceleration of the charge, which is $-i\omega^2 x_1 e^{i\omega t}$ (and is the same for every charge). The electric field that we want at the point P due to a charge at the point Q is proportional to the acceleration of the charge q , but we have to remember that the field at the point P at the instant t is given by the acceleration of the charge at the earlier time $t - \tau = t - c/\rho$, where c/ρ is the time it takes the wave to travel the distance ρ from Q to P . Therefore the field at P is proportional to

$$-i\omega^2 x_1 e^{i\omega(t-\tau)} \quad (30.10)$$

Using this value for the acceleration to scan from P in our formula for the electric field at large distances from a radiating charge, we get

$$\left(\frac{\text{Electric field at } P}{\text{from charge at } Q} \right) = \frac{q}{4\pi\varepsilon_0 c^2} \frac{i\omega^2 x_1 e^{i\omega(t-\tau)}}{r} \text{ (approx.).} \quad (30.11)$$

Now this formula is not quite right, because we should have used c for the acceleration of the charge but for components perpendicular to the line QP . We shall suppose, however, that the point P is so far away, compared with the distance of the point Q from the axis (the distance ρ in Fig. 30-9), that those changes that we need to take into account, that we can leave out the cosine factor (which would be nearly equal to 1 anyway).

To get the total field at P , we now add the effects of all the charges in the plane. We should, of course, make a vector sum. But since the direction of the electric field is nearly the same for all the charges, we may, in keeping with the approximation we have already made, just add the magnitudes of the fields. To an approximation the field at P depends only on the distance ρ , so all charges at the same ρ produce equal fields. So we add, first, the fields of those charges in a ring of width $d\rho$ and radius ρ . Then, by taking the integral over all ρ , we will obtain the total field.

The number of charges in the ring is the product of the surface area of the ring, $2\pi\rho d\rho$, and n , the number of charges per unit area. We know, then,

$$\text{Total field at } P = \int_{\rho=0}^{\infty} \frac{q}{4\pi\varepsilon_0 c^2} \frac{i\omega^2 x_1 e^{i\omega(t-\tau)}}{r} \cdot n \cdot 2\pi\rho d\rho. \quad (30.12)$$

We wish to evaluate this integral from $\rho = 0$ to $\rho = \infty$. The variable r , of course, is to be held fixed while we do the integral, so the only varying quantities are ρ and n . Leaving out all the constant factors, involving the factor $i\omega^2$ for the moment, the integral we wish is

$$\int_{r=0}^{\infty} \frac{e^{-i\omega\tau}}{r} r^2 dr. \quad (30.13)$$

To do this integral we need to use the relation between r and ρ ,

$$r^2 = \rho^2 + z^2. \quad (30.14)$$

Since ϕ is independent of p , when we take the differential of this equation, we get

$$2r dr = 2a dp,$$

which is lucky, since in our integral we can replace $a dp$ by $r dr$ and the r will cancel the one in the denominator. The integral we want is thus the simpler one

$$\int_{r=0}^{r=\infty} e^{-\sigma r^2} dr. \quad (30.15)$$

To integrate an exponential is very easy. We divide by the coefficient of r in the exponent, and evaluate the exponential at the limits. But the limits of r are not the same as the limits of p . When $p = 0$, we have $r = \infty$, as the limits of r are $\pm\infty$ to infinity. We get for the integral

$$\int_0^\infty [e^{-\sigma r^2} - e^{-\sigma(\infty)^2}] dr, \quad (30.16)$$

where we have written ∞ for $(\infty)r$, since these both just mean a very large number!

Now $e^{-\infty}$ is a mysterious quantity. Its real part, for example, is $\cos(-\infty)$, which, mathematically speaking, is completely indefinite (although we would expect it to be everywhere—or everywhere (δ) —between -1 and $+1$). But in a physical situation, it can mean something quite reasonable, or usually can just be taken to be zero. To see that this is so in our case, we go back to consider again the original integral (30.15).

We can understand (30.15) as a sum of many small complex numbers, each of magnitude Δr and with the angle $\theta = -\sigma r^2$ in the complex plane. We can try to evaluate the sum by a graphical method. In Fig. 30-11 we have drawn the first five pieces of the sum. Each segment of the curve has the length Δr and is placed at the angle $\theta = -\sigma r^2$ with respect to the preceding piece. The sum for these first five pieces is represented by the arrow from the starting point to the end of the fifth segment. As we continue to add pieces we shall race off clockwise until we get back to the starting point (approximately) and then start around once more. Adding more pieces, we just go round and round, staying close to a circle whose radius is easily shown to be c/a . We can see now why the integral does not give a definite answer!

But now we have to go back to the effects of the situation. In any real situation the plane of charges cannot be infinite in extent, nor must it sometime stop. If it slopes suddenly, and was exactly circular in shape, our integral would have some value on the circle in Fig. 30-11. (If, however, we let the number of charges in the plane gradually taper off at some large distance from the center, or else stop suddenly but in an irregular shape so far larger p , the entire ring of width dp no longer contributes), then the coefficient a in the exact integral would decrease toward zero. Since we are adding smaller pieces but still going through the same angle, the graph of our integral would then become a curve which is a spiral. The spiral would eventually end up at the center of our original circle, as drawn in Fig. 30-12. The physically correct integral is the complex number A , in the figure represented by the interval from the starting point to the center of the circle, which is just equal to

$$\frac{c}{\infty} e^{-\sigma(\infty)^2}, \quad (30.17)$$

as you can work out for yourself. This is the same result we would get from Eq. (30.16) if we set $a = 0$.

(There is also another reason why the contribution to the integral drops off for large values of r , and that is the factor we have omitted for the projection of the acceleration in the plane perpendicular to the line $F(r)$.)

We are, of course, interested only in physical situations, so we will take c/a equal to zero. Returning to our original formula (30.12) for the field and putting

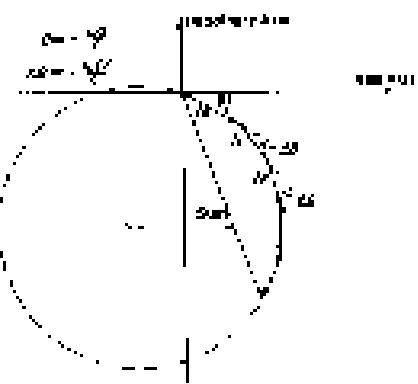


Fig. 30-11. Graphical solution of $\int_0^\infty e^{-\sigma r^2} dr$.

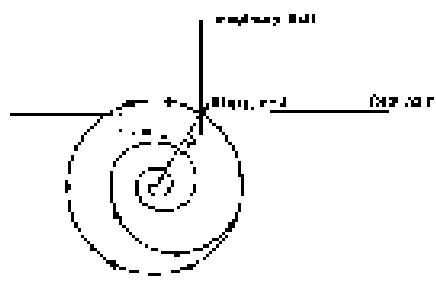


Fig. 30-12. Graphical solution of $\int_c^0 e^{-\sigma r^2} dr$.

Add all of the factors along with the integral, we have the result

$$\text{Total field at } P = -\frac{\lambda^2}{2\epsilon_0 c} \cos(\omega t) \quad (30.18)$$

(remembering that $\lambda = -\lambda$)

It is interesting to note that $\cos(\omega t)$ is just equal to the velocity of charge, so that we can also write the equation for the field as

$$\text{Total field at } P = -\frac{\lambda v}{2\epsilon_0 c} (\text{velocity of charge}) \cos(\omega t) \quad (30.19)$$

which is a little strange, because the reference is not to the distance r , which is the distance from P to the plane of charges. But that is the way it comes out—fortunately a rather simple formula. (We may add, by the way, that although our derivation is valid only for distances far from the source of oscillatory charges, it turns out that the formula (30.18) or (30.19) is correct at very distance r , even for $r < \lambda$.)

The Optics of the Refractive Index

31-1 The index of refraction

We have seen that light goes slower in water than in air, and slower again in air than in vacuum. This effect is described by the index of refraction n . Now you would like to understand how such a slower velocity could come about. In your lecture, we may try to see whether our current known physical assumptions or statements are violated; nothing, what were the following:

- That the total electric field in any physical circumstance can always be represented by the sum of the fields from all the charges in the universe.
- The charge field from a single charge is given by a mathematical calculation with a relationship to the speed c , namely (from the Coulomb field):

But, for a piece of glass, you might think, "Oh, no, you could modify all this. You might say it is retarded at the speed c/n ," thus, however, is not right, and we have no understanding why it is not.

It is approximately true that light or any other real wave does appear to travel at the speed c in a vacuum, whose index of refraction is 1. The field is proportional to the motion of all the charges. Including the charges moving in the material—and with these basic contributions of the field travelling at the speed of light c . Our problem is to understand how the apparently slower velocity comes about.

We shall try to understand the effect in a very simple case. A source which we shall call "the central source" is placed a large distance away from a thin plane of transparent material, say glass. We imagine that the field at a large distance R is the opposite side to the plane. The situation is illustrated by the diagram of Fig. 31-1, where S and P are imagined to be very far away from the plane. According to the principles we have stated earlier, the electric field E_0 where S is located, due to moving charges is the vector sum of the fields produced by the external source on S and the fields produced by each of the charges in the sheet of glass, every one due to propagation according to the retardation. Remember that the contribution of each charge is not changed by the presence of the other charges. Therefore our basic p analysis. The field at P can be written thus:

$$\mathbf{E} = \sum_{\text{charges}} \mathbf{E}_{\text{due to } S} \quad (31-1)$$

$$\mathbf{F} = \mathbf{E} = \sum_{\text{charges}} \mathbf{F}_{\text{due to } S} \quad (31-2)$$

where E_0 is the field due to the source alone and would be precisely the field if S were a stationary particle. We expect the field at P to be different from E_0 if there are any other moving charges.

Why should there be charges moving in the glass? We know that all the metal consists of atoms which contain e electrons. When the electric field of the source acts on these atoms it drives the electrons up and down, because it can be a force on the electrons. And moving charges can generate a field. They constitute new currents. These new currents are related to the source as E because they are driven by the field of the source. The total field is not just the field of the source E_0 but it is modified by the additional contribution from the other moving charges. This means that the field is not the same as the one at S was there before the glass was there but is modified. And it turns out that it is modified in such a way that

31-1 The index of refraction

31-2 The field due to the material

31-3 Dispersion

31-4 Absorption

31-5 The energy carried by an electric wave

31-6 Diffraction of light by a screen



Fig. 31-1. Electric waves passing through a sheet of transparent material.

the field inside the glass appears to be moving at a different speed. That is the β which we would like to work out quantitatively.

Now this is, in the exact case, actually impossible, because although we have said that all the other moving charges are driven by the source field, but is not quite true. If we think of a permanent charge, it feels only the source, but the rest of the world, it feels all of the other charges that are moving. It feels, in particular, the charges that are moving somewhere else in the glass. So the total force which it feels on a particular charge is a sum of two fields from the outside space, whose relative magnitudes from the permanent source is β/β_0 . You can see that it would need a complete set of equations to get the complete and exact formula. I say only today that we postpone this problem till next year.

Instead we shall work out a very simple case in order to understand what the physical principles very clearly. We take a circumstance in which the effects from the other source is very small compared to the effects from the source. In other words, we take a medium, in which the field is not modified very much by the motion of the other charges. That is, it appears to a charge in which the index of refraction is very close to 1, which will happen, for example, if the density of the source is very low. Our calculation will be valid for any case in which the index is for any reason very close to 1. In this way we shall avoid the complications of the more general, complete equation.

Finally, you should notice that there is another effect caused by the motion of the charges in the plane. These charges will also cause waves back toward the source. This backward-going field is the light we see reflected from the surfaces of prepared materials. It does not come from just the surface, but backward radiation comes from everywhere in the interior, but it has to escape the bulk after being equivalent to reflection from the surface. These reflection effects are beyond our approximation at the moment, because we shall be limited to calculating the material with an index so close to 1 that very little light is reflected.

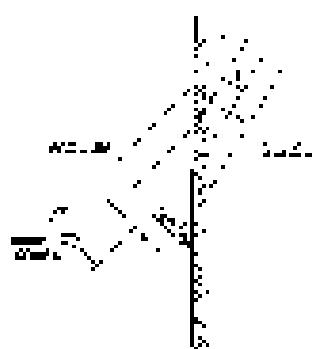


Fig. 37-2. Relation between reflection and refraction and velocity changes.

Let us proceed with our theory of how the index of refraction corresponds. We should understand first of all that it is caused by multiple reflections to understand why the apparent wave velocity is different in different materials. The bending of light rays comes when just because the effective speed of the waves is different in the materials. To remind you how that comes about we have drawn in Fig. 37-2 several successive parts of an acoustic wave which arises from a wave source not too far from a block of glass. The time is proportional to the wave crest indicates the direction of travel of the wave. Now, oscillations in the wave crest have the same frequency. (We have seen that light oscillates over the same frequency as the living source.) This means, then, that the wave crests in the wave can both cross the surface without loss of speed along the surface since they must travel together, so that a charge sitting at the boundary will feel only one frequency. The average distance between crests of the wave, however, is the wave length which is the velocity divided by the frequency. On the air side it is $\lambda_0 = v/v_0$, and on the other side it is $\lambda_1 = v/v_1$. From previous, $v_1 = c/n$ is the velocity of the wave. Since the light is taken out this is the only way for the wave to "fit" properly at the boundary is for the waves in the material to be traveling at a definite angle with respect to the surface. From the geometry of the diagram you can see that for $n > 1$ we must have $\lambda_1 < \lambda_0$, or $v_1 < v_0$, or $c/n < v_0$, which is usually true. We shall, for the rest of our discussion, consider only c/v , light has an effective speed of c/n in the rest of index n , and an longer way, in this chapter, about the bending of the light directions.

We go back now to the situation shown in Fig. 37-1. We see that what we have to do is to calculate the field produced by all the oscillating charges in the glass plate. We shall call this part of the field E_1 , and it is just as was written, as the second term in Eq. (37-2). When we add it to the field E_0 , due to the source, we will have the total field as E .

This is probably the most common misconception among students going into this year, but it is important to note that there are many pieces and have to be put together, each piece, however, is very simple. It's like when we build a model we say, "Forget the orientation, just look at the overall," in this case we don't need the orientation of each of the directions. In other words, the drag or retardation now is the physical machinery for the production of the index.

To see what we are going to do, let us take first of all what the "true current field" E_0 would have to be if the plate had no E_1 . It is going to look like this because from the source that is skewed down while passing through the thin plates. If the plate had no effect on it, the field of a wave travelling to the right (along the x -axis) would be

$$E_0 = E_0 e^{i(kx - \omega t)} \quad (31.3)$$

or, using the exponential notation,

$$E_0 = E_0 e^{i(kx - \omega t)} \quad (31.4)$$

Now what would happen if we were to pull more slowly in going through the plate? Let us call the slowness of the plate α . If the plate is constant there will be wave without train. It would not be in the true field. But it appears to travel at the speed of light it would take the longer time $\Delta t = \alpha$ in the additional time $\Delta t = \alpha = \Delta x/c$. After that it would continue to travel at the speed of light. We can take into account the extra delay by putting enough the index by replacing β in Eq. (31.4) by $\beta + \Delta t/c$ or by $\beta + (\alpha - 1) 2\pi/c$. So the wavefield in front of the plate should be written

$$E_{\text{travelled}} = E_0 e^{i(k(x - \Delta x) - \omega(t - \Delta t))} \quad (31.5)$$

We can also write this expression as

$$E_{\text{travelled}} = e^{-i\omega\Delta t} E_0 e^{i(kx - \omega t)} \quad (31.6)$$

which says that the wave after the plate is obtained from the wave which would exist without the plate, i.e., from E_0 , by multiplying by the factor $e^{-i\omega\Delta t}$. Now we notice that multiplying or dividing something like $e^{\beta x}$ by a factor $e^{\alpha x}$ just gives us the change the phase of the wave divided by the angle α , which is of course, where the extra delay is passing through the medium. So instead of this retarded field please let the amount add $(\alpha - 1) 2\pi/c$ (neglected, however, of the minus sign) in the exponent.

We have said earlier that the plate should add a field E_1 to the original field $E_0 = E_0 e^{i(kx - \omega t)}$, but we have found instead that the effect of the plate is to multiply the field by a factor $e^{-i(\omega\Delta t - k\Delta x)}$. However, that is really a right because we can get the same result by adding a suitable complex number. It is particularly easy to find the right number to add in this case since α is small. Do you still remember that if a is a small number then e^a is nearly equal to $1 + a$? We can write, therefore,

$$e^{-i(\omega\Delta t - k\Delta x)} = 1 - ik\Delta x - \frac{1}{2}k^2\Delta x^2 \quad (31.7)$$

Using this equality in Eq. (31.6), we have

$$\text{Cylinder} = E_0 e^{i(kx - \omega t)} - \frac{i\omega\Delta x - 1}{2}k^2\Delta x^2 E_0 e^{i(kx - \omega t)} \quad (31.8)$$

The first term is just the field from the source and the second term must just be equal to E_1 , the field produced to the right of the plate by the net trailing charges on the plate—represented here in terms of the index of refraction α , and depending, of course, on the slowness of the wave from the source.

What we have seen today is easily visualized. See back of the cylinder containing E_0 in Fig. 31.3. We first draw the number E_0 (not drawn to scale, but it is not too far from horizontal), but this is unnecessary). The easy thing to



Fig. 31.3. Diagram for the transmission of a wave through a cylinder.

slowing down in the plates would affect the phase of this number. That is, it would rotate E_0 through a negative angle. But this is equivalent to adding a small vector to E_0 roughly right angles to E_0 . But that is just what the factor $-is$ there in the second term of Eq. (31.8). It says that if E_0 is real, then E_0 is negative imaginary or, in general, E_0 and E have a right angle.

31-2 The field due to the anodes.

We now have to work out the field E_0 , which fits in the second term of Eq. (31.8) the kind we would expect from oscillating charges in the plates. If we can carry this on, we will then have calculated what the index s should be! (Since s is the only nondimensional number in Eq. (31.8).) We don't know exactly what field E_0 the charges in one anode will produce. (To help you keep track of the many symbols we have used up to now, and without having to do most of the calculation, we have put them all together in Table 31-1.)

Table 31-1
Symbols used in the calculations

E	= field from the source
E_0	= field produced by charges in the plate
dr	= distance of the plate
a	= perpendicular distance from the plane
n	= index of refraction
ω	= frequency (angular) of the oscillation
N	= number of charges per unit volume in the plate
q	= number of charges per unit area of the plate
q_0	= charge on an electron
m	= mass of an electron
v_0	= received frequency of an electron beam in an atom

If the source S (of Eq. 31-1) is to drift to the left, then the field E_0 will have the same phase everywhere in the plane, so we can write that to the right hand side of the plate

$$E = E_0 e^{i(\omega t - \phi)} \quad (31.9)$$

Right of the plate, where $a = 0$, we will have

$$E_0 = E_0 e^{i\omega t} \text{ (at the plate)} \quad (31.10)$$

All of the electrons in the chains of the plate will feel this electric field and will be driven up and down (assuming the direction of v_0 is vertical) by the electric force qeE . To find who moves we expect the electrons, we will assume again among the little oscillators, that is, that the electrons are confined elastically to the plate, which means that if a force is applied to an electron its displacement from its normal position will be proportional to the force.

You may think that this is a fairly simple model of an atom. You have heard some electrons whirling around in orbits. But that is just an oversimplified picture. The correct picture of an atom, which is given by the theory of wave mechanics, says that, as far as problems involving light are concerned, the electrons behave as though they were held by springs. So we shall suppose that the electrons have a linear restoring force, which, together with their mass m , makes them behave like little oscillators with a resonant frequency ω_0 . We have already studied such oscillators, and we know that the equation of their motion is written this way:

$$m \left(\frac{d^2y}{dt^2} + \omega_0^2 y \right) = F_0 \quad (31.11)$$

Where F_0 is the driving force.

For our problem, the driving force comes from the electric field of the wave from the source, given by (31.1), we

$$F = q_0 E_0 = q_0 \epsilon_0 E_0^{\text{ext}}, \quad (31.12)$$

where q_0 is the electric charge on the electron and for E_0 , we use the expression $E_0 = E_0^{\text{ext}}$ from (31.1). Our equation of motion for the electron is then

$$m \left(\frac{d^2 r}{dt^2} + \omega_0^2 r \right) = q_0 \epsilon_0 E_0^{\text{ext}}. \quad (31.13)$$

We have solved this equation before, and we know that the solution is

$$r = r_0 e^{i\omega t}, \quad (31.14)$$

where, by substituting in (31.13), we find that

$$\omega_0 = \frac{q_0 E_0}{m(\omega_0^2 - \omega^2)}, \quad (31.15)$$

so that

$$r = \frac{q_0 E_0}{m(\omega_0^2 - \omega^2)} e^{i\omega t}. \quad (31.16)$$

We have just we needed to know—*the motion* of the electrons in the plate. And it is the same in every electron wave, that the mean position (the “*dc*” of the motion) is, of course, different for each electron.

Now we are ready to find the field E_0 that these atoms produce at the point x . Because we have already worked out (in the end of Chapter 20) what field is produced by a sheet of charges that all move together. Referring back to \mathbf{E}_0 (30.19), we see that the field E_0 at x is just a negative constant times the velocity of the charges, divided by time the amount a_0 . Substituting r_0 in Eq. (31.16) to get the velocity, and sticking in the substitution r_0 just putting r_0 from (31.15) into (30.19) yields

$$E_0 = -\frac{q_0}{2a_0} \left[\ln \frac{q_0 E_0}{m(\omega_0^2 - \omega^2)} e^{i\omega_0 a_0/\omega} \right]. \quad (31.17)$$

Just as we expected. To draw the line of the electrons produced an excited wave which travels to the right (that is what is factor $e^{i\omega_0 a_0/\omega}$ does), and the amplitude of the wave is proportional to the number of atoms per unit area in the plate (proportional to n) and is proportional to the strength of the electric field (the factor E_0). Then there are some factors which depend on the atomic properties (e.g., m) and which we should expect.

The most important thing, however, is that this formula (31.17) for E_0 looks very much like the expression for E_0 that we got in (27.10.6) by solving (27.10.6) when we was defined a positive dielectric material with an index of refraction n . The two expressions are, in fact, identical if

$$(n - 1)^2 = \frac{\omega^2}{2\epsilon_0 m(\omega_0^2 - \omega^2)}. \quad (31.18)$$

Hence the field E_0 is proportional to Δn , since n , which is the number of atoms per unit area, is equal to $\Delta n N$, where N is the number of atoms per unit volume of the metal. So, returning to (31.16) and changing the frequency from ω , result, a formula for the index of refraction in terms of the properties of the atoms of the material—and of the frequency of the light:

$$n = 1 + \frac{q_0^2}{2\epsilon_0 m(\omega_0^2 - \omega^2)}. \quad (31.19)$$

This equation is very similar to the “*drift-disk*” of the index of refraction that we wished to obtain

31-3 Dispersion

In the last section we have obtained something very interesting. For we have not only assumed that the index of refraction which can be computed from the basic optical quantities, but we can also learned how the index of refraction should vary with the frequency ω of the light. This is something we could never understand from the simple statement that "light travels slower in a transparent material." We still have the problem, of course, of knowing how many oscillations per unit of time there are, and what is their natural frequency ω_0 . We do not know this just yet, because it is different for every different material, and we cannot yet get a general theory of the way. Recomputing our general theory of the properties of different substances their natural frequencies and comparing with quantum atomic mechanics. Also, different materials have different properties and different indexes, so we cannot expect, anymore, to get a general formula for the index which will apply to all substances.

However, we shall discuss the formula we have obtained in various possible circumstances. First of all, for most ordinary gases (for instance, for air), these values ω_0 (hydrogen, helium, and so on) the natural frequencies of the electron orbits will correspond to visible light. These frequencies are higher than the frequencies of visible light, that is, ω_0 is much larger than visible light, and so if this appears more or less transparent, we can dispense ω^2 in comparison with ω_0^2 . Then we find that the index is nearly constant. So for a gas, the index is nearly constant. This is also true for some other transparent substances like glass. If we have a gas, especially a little more closely, however, we notice that as we take away a little bit more away from the denominator, the index decreases. It decreases slowly with frequency. The index is higher for blue light than for red light. That is, the reason why we perceive the light more in the blue than in the red.

The phenomenon that the index depends upon the frequency is called the phenomenon of dispersion, because it is the basis of the fact that light is "dispersed" by a prism into a spectrum. The expression for the index of refraction as a function of frequency is called a dispersion equation. So we've now derived a dispersion equation. (In the past few years "dispersion equations" have been finding a new use in the theory of elementary particles.)

Our dispersion equation suggests other interesting effects. If we have a natural frequency ω_0 which lies in the visible region, or if we measure the index of refraction of a material like glass in the ultraviolet, where ω_0 is probably ω_0 , we can take frequencies which are lower than the natural frequency. Then the index will automatically be large because the denominator will go to zero. Next, suppose that ω is greater than ω_0 . This would mean, for example, if we take a material like glass, say, and shingle it with aluminum. In fact since many metals which are opaque to visible light, like graphite for instance, are transparent to X-rays, we can also talk about the index of refraction of graphite to X-rays. All the natural frequencies of the carbon atoms would be moved lower than the frequency we are using in the process, since ω_0 is still too small at a very high frequency. The index of refraction is thus given by our dispersion equation if we set ω_0 equal to zero (we neglect ω^2 , compared with ω_0^2).

A similar situation could occur if we beam electrons (or light) on a gas of free electrons. Let us again suppose electrons and like nuclei to be stationary by ultraviolet light from the sun. They will strip there as free electrons. Our free electrons are subject to electric accelerating forces. Setting $m_e = 1$, the dispersion equation after years the correct formula for the index of refraction for ultraviolet to the ultraviolet, where this corresponds to the density of free electrons (must be zero in practice) in the ultraviolet. (C. let me look again at the calculation. If we take ω to be the frequency of radio waves, or the ultraviolet, the term $(\omega - \omega_0)^2$ becomes negative, and we obtain the result that it is less than one. That means that the effective speed of the wave in the substance is faster than c. Can that be correct?)

It is correct. Despite of the fact that it is said that you cannot send signals with the speed of light, it is nevertheless the case that a signal of frequency ω can travel at a particular frequency even faster than c. See this page.

just receive the "radio wave" which is produced by the two train lights now to side, positive or negative. It can be shown, however, that the speed at which you can send a signal is not determined by the frequency or the frequency, but depends on what the index is for waves frequencies. What the index tells us is the speed at which the waves (or crests) of the wave travel. The wave you have is not a signal by itself; it is a radio wave, which has no oscillations of its kind, i.e., which is a steady oscillation, you cannot call it a radio wave; if so, you cannot send them a timing signal. In order to send a signal you have to change the wave somehow, add to it something in it, make it a little bit faster or slower. That means that you have to have more than one frequency in the wave and it can be seen that the speed of which which travels is not dependent upon the radio signal but upon the wave, that the index changes with the frequency. This subject we must also delay until Chapter 45. Then we will calculate for you the exact speed of signals through such a piece of glass, and you will see that it will not be faster than the speed of light, although the index, when the mathematical point, does not exceed twice the speed of light.

Now, to give a slight hint as to how this happens, you will note that the field definitely goes with the fact that the responses of the charge are opposite to the field, i.e., the sign has gone around well. Thus in our expression for \ddot{x} (Eq. 31.15) the displacement of the charge is in the direction opposite to the driving field. And it is there that the phase of the transmitted field can appear in the advanced wave respects the source wave. It is this advance property which is true, either we say that the "phase velocity" or velocity of the nodes is greater than c . In Fig. 31-4 we give a schematic idea of how the wave might look for a case where the wave is suddenly increased to take a signal. You will see from the diagram that the signal (i.e., the part of the wave) is run earlier in the wave which ends up with an advance in phase.

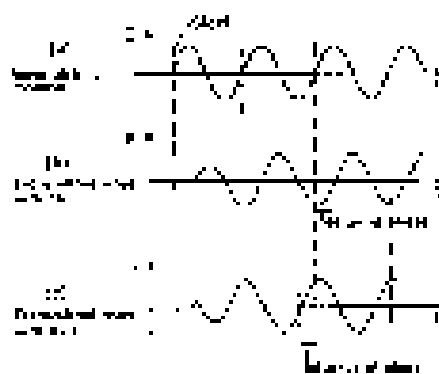


Fig. 31-4. Wave "signals."

Let us now look again at our previous equation. We should remark that on any kind of quantitative index gives a result that is somewhat simpler if we could actually find in nature. To be completely accurate we ought to make refinements. First, we should expect that our mass of the charge in oscillation should have some damping forces (like viscous forces), if we go with a force, and we do not expect that to happen. We have worked out before (Eq. 31.8) the motion of a damped oscillator and the result is not the last number in Eq. (31.6b), and therefore in (31.10), is changed from $(\omega_0^2 - \omega^2)$ to $(\omega_0^2 - \omega^2 + \gamma\omega)$, where γ is the damping coefficient.

We shall now add (31.10) to take care of this. In fact that there are several resonant frequencies for a particular kind of system. It is very important

Dispersion is got by imagining that there are several different kinds of oscillators, but each has a different frequency, and so we simply add the contributions of all the oscillators. Let us say that there are N_1 electrons per unit of volume whose natural frequency is ω_1 and whose damping factor is γ_1 . We would then have the sum dispersion equation

$$n = 1 + \frac{c^2}{\omega_1^2} \sum_{k=1}^{\infty} \frac{N_k}{\omega_k^2 + \gamma_{k,0}^2}. \quad (31.20)$$

We have finally a complete expression which describes the index of refraction that is observed for many substances.⁴ The index described by this formula varies with frequency non-linearly, as the curve shows in Fig. 31-3.

You will notice that so long as it is close to one of the natural frequencies, the slope of the curve is positive. Such a positive slope is referred to as "normal" dispersion (as it is clearly the most common occurrence). However, the natural frequencies, however, there is a range of ω for which the slope is negative. Such a negative slope is often referred to as "anomalous" dispersion (also called dispersive), because it was not known when it was first observed, long before anyone ever knew there were such things as electrons. From our point of view both slopes are quite "normal".

31-4 Alternating

Perhaps you have noticed something a little strange about the form (Eq. 31.20) we obtained for our dispersion equation. Because of the term (γ) we put in to take account of damping, the index of refraction is now a complex number. What does that mean? By working out what the real and imaginary parts of n are we could write

$$n = n' - i n'', \quad (31.21)$$

where n' and n'' are real numbers. (We use the minus sign in front of the $i n''$ because then, if n'' turns out to be a positive number, as you can show for yourself, n

will be a *real* number and a complex index of refraction, by going back to Eq. (31.6), which is the equation of the phase shift in passing through a plate of material with thickness d . If we put $n = n' + i n''$ in the equation, and do some rearranging, we get

$$\text{Phase shift} = e^{-i k d} = e^{-i \frac{2 \pi}{\lambda} n' d} e^{-\frac{2 \pi}{\lambda} n'' d \cos^2 \theta}, \quad (31.22)$$

The last factors, marked " i " in Eq. (31.22), are just the form we had before, and again describe a wave whose phase has been delayed by the angle $\omega n'' = -1/2 \pi$ in traversing the material. The first term (n') is now written as exponential factor with a non-exponent, because there are γ 's that are canceled. Also, the exponent is negative, so the factor is a real number less than 1. It describes a decrease in the magnitude of the field due to the absorption by an amount which is more the larger ω is. As the wave goes through the material, it is weakened. The material is "dissipating" part of the wave. The wave carries no energy, so it will lose energy. You won't be surprised at this, because the damping we put in (γ) for one reason is indeed a loss of energy and must be expected to cause a loss of energy. We see that the "imaginary part" of a complex index of refraction represents an absorption (or "dissipation") of the wave. In effect, n'' is sometimes referred to as the "absorption index".

We may also point out that an imaginary part of the index corresponds to breaking the arrow \vec{E}_0 in Fig. 31-3 toward the origin. It is clear why the unamplified field is then decreased.

⁴ Actually, although the present analysis (Eq. 31.20) is still valid, its interpretation is somewhat different. In general, oscillators can have very wide band widths and often have several natural frequencies. Therefore N_k is not really the number of electrons having the frequency ω_k ; it is N_k divided by N , where N is the number of electrons, with N_k small. It is called a *weight* or *strength* that tells how strongly the mean velocity v of electrons of frequency ω_k .

Normally, the index as in glass, the absorption of light is very small. The δ factor reported from our Eq. (11.20), however, the imaginary part of the dielectric, δ_{abs} , is much smaller than the term $(\omega - \omega_0)$. But if the light frequency is very close to ω_0 , then the resonance term $(\omega - \omega_0)^2$ can become large compared with δ_{abs} and the index becomes almost completely imaginary. The absorption of the light becomes very dominant. It is just this effect that gives the dark lines in the spectrum of light which we receive from the sun. The light from the solar surface has passed through the sun's atmosphere (as well as the earth's) and the light has been strongly absorbed at the resonant frequencies of the atoms in the solar atmosphere.

The observation of such spectral lines in the sunlight allows us to tell the chemical composition of the atoms and hence the chemical composition of the atmosphere. These multi-wavelength observations tell us about the materials at the stars. From such measurements we know that the chemical elements in the sun and in the stars are the same as those we find at the earth.

11-5 The energy carried by an electric wave

We have seen that the imaginary part of the index means absorption. We shall now use this knowledge to find out how much energy is carried by a light wave. We have given earlier an argument that the energy carried by light is proportional to $\overline{E^2}$, the time average of the square of the electric field in the wave. The decrease in Δ due to absorption must mean loss of energy, which would go into the motion of the electrons and, we might guess, would end up as heat in the material.

If we consider a light arriving at a unit area, say one square centimeter, to compute in Eq. (31.1), then we obtain the following energy input per unit time (if we assume that energy is conserved, as we do):

$$\text{Energy in per sec} = \text{energy out per sec} = \text{work done per sec.} \quad (11.21)$$

For the first term we can write $\alpha \overline{E^2}$, where $\alpha \gg 1$ is yet another constant of proportionality which defines the average value of the energy being carried by the second term we must include the part from the radiating atoms of the material, so we should use $\alpha \overline{E^2} + N \sqrt{\epsilon}$, or rewriting the equation slightly $\overline{E^2} (\overline{E^2} + \Sigma E)$.

A lot of calculations have been made for a thin layer of material whose index is not too far from 1, so that δ , which is much less than E , just to make the calculations easier. In keeping with our approximate terms, we should, therefore, leave out the term ΣE , because it is much smaller. Only $\overline{E^2}$. You may say: "Then you should leave out $\alpha \overline{E^2}$, too, because it is much smaller than E ." I agree that $\overline{E^2}$ is much smaller than E , but we must keep $\alpha \overline{E^2}$ or our approximation will not be exact. Just as well apply if we neglected the presence of the material completely! One way of checking the calculations is to consider Δ to see that we always keep terms which are proportional to $N \delta z$, the areal density of atoms in the material, but we know a lot more which are proportional to $(N \delta z)^2$ or are higher powers of $N \delta z$. That is what would be called a "second-order approximation."

In the same spirit, we might remark that an energy injection has reflected the energy in the reflected wave. But this is OK because this term is also proportional to $N \delta z$, since the amplitude of the reflected wave is proportional to δz .

For the last term in Eq. (11.21) we wish to compute the rate at which the incoming wave is doing work on the electrons. We know that work is force times distance, so the rate of doing work (or emitted power) is $F \cdot V$, where V is the velocity. I, I, I, and $F \cdot V$, but we do not need to worry about the dot product when the velocity and force are along the same direction as they are here (except for a possible minus sign). So for each atom we have $\frac{1}{2} m v^2$ for the average rate of

doing work. Since $\nabla \cdot \mathbf{E} = 0$ at ∞ , as in a free space, the last term in Eq. (31.27) drops to 0 . Thus, our energy equation looks like

$$dE_t = dE_0 + 2\pi k_e E_0 + N \Delta \epsilon_0 E_0. \quad (31.28)$$

The E_0 is measured, and we have

$$2\pi k_e E_0 = N \Delta \epsilon_0 E_0. \quad (31.29)$$

We now go back to Eq. (31.27), and we notice that due to $\nabla \cdot \mathbf{E} = 0$,

$$\Delta \epsilon_0 = \frac{N \Delta \epsilon_0}{2k_e}, \text{ neglecting } \epsilon_0^2. \quad (31.30)$$

(noticing that $\epsilon_0 = N \Delta \epsilon_0$). Putting Eq. (31.30) into the left-hand side of (31.27), we get

$$2\pi \frac{N \Delta \epsilon_0}{2k_e} \overline{\epsilon_0(\vec{r}, t)} = \text{neglect } \epsilon_0^2.$$

However, $\epsilon_0(\vec{r}, t) \leq \mathcal{E}_0$ (at $t = 0$) multiplied by ϵ_0/ϵ_0 . Since the average is independent of time, it is the same quantity as calculated by ϵ_0/ϵ_0 , or is $\mathcal{E}_0(t)$ along \vec{r} , the same average that appears on the right-hand side of (31.28). The two sides are therefore equal to

$$\frac{1}{2} \epsilon_0 \mathcal{E}_0^2 = 1, \quad \text{or} \quad \mathcal{E}_0^2 = 2\epsilon_0. \quad (31.31)$$

We have discovered that if energy is to be conserved, the energy started in an infinite wavelet at $t = 0$ is now per unit time (frequency) we have called the intensity, and is given by $\epsilon_0/2\epsilon_0$. If we call the intensity I , we have:

$$I = \begin{cases} \text{Intensity} \\ \text{or} \\ (\text{energy}/\text{area}/\text{time}) \end{cases} = \epsilon_0/2\epsilon_0. \quad (31.32)$$

which further proves the law of energy. We have a similar message from quantum theory of the electrons indeed!

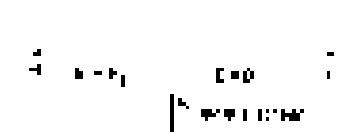


Fig. 31-4. Diffraction by a screen.

31-4. Diffraction of Light by a screen

It is now a good time to take up a somewhat different question which we can handle well: the machinery of this chapter. In the last chapter we said that when you have a discontinuity of the light cone coming through a screen, the direction of an electric field, the diffraction pattern, could be obtained by imagining instead that the holes were replaced by sources (positive or negative) distributed over the holes. In other words, the reflected wave is the same as though the hole were a new source. We learned earlier that the screen functioned, because the hole is, of course, just where there are no charges, where there are no canceling charges.

Let us first ask, "What is an opaque screen?" Suppose we have a completely opaque screen between a source S and an observer at P , as in Fig. 31-4(a). If the screen is "opaque" there is no field at P . Why is there no field there? According to the basic principle of superposition the field \mathbf{E}_{ext} at the point P is the sum of the fields delayed since the field from S the other charges provided. \mathbf{E}_{ext} , as we have seen above, the charges at the screen will be set to motion by the field \mathbf{E}_0 , and these charges generate a new field which, if the screen is opaque, must cancel exactly the field \mathbf{E}_0 on the back side of the screen. You say: "Well, it's a miracle that it becomes opaque!" Suppose it was not exactly right! If it were not exactly right, (remember that this opaque screen has some thickness), the field beyond the rear part of the screen would not be exactly zero. So, we being wise, it would set into motion some other charges on the inner side of the screen, and this makes \mathbf{E}_{ext} more \mathbf{E}_0 trying to get the total field near zero. So if we make the screen more and more opaque, there is no residual field, because there is enough opportunity to totally get the thing quieted down. Just as when S on page 5 above we should say that is 31-10

screen does not yield image rays, so we have a reflected expansion of the light source. This shows, however, that this enough of the incident material, even gold, is transparent.

Now let us see what happens with an opaque screen which has holes in it, as in Fig. 37.3(b). What do we expect for the field? The field at P can be approximated as a sum of two parts: the first due to the source, E_{source} , the field E_{wall} of the wall, i.e., due to the motions of the charges in the walls. We might expect the motions of the charges in the wall to be negligible, but we can find out this first very quickly in a rather simple way.

Suppose that we were to take the same screen but just up to the index, illuminated in just the left of the figure. We imagine that no charge moves by the entire distance to the wall. And if you look at the places where the holes were at least λ . Now let us calculate the field at P . The field E_{source} is necessarily zero in case (b), but there *is* a change in the field from the source plus the field due to the motions of the atoms in the wall and in the plug. We can write the following equations:

$$\text{Case (b)}: E_{\text{at } P} = E_{\text{t}} + E_{\text{wall}},$$

$$\text{Case (c)}: E_{\text{at } P} = 0 = E_{\text{t}} + E_{\text{wall}} - E_{\text{plug}},$$

where the signs refer to the case where the plug is in place, b, or is of no use, the same as both cases. Now, if we subtract the two equations, we get

$$E_{\text{at } P} = (E_{\text{wall}} - E_{\text{t}}) - E_{\text{plug}},$$

Now if the holes are not too small (say many wavelengths in size), we would not expect the presence of the plug to change the backward arrow of the wall except possibly for a little bit around the edges of the holes. Neglecting this small effect, we expect $E_{\text{wall}} = E_{\text{t}}$ and obtain that

$$E_{\text{at } P} = -E_{\text{plug}},$$

We have the result that the field at P when there are holes in a screen (case (b)) is the same (except for sign) as the field that is produced by that portion of a complete screen with which is located near the hole and of the sign is just right which, since we are usually interested in intensity, which is proportional to the square of the field. It seems like an amazing result and it is no argument. It is, however, not only this approximately true, but too sure holding, but useful, and is the justification for the usual theory of diffraction.

The field E_{plug} is computed in any particular manner by considering all the motion of the charge wherever in the screen is just right which will cancel out the field E_{t} on the back of the screen. Once we know the motions, we add the radiating fields at P due just to the charges in the plug.

We remark again that this theory of diffraction is only approximate, and will be good only if the holes are not too small. For holes which are too small the E_{plug} will not be small and then the difference between E_{wall} and E_{t} , (which difference we have to be zero) will be comparable to or larger than the small E_{t} , hence our approximation will no longer be valid.

Radiation Decaying, Radio Scattering

32-1 Radiation redshift

In the last chapter we learned that when a system is oscillating, energy is carried away, and we derived a formula for the energy which is radiated by an oscillating system. If we know the electric field, then the average of the square of the force times a_0 is the amount of energy that passes per square meter per second through a surface normal to the direction in which the radiation is going.

$$S = \text{cav}(E^2). \quad (32.1)$$

Any oscillating charge radiates energy; for instance, a short antenna radiates energy. If the system radiates energy, then in order to account for the conservation of energy we must add that power is being delivered along the wires which feed into the antenna. That is, in the driving circuit the antenna acts like a transformer, or a power source, energy can be "lost" if the energy is not really lost, it's really radiated out, but as far as the circuit is concerned, the energy is lost. In an ordinary resistance the energy which is "lost" goes into heat. In this case the energy which is "lost" goes out in a wave. So, from the standpoint of circuit theory, without considering how the energy goes, the net effect on the circuit is heating—energy is "lost" from the circuit. Therefore the circuit appears to the generator as having a resistance, even though it may be made will be hard, good copper. In fact, if it is well built it will appear as almost a pure inductor with very little inductance or capacitance because we would like to radiate as much energy as possible out of the system. This resistance that an antenna shows is called the radiative resistance.

If a current I is going in the antenna, then the average rate of stored power S delivered to the antenna is the average of the square of the current times the resistance. The rate at which power is radiated by the antenna is proportional to the square of the current, in the antenna, of course, because all the fields are proportional to the currents, and the energy losses are proportional to the square of the field. The coefficient of proportionality between radiated power and (I^2) is the radiative resistance.

An interesting question is, what is this radiative resistance due to? Let us do a simple calculation, let us say that currents are carried up and down in an antenna. Well, first that we have to put work in. If the antenna is to radiate energy, if we take a charged body and accelerate it up and down it radiates energy; if a proton, charged, would radiate energy. It is one thing to calculate from the conservation of energy that energy is lost, but another thing to answer the question, against whom are we doing the work? That is an interesting and very difficult question which has been incompletely and satisfactorily answered for electrons, although it has been for photons. What happens is this: in an antenna, the fields generated by the moving charges in one part of the antenna react on the moving charges in another part of the antenna. We can calculate these forces and find out how much work they do, and so find the right rule for the radiation resistance. When we say "We can calculate," that is not quite right—we cannot, because we cannot yet calculate the laws of electrodynamics, at least not now, and, at best, certainly do we know what the electric field is. We know the formula (23.1), but in present

is not considered how to calculate the fields inside the wave zone. Of course, energy conservation of energy is valid, we can calculate the result all right without knowing the fields at short distances. This is another of the reasons why using this argument is difficult. Thus, we can't find the formula for the forces at short

32-2 Radiation resistance

32-2 The rate of radiation of energy

32-3 Radiation damping

32-4 Independent sources

32-5 Scattering of light

distance a , by knowing the field at every large distance, by using the law of conservation of energy, he can go larger or smaller than zero.

The problem is the case of a single electron & this, if the e is only one charge, what can the force be? It has been proposed in current string theory, that the charge e is a little ball and that one part of the charge acts on the other part. Because of the decay in the collision across the fine electron, the front is not exactly in phase with the medium. That is, if we know the electric scattering shift, we know the "position" of the source. So the various internal forces are constant and there is no net force. But if the electron is accelerating, the front of it, the time delay Δt , the front which is moving on the front from the back, the back is not exactly due to the force on the back & the front, because of the delay in the effect. This leads to the situation, that is a lack of the effect, or a small effect, the thing holds itself back by its own traps. This model of the origin of the resistance to acceleration, the radiation resistance of a moving particle, has now many difficulties, because this present view of the electron is, that it is very little ball. If this problem has never been raised. Nevertheless we can conclude exactly, what the net radiation resistance is, and that is, how much loss. It is, must be with the accelerated charge, in spite of not knowing directly the mechanism of how that loss comes.

32.2 The rate of radiation of energy

Now we can calculate the total energy radiated by an accelerating charge. To keep the discussion general, we shall take the case of a charge moving in any which way, but nonrelativistically. At a moment when the acceleration a , say, vertical, we know that the electric field that is generated is the charge multiplied by the projection of the velocity vector, divided by the dielectric. So we know the electric field at any point, and we therefore know the square of the electric field and also the energy $c^2 E^2$ passing through a unit area per second.

This quantity $c^2 E^2$ is quite often in experiments involving radiation proportional to the square of frequency, and it is an easy number. For instance, it is the result for $\nu = 337$ ohm. So the power, or work per second, is equal to the average of the field squared, divided by 4π .

Using our expression (32.1) for the electric field, we find that

$$S = \frac{e^2 c^2 \sin^2 \theta}{mc^2 \epsilon_0 \pi r^2} \quad (32.2)$$

is the power per second, radiated in the direction θ . We notice that it goes down as the square of the distance, as we said before. Now, if we wanted the total energy radiated at all directions, then we must integrate (32.2) over all directions. First, we multiply by the $d\Omega$. To find the solid angle, it flows $d\Omega$ in a little angle $d\Omega$ (F_θ). Now, we need the area of a spherical sector. The way to think of it is this: there is the radius, then the width of the annular segment is $r d\theta$, and the circumference is $2\pi r \sin \theta$, because $\sin \theta$ is the radius of the circle. So the area of the little piece of the sectors is $2\pi r \sin \theta d\theta \cdot r d\theta$.

$$dA = 2\pi r^2 \sin \theta d\theta \quad (32.3)$$

By multiplying the term (32.2) the power per surface meter, by the area in square meters included in the small angular section $d\Omega$, the amount of energy that is liberated in this small bin between $\theta = 0 - d\theta$, then we integrate this over all the angles from 0 to 2π :

$$J = \int_{0}^{2\pi} S d\Omega = \frac{c^2 e^2}{512 \pi \epsilon_0 r^2} \int_{0}^{\pi} \sin^2 \theta d\theta \quad (32.4)$$

By writing $\sin^2 \theta = (1 - \cos^2 \theta)/2$ and using $\int_0^\pi \cos^2 \theta d\theta = \pi/2$. Using this fact, we finally get

$$P = \frac{dE/dt}{4\pi r^2 c^2} \quad (32.5)$$

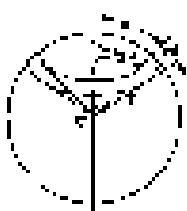


Fig. 32.1. The case of a spherical object of $2\pi r^2 \sin \theta d\theta$.

This expression deserves some remarks. First of all, since the value of $\langle \hat{w}^2 \rangle$ depends on the a^2 in (32.5) we shall be too scared of the vector \mathbf{w} , that is, \sqrt{a} , the length of the vector squared. Secondly, the flux (32.2) was calculated using the mean ω , acceleration; that is, the acceleration at the time t at which the energy was passing through the sphere was radiated. We might like to say that the energy was in the lines of \mathbf{E} at this time. This is not exactly true, it is only an approximation. The exact time when the energy is released can not be deduced precisely. All we can really calculate precisely is what happens in a complete motion, like an oscillation or something where the acceleration finally comes. Then when we find is that the total energy flux per cycle is the average of acceleration squared, for a complete cycle. This is what should really appear in (32.2), i.e., if it is radiation with an acceleration that is finite, y and finally zero, then the total energy loss per second is the time integral of (32.2).

To illustrate the consequences of formula (32.2) let us have an oscillating system in mind, what happens if the displacement x of the mass goes oscillating so near the acceleration $a = -\omega_0^2 x/c^2$? The average of the acceleration squared over a cycle trembles and we have to be very careful with the square things that get arising in our calculation. It really is the cosine, and the average of \cos^2 is obviously unity.

$$\langle x^2 \rangle = \langle x^2(t) \rangle$$

Integration:

$$P = \frac{q^2 \omega_0^2 / c^2}{12 \pi^2 f^3}. \quad (32.6)$$

The formulas we are now discussing are relatively advanced and more or less modern; they date from the beginning of the twentieth century, and they are very famous. Because of their historical value, it is important for us to be able to understand them in older books. In fact, the older books also used a system of units different from our present-day system. However, all these complications can be avoided easily in the final formulas derived with electrons by the following rule. The quantity q^2/c^2 , where q is the electronic charge (in coulombs), and, naturally, being greater or c^2 , it is very easy to calculate it even in the old system as numerically equal to 1.5158×10^{-16} , because we know that, numerically, $q_e = 1.60205 \times 10^{-19}$ and $c = 2.99792 \times 10^8$. Therefore we shall often use the convenient old notation

$$P = \frac{q^2}{c^2 \pi r}. \quad (32.7)$$

If we use the above numerical values of q at the basic frequencies and frequencies though they were written in mks units, we will get the right numerical results. For example, the older than 25 years is $P = 90^2 \pi^2 / c^2 \cdot 4 \text{ cm} \cdot 1 \text{ psec} = 6.67 \text{ erg/sec}$ at a polder and at the most distance $r = d^2/4\pi c^2$ or d^2/λ , with $d = 1.3183 \times 10^{-14}$ mks.

32.3 Radiation damping

Now, let's first consider a system that loses a certain energy without reason. If we put a charge on the end of a spring (one electron in an atom which has $\omega_0^2 \gg 1$) imagine ω_0 , and we start it oscillating and let's say it will do oscillate forever, even if it is in empty space nothing comes from anything. There is no oil, no resistance, in an empty space we "vibrate". But, nevertheless it will not oscillate, as we might once have said. However, however if it is charged it is radiating energy, and therefore the oscillation will slowly stop on. How about? What is the ζ of such an oscillator caused by the electrodynamic effects, the so-called radiation resistance or radiation damping of the system? The ζ of any oscillating system is the least energy content of the oscillation at any time divided by the energy loss per radian:

$$\zeta = \frac{\dot{E}}{dE/dt}.$$

Or another way to write it, since $dE/dt = -dE/dt$, $(dE/dt)/(\omega_0/\Delta\omega) = -dE/dt/\Delta\omega/2$.

$$R = \frac{dE}{d\omega_0/\Delta\omega} \quad (12.8)$$

For a given dE/dt it tells us how the energy of the oscillator varies with $\omega_0/\Delta\omega = -\langle E \rangle / Q^2 \omega_0^2$, which are the variables $R^2 = -dE/d\omega_0/\Delta\omega$. If E_0 is the initial energy at $t = 0$,

To find the $\langle E \rangle$ for a radiator, we go back to (12.8) and use (12.6) too and find

Now what do we use for the energy E_0 of the oscillator? The kinetic energy of the oscillator is $\frac{1}{2}mv^2$, and the zero kinetic energy is $m\omega_0^2 r_0^2/4$. But we remember that for the total energy of an oscillator, we have one-half its kinetic and half its potential energy, and so we divide our result, and find for the total energy of the oscillator

$$E = \frac{1}{2}mv^2 \omega_0^2 \quad (12.9)$$

What do we use for the frequency in our formulae? We use the natural frequency ω_0 , because for all practical purposes, that is the frequency at which our atom is radiating, and for v we use the electron mass m_e . Then, making the necessary corrections and cancellations, the formula comes down to

$$\frac{1}{Q} = \frac{4\pi v^2}{\lambda c m_e^2} \quad (12.10)$$

In order to see it better and in a more justified form, we write it using in place of v^2/m_e , c^2 , and the factor $4\pi^2/3$, which has been written as $2\pi^2/3$, since Q is dimensionless. The combination c^2/m_e^2 would be a quantity only of the electron charge and mass, an intrinsic property of the electron, and it must be a length. It has been given a name, the *characteristic radius*, because its early atomic students at first were inclined to equate the radiation resistance to the size of the atom. One part of the electron is acting, on the other parts, it needs to have an electron whose dimensions are of this general order of magnitude. However, this quantity no longer has the significance that we had, but the electron really has size 4 radius. Numerically, the magnitude of this radius is

$$r_0 = \frac{c^2}{m_e^2} = 3.82 \times 10^{-16} \text{ m} \quad (12.11)$$

Now let us actually calculate. In Q of an atom that is emitting light, λ , say 6500 angstroms. For a sodium atom, the wave length is roughly 6000 angstroms, in the yellow part of the visible spectrum, and has a typical wavelength. This

$$Q = \frac{c^2}{\lambda m_e^2} \approx 5 \times 10^9 \quad (12.12)$$

so the Q of an atom is of the order 10^9 . The mean total resistance of an oscillator will oscillate for 10^9 radiations about 10^9 oscillations, before its energy falls by a factor $1/e$. The frequency of oscillation of light corresponding to 6000 angstroms, $\nu = c/\lambda \approx$ on the order of 10^{14} cycles/sec, and therefore the lifetime, the time it takes for the energy of a radiating atom to fall off by a factor $1/e$, is on the order of 2×10^{-10} sec. In ordinary circumstances, radio emitting sources usually take about 10¹⁰ sec to radiate. This is only for atoms which are in empty space, not being disturbed in any way. If the electron is in a solid and it has to go through some of other electrons, then there are additional resistances and different damping.

The effective resistance term, γ , in the resistor $R = R_0 + \gamma$, the resistance can be found from the relation $1/Q = \gamma \omega_0^2$, and we remember that the size of γ determines how wide the resonance curve is (Fig. 11-2). Thus we have just computed the width of spectral lines for freely radiating atoms. Since $R = R_0 + \gamma \omega_0^2$, we find

$$\begin{aligned} \gamma \omega_0 &= 2\pi \cdot 6000^2 \cdot 10^9 \cdot 4\pi^2/3 = 2\pi c^2/3\omega_0 \\ \gamma \cdot 2\pi \cdot 6000/3 &= 1.18 \times 10^{-11} \text{ m} \end{aligned} \quad (12.13)$$

9.1 Interference sources

In preparation for our second topic, the scattering of light, we must now discuss a certain feature of the phenomenon of interference that we neglected or discussed previously. This is the question of when interfering does not occur. If we have two sources S_1 and S_2 , with amplitudes A_1 and A_2 , and we make an observation in a direction \hat{r} (in which the phases ϕ_1 and ϕ_2 of the two signals are ϕ_1 and ϕ_2 (a combination of the actual time of arrival from and the delay time, depending on the position of observation), then the energy that we receive can be found by summing over the two complex numbers $A_1 e^{i\phi_1}$ and $A_2 e^{i\phi_2}$ (as in Chapter 6) and we find that the resultant energy is proportional to:

$$A_R^2 = A_1^2 + A_2^2 + 2A_1 A_2 \cos(\phi_1 - \phi_2). \quad (9.1.1)$$

Now if the cross term $2A_1 A_2 \cos(\phi_1 - \phi_2)$ were not there, then the total energy that would be emitted in a given direction would simply be the sum of the energies, $A_1^2 + A_2^2$, and it would be like adding two waves separately, which is what we usually expect—that is, the combined intensity of light coming to somewhere from two sources is the sum of the intensities of the two lights. On the other hand, if we have things just right and we have some form of interference, because there is also some coherence, then there are circumstances in which this term is of no importance; that is, we could say the interference is approximately lost. Of course, in nature it is always there, but we may not be able to detect it.

Let us consider some examples. Suppose first that the two sources are $7000 \text{ km}/\text{sec}$ apart, as in an interplanetary experiment. Then in a given direction it is true that there is a very definite value of cross source elements. And, on the other hand, if we are just using, i.e., one dimension of wave-lengths, which is no distance at all (or a very, almost, but not quite, negligible one), we are averaging the effects over a range very wide compared with one wavelength, from one edge of the celestial plane and the other, and so very rapidly. If we take the average of the intensity over a long enough time interval, which goes plus minus, plus, minus, as we move around, averages to zero.

So if we average over a long time, when the phases change very rapidly with position, we get no interference.

Another example. Suppose that the two sources are two independent radio transmitters—no single one star being fed by two wires, which you know that the phases are kept together—but two independent sources—and that they are, for example, constant at the same frequency f . It is very hard to make them at exactly the same frequency without slightly varying them together. In this case we have what we call two independent sources. Of course, since the frequencies are not exactly equal, although they started in phase, one of them begins to get a little ahead of the other, and pretty soon they are out of phase, and then it goes off further ahead, and so the two start to be in phase again. So the phase difference between the two is gradually changing with time. But if one observes for a long time, we cannot see that little time. If we average over a much longer time, even with c the velocity very well and fully like δt at the 10^{-12} "heads" in seconds, if these settings and turnings are too rapid for our equipment, before changing this way and coming

back to the first one, the phases shall average out, we get no interference.

One has many books which say that two random light sources never interfere. This is not a statement of physics, but is merely a statement of the degree of consistency of the technique of the experiments at the time the book was written. What happens is a light source is radiating one atom radius, then another atom radius, and so forth, and we have just seen that atoms radiate a train of waves only for about 10^{-12} sec; after 10^{-12} sec, some atoms have probably taken over, then another atom takes over, and so on. So the phases can really only stay the same for about 10^{-12} sec. Therefore, if we average for very much more than 10^{-12} sec, we do not see any interference from two different sources, because they cannot link their phases steady for longer than 10^{-12} sec. With phototubes, very high-speed ones, in

is possible, and one can show that there is an interference which varies with time, up and down, in about 10^{-8} sec. But most detection equipment, of course, does not look at such fine time intervals, and thus was no interference. Only with the eye, which has a length of a second, averaging time, there is no chance whether or not seeing an interference, because two coherent waves may cancel.

Recently it has become possible to make light sources which get around this difficulty by making all the photons fit together to travel. The laser which does this is very complicated thing, and has to be understood in a quantum-mechanical way. It is called coherent, and it is possible to produce from a laser a source in which the interference frequency, the instant which the phase is kept constant, is very much larger, say 10^{12} sec. This is sort of the order of a hundredth, a tenth, or even one second, and so, with ordinary binoculars one can pick up the frequency between two different lasers. One can easily detect the pulsing of the lasers between two sources. Some day, right, someone will have to demonstrate an audience through a wall, in which the laser can show that there is some change. In you are asked.

Another case in which the interference averages out is that in which, instead of having only two sources, we have many. In this case, we would write the expression for A_0 as the sum of a whole lot of amplitudes, complex numbers, a_1 to a_n , and we would get the square of each one, all added together, if there were no interaction between each pair, and if the a_i 's were not real, that the total average out, there were no interference of individual ones. It may be that the various sources are arranged in such fashion that this may, although the phase difference between a_1 and a_2 , etc., is considerable, it is very different from that between a_1 and a_3 , etc. So we might get a whole lot of waves, many plus minus signs all averaging out.

So it is that in many cases, when we do not see the effects of interference, this is only a collective effect in every equal to the sum of all the intensities.

22-5 Scattering of light

The above leads us to the fact which occurs in all cases as a consequence of the interaction of photons with charges. When we were discussing the index of refraction, we saw that an incoming beam of light will make the atoms vibrate, the electric field of the incoming beam drives the electrons up and down, and they radiate because of their motion, too. This scattered radiation contributes a wave in the same direction as the incoming beam, but at somewhat different phases, and this is the origin of the index of refraction.

But what can we say about the interaction of scattered light in some other direction? Obviously, if the atoms are very haphazardly located in a rock for example, it's easy to find that we get nothing in a long distance. However, if we are adding a lot of waves, and their phases always changing, and the result comes to zero. If the object is, however, located, then the total intensity in any direction is the sum of the intensities that are scattered by each atom, as we have just discussed. You remember, the atoms in a gas are in actual motion, so that although the relative phase of wavelet to a definite atom does the same, will be quite different and therefore the wave coming from will average out. Therefore, the final intensity in light is scattered in a given direction by a gas, we get that only the effects of one atom is multiplying 10²³ intensity, it comes by the number of atoms.

Earlier we remarked that the propagation of scattering of light of various wavelengths along the line of sight. The sun light goes through the air, and when we look to one side of the sun, say at 40° to the sun, we see blue light, and we now have to calculate whether such light we see are with it is also.

If the incident beam has the electric field $\mathbf{E} = E_0 e^{i\omega t}$ at the point where the atom is located, we know that an atom in the atom will vibrate up and down in response to this \mathbf{E} (Fig. 22-2). From Eq. (22.3) the amplitude will be

$$\mathbf{a} = \frac{e}{m(\epsilon_0 + m^2/\omega^2)} \mathbf{E} \quad (22.13)$$

We could include the damping, and the possibility for the wave to be scattered by different frequency and size over the various frequencies, but the simplicity let us just take one oscillation and neglect the damping. Then the response to this constant electric field, which we have already read in the calculation of the index of refraction, is simply

$$\mathbf{x} = \frac{q_0 E_0}{m\omega^2} \quad (32.16)$$

We could now easily calculate the intensity of light that is emitted if we know ω , because, using formula (32.5) and the answer to our question in the diagram, it is given by

For the time being this however, we shall simply calculate the total amount of energy scattered in all directions just to save time. The total amount of light energy scattered, scattered in all directions by the single atom, is of course given by Eq. (32.7). So, putting together the various pieces and regrouping them, we get

$$\begin{aligned} P &= (\epsilon_0^2 n^4 / 16 \pi c^2) q_0^2 E_0^2 / (m^2 \omega^4 - \lambda_0^2) \\ &= 1.602 \times 10^{-24} \text{ ergs} / (16 \pi^2 n^4 m^2) \lambda_0^4 / (\omega^4 - \lambda_0^2) \\ &\propto \epsilon_0^2 n^4 \lambda_0^4 / (\omega^4 - \lambda_0^2) \end{aligned} \quad (32.17)$$

In other words, power is emitted in all directions.

We have written the result in the above form because it is then easy to remember: First, the total energy that is scattered is proportional to the square of the incident wave. What does that mean? Obviously, the source of the incident field is zero orthogonal to the energy which is being transferred. In fact, the energy incident on ω , at a certain per cent, is sent into the energy (ω^2) of the scattered field, and if Δ is the maximum value of Δ , then $\omega^2 = \frac{1}{2}\Delta^2$. In other words, the total energy scattered is proportional to the energy per second into that angle i.e., the "light" density of the scattering is the very bright or the very bright to low.

Next, what fraction of the incoming light is scattered? Let us imagine "the angle" with a certain area, let us say a , in the beam to be the material target because this would catch light, and so on; we need an imaginary area a in the space. The total amount of energy that would pass through this surface in a given dimension is proportional, both to the incoming intensity and to a , and it would be

$$P = I_0 A \epsilon_0 N / 4\pi \quad (32.18)$$

Now we iterate on this, we say that the amount of the total amount of intensity which is the amount which would fall on a certain geometrical area, and we give the answer by giving two rates. That answer, then, is independent of the incident intensity, it gives the "rate of intensity scattered" in the same incident per second per meter. In other words, the rate

$$\frac{\text{total energy scattered per second}}{\text{energy incident per square meter per second}} \text{ is an area}$$

The significance of this area is that, (1) the area product is equal to that area we were to be quoted in cm^2 sometime. (2) that is the amount of energy that would have been by the area.

This area is called a cross section for scattering. The idea of a cross section is well established, when you write a minimum cross section proportional to the intensity of the beam. In such cases one always describes the amount of the absorption by saying what the effective area would have to be to pick up the amount of the beam. If there were no mass in any way that this resistance actually has such an area. If there were no mass present, but a mass which is shaking up and down, there would be no mass density associated with it, physically. It is merely a way of expressing the amount to a certain kind of problem; it tells us what area the equivalent beam would

have no air to scatter to account for the much energy scattering at. Thus, for our case,

$$\tau = \frac{8\pi^4}{3} \cdot \frac{\alpha^4}{(\omega^2 - \omega_0^2)^3} \quad (2.19)$$

The subscript is to "scattering".

Let's look at some numbers. First, if we go to a very low natural frequency ω_0 , or to completely unbound electrons, for which $\omega_0 = 0$, then the frequency is constant, and the wave function is a constant. This low-frequency term, or the first electron current density, is known as the Rayleigh scattering cross section. It has area whose dimensions are approximately 10^{-13} meter, times or less, on a side, i.e., 10^{-31} square meter, which is quite small.

On the other hand, if we take the case of light to be ω , we remember that for the excited frequencies of the oscillators are higher than the frequency of the light that we use. This means, for a fine approximation, we can disregard ω^2 in the denominator, etc. we find that the scattering is proportional to the fourth power of the frequency. That is to say, light which is of higher frequency by, e.g., a factor of $\omega_0 > 10^{10}$ times, more intensely scattered, which is a considerable difference. This means that the light which fluctuates twice the frequency of the incident end of the spectrum, is scattered as a far greater intensity than light which we look at the sky in only that glorious time that we see all the stars!

I have got several points to be stated about the atmosphere. One interesting question is why do we ever see the clouds? When in the clouds some friend Franklykilly knows it is the condensation of water vapor. But, of course, the water vapor is always in the atmosphere, however, condenses on why don't we see it, i.e.,? After it condenses it is perfectly obvious, it won't there, now it is clear that the mystery of where the clouds come from is not really such a difficult mystery as "Where does the water come from?" but has to be resolved.

We have just explained that every atom scatters light, and of course the water vapor will scatter light too. The trouble is, if you when the water is combined into clouds, does it scatter such a correspondingly greater amount of light?

Consider what would happen if, instead of a single atom, we had an aggregate of atoms, say two, very close together and spaced with the wavelength of the light. Recall that atoms is only an angstrom or so across, while the wavelength of light is, say, 500 angstroms, so when they form a dumbbell-like atoms together, they can be very close together compared with the wavelength of light. Then when the electric field is a sum of the two electric fields in phase, i.e., consider the amplitude that were ψ_1 with a single atom, and the energy which is proportional is the square of $|\psi_1|$ with one atom, it is with a single atom, not twice! So twice of atoms require no greater more energy than the single atom. Our argument, i.e., the pictures are independent is based on the assumption that there is a real and large difference in phase between any two atoms, which is true only if they are several wavelengths apart and randomly spaced, in moving. But, "they are right next to each other, they necessarily scatter in phase, and they have a coherent interference which gives us an increase in the scattering."

If we have N atoms in a line, which is a tiny droplet of water or other substance, it will be driven by the electric field in about the same way as before (the effect of one atom on the others is not important; it is just to get the idea anyway), and the amplitude of scattering from each one is the same, so the total field which is scattered is N -fold increased. The intensity of the light which is scattered is then the square, or N^2 -fold, increased. We would have imagined, if the atoms were spread out in space, only N times as much as intensity we get N^2 times as much as 1. That is to say, the scattering of water in terms of N molecules is not $\sim N$ times more intense than the scattering of the single atoms. Since however approaching the scattering increases, there is increase not different. Now, when does the analogy begin to fail? How many atoms are we put together before we cannot ignore the aggregate and consider it a dumbbell? When the water droplets are big, but still the size of the wavelength or so, then the atoms are no longer close

place because they will scatter light. So as we keep increasing the size of the droplets we get more and more scattering until we have about $10\text{ }\mu$ in diameter. At this size of a wave length, and then the scattering does not increase anywhere nearly as rapidly as the unperturbed light. Furthermore, the blue disappears, because the long wavelength the charge can be higher before this limit is reached, then they can be for short wavelengths. Although the short waves scatter more per atom than the long waves, there is a larger enhancement for the red end of the spectrum than for the blue end where all the charges are bigger than the wavelength, so the color is shifted from the blue toward the red.

Now we can make an experiment that demonstrates this. We can make particles that are very small at first, and then grow, if you like. We used a solution of sodium thiosulfate that was sulphuric acid, which precipitates very finely divided sulphur. As the sulphur precipitates the grains get continually smaller, and the scattering is a little clumpy. As it precipitates more it gets more clusters, and then will get sulphur crystals yet bigger. In addition, the light which goes straight through will have the blue taken out. That is why the sunsets are red, of course, because the light that comes through a lot of air to our eye has had a lot of blue taken out, and so is yellow-red.

Finally, there is one other important fact which really belongs in the next chapter, on polarization, but it is so interesting that we put it in now. This is the photoelectric effect of the scattered light which is a particle interaction. The electric field in the incoming light is oscillating in some way, and the driven oscillations of the atoms in the same direction, and if the atoms are scattered occur at right angles to the beam, we will see polarized light, that is to say, light in which the electric field is going only one way. In general, the atoms can vibrate in any direction at right angles to the beam, but if they are driven coherently toward or away from us, we do not see this. So if the incoming light has an electric field which changes and oscillates in any direction, which we call unpolarized light, then the light which is coming out at 90° to the beam will also be only one direction. (See Fig. 32-3.)

There is a substance called galena which has the property that when light goes through it, only one piece of the electric field which is along one particular axis can get through. We can use this to test for polarization, and indeed we find the light scattered by the hydrosolids to be a singly polarized.

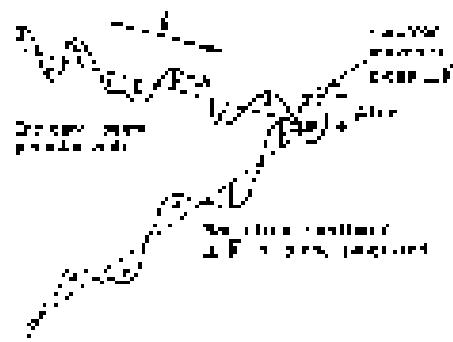


Fig. 32-3. Illustration of the origin of the polarization of scattered scattered straight engine to the medium beam.

Polarization

33-1 The electric vector of light

In this chapter we shall consider those phenomena which depend on the fact that the electric field E describes the light as a vector. In previous chapters we have not been concerned with the direction of oscillation of the electric field except to note that the electric vector E is always perpendicular to the direction of propagation. The particular direction in the plane does not concern us. We now consider those directions whose control factor ϵ_{xy} is the particular direction of oscillation of the electric field.

In ordinary monochromatic light, the electric field E oscillates at a definite frequency, but since the x-component and the y-component are not independently at a definite frequency, we must first consider the resultant effect produced by superposing two independent oscillations at right angles to each other. What kind of effect is this? It's much like that of a longitudinal and a transverse wave which oscillate in the same frequency. If one tends to an vibration along the x -axis and the y -vibration in the same phase, the result is a vibration in a new direction in the system. Figure 33-1 illustrates the superposition of different amplitudes for the x-vibration and the y-vibration. But the results shown in Fig. 33-1 are not the only possibilities; in all of these cases we have assumed that the x-vibration and the y-vibration are in phase, but this need not be the way. It could be that the x-vibration and the y-vibration are out of phase.

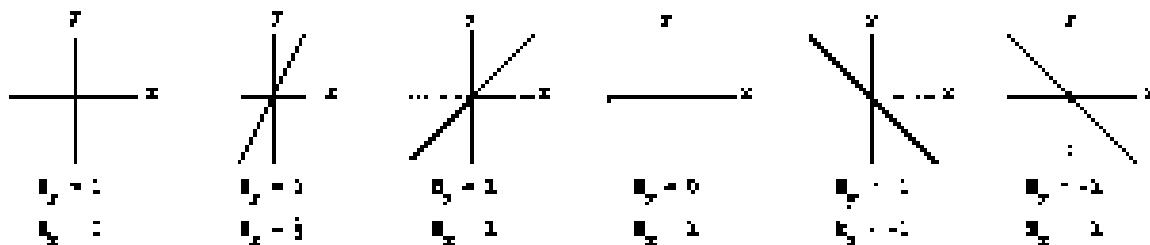


Fig. 33-1. Superposition of x-vibration and y-vibration in phase.

When the x-vibration and the y-vibration are out of phase, the electric field vector moves around in an ellipse, and we can illustrate this in a familiar way. If we hang a ball from a support by a long string, so that it can swing freely in a horizontal plane, it will execute "circular" oscillations. If we imagine horizontally x - and y -coordinates with their origin at the rest position of the ball, the ball will swing in either the x or y direction with the same fundamental frequency. By selecting the proper initial displacement and initial velocity v_0 , we can let the ball oscillate in straight paths, back and forth, or along any straight line in the xy -plane. These motions of the ball are analogous to the oscillations of the electric field shown illustrated in Fig. 33-1. In such line cases, since the x-vibration and the y-vibration both have constant frequency at the same time, the x and y oscillations are in phase. But we know that the actual motion of the ball is motion in an ellipse, which corresponds to oscillations in which the x - and y -oscillations are out of phase. The superposition of x and y vibrations which are not in phase is illustrated in Fig. 33-2. On a slightly elliptical orbit in the plane of the vibration are plotted two vibrations. The path of the ball is that the electric vector moves around in an ellipse. The motion is a straight line in a particular

33-1 The electric vector of light

33-2 Polarization of scattered light

33-3 Birefringence

33-4 Polarizers

33-5 Optical activity

33-6 The Intensity of reflected light

34-1 Anomalous refraction

corresponding to a phase difference of zero (or an integral multiple of π); modulated electric waveforms have amplitudes with a phase difference of $\pi/2$ (or any odd integral multiple of $\pi/2$).

In Fig. 33-2 we have labeled the electric field vectors in the x - and y -directions with complex numbers, which is a common representation at higher frequencies. We project the real and imaginary components of the complex electric vector to this notation with the x - and y -components of the field. The x - and y -components plotted in Fig. 33-1 and Fig. 33-3 are actual electric fields that we can measure. The real and imaginary components of a complex electric field vector are only mathematical conveniences and have no physical significance.

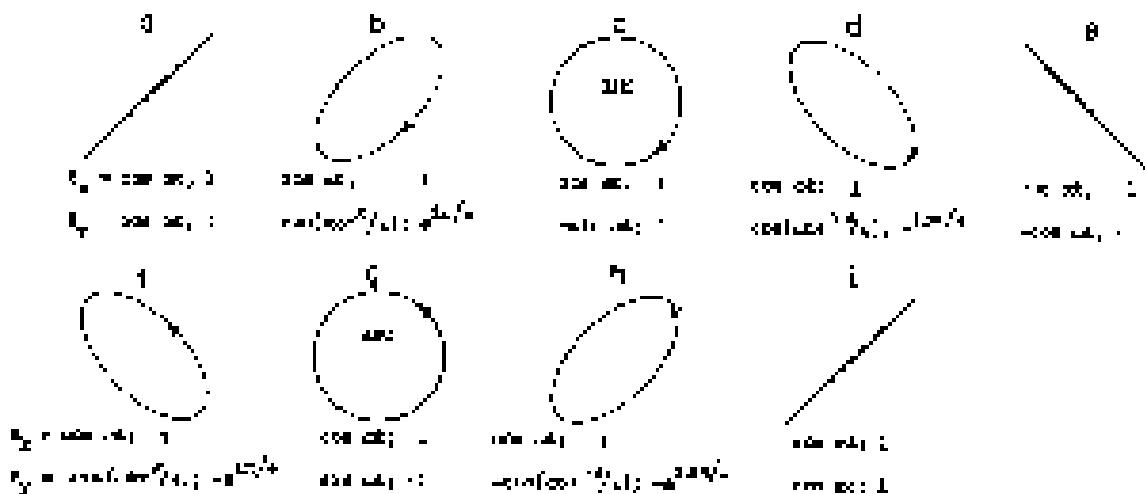


Fig. 33-2. Superposition of x and y vibrations with equal amplitude but vertical phase. The components E_x and E_y are expressed in both real and complex notation.

Now for some terminology. Light is linearly polarized (sometimes called plane polarized) when the electric field oscillates on a straight line. Fig. 33-1 illustrates linear polarization. When the end of the electric field vector revolves in an ellipse, the light is elliptically polarized. When the end of the electric field vector rotates around a circle the light is circular polarized. If the end of the vector is nearer when we look at it to the right, comes straight toward us goes around to the counterclockwise direction, we call it right-hand circular polarization. Figure 33-3(a) illustrates right-hand circular polarization, and Fig. 33-3(b) shows left-hand circular polarization. In both cases the light is coming out of the page. Our convention for labeling left-hand and right-hand circular polarizations is consistent with that which is used today for the outer electrons in atoms which exhibit polarization (e.g., electrons). However, in some books one finds the opposite convention is used, so one must be careful.

We have considered linearly, circularly, and elliptically polarized light which covers everything except for the case of unpolarized light. Now how can we light be unpolarized when we know that it must exhibit in one or another of these effects? If the light is not absolutely monochromatic, i.e., if the x - and y -phases are not kept perfectly together, or if the electric vector just vibrates in one direction, the polarization is constantly changing. Remember that one atom emits during 10^{-9} sec, and if a beam emits over air polarization, and then another atom emits light with a different polarization, the polarizations will change every 10^{-9} sec. If the polarization changes more rapidly than we can detect it, then we call the light unpolarized, because all the effects of polarization find average out. None of the few simple effects of polarization would show up with unpolarized light. But if we look from the definition, light is unpolarized only if we are unable to find out whether the light is polarized or not.

20-2 Polarization of scattered light

The first example of the polarization effect that we have already discussed is the scattering of light. Consider a beam of light, for example from the sun, passing through air. Electric field will produce oscillations of charges in the air, and motion of these charges will radiate light with its maximum intensity in a plane normal to the direction of vibration of the charges. The beam from the sun is unpolarized, so the direction of polarization changes randomly. If we consider light scattered at 90°, i.e. vibration of the charged particles relates to the observer only when the vibration is perpendicular to the characteristic line of sight, and then light will be polarized along the direction of vibration. So scattering is an example of one more of the many polarization effects.

20-3 Birefringence

Another interesting effect of polarization is the fact that there are substances for which the index of refraction is different for light linearly polarized in one direction and nearly polarized in another. Suppose that we had some medium which consisted of long, nonspinning molecules, longer than they are wide, and suppose that these molecules were arranged in the substance with their long axes parallel. Then what happens when the oscillating electric field passes through this substance? Suppose that because of the structure of the molecule, the vibration in the substance requires more energy to oscillate in the direction parallel to the axes of the molecules than they would require if the electric field tried to push them straight along the molecular axis. In this way we expect a different response for polarization in one direction from the polarization in other angles to that direction. Let us call the direction of the axes of the molecules the optic axis. When the polarization is in the direction of the optic axis the index of refraction is different than it would be if the direction of polarization were at right angles to it. Such a substance is called birefringent. It has two refractive indices, i.e. two indices of refraction depending on the direction of the polarization inside the substance. This kind of a substance can be calcite. In a birefringent substance there must be a certain amount of anisotropy, for one reason or another, in the medium itself. Our usual cubic crystal, which has the symmetry of a cube, cannot be birefringent. Our long needle-like crystals and rod-like molecules that are very metric, and one observes this effect very easily.

Let us see what effects we would expect if we were to shine polarized light through a plate of a birefringent substance. If the polarized wave is parallel to the optic axis the light will go through with one velocity; if the polarization is perpendicular to the axis, the light is transmitted with a different velocity. An interesting situation arises when, say, light is linearly polarized at 45° to the optic axis. Now the 45° polarization, we have already noticed, can be regarded as a superposition of the x- and the y-polarizations of equal amplitudes and in phase, as shown in Fig. 20-2(a). Since the x- and y-polarizations travel with different velocities, their phases change at a different rate as the light passes through the substance, but although the x- and y-polarizations are in phase, inside the material, the phase difference between x- and y-polarizations is proportional to the length of the substance. As the light proceeds through the material the polarization changes as shown in the series of diagrams in Fig. 20-2. If the thickness of the plate is just right to introduce $\pi/2$ phase shift between the x- and y-polarizations, as in Fig. 20-2(c), the light will come out elliptically polarized. Such a thickness is called a quarter-wavelength, because it introduces a quarter-cycle phase difference between the x- and the y-polarizations. If linearly polarized light is sent through two such thin plates, i.e. with some net plane polarization again, but at right angles to the original direction, as we can see from Fig. 20-2(d).

One can easily illustrate this phenomena with a piece of cellophane. Cellophane is made of long, fibrous molecules, and is not isotropic, since the fibers lie preferentially in a certain direction. To demonstrate this property we take a

is a linearly polarized light, and we can obtain it by sending unpolarized light through a sheet of polaroid. Polycryl, which we will discuss later in more detail, has the useful property that it transmits light that is linearly polarized parallel to the axis of the polaroid very well. Consequently, the light polarized in a direction perpendicular to the axis of the polaroid is strongly absorbed. When we pass unpolarized light through a sheet of polycryl, only one part of it emerges unchanged which is either (if parallel to the axis) or the polaroid goes through so that the transverse beam is identity polarized. This same property of polaroid is also useful in detecting the direction of polarization of a linearly polarized beam, or in determining whether a beam is linearly polarized or not. One simply passes the beam of light through the polaroid sheet and rotates the polaroid in the plane defined by the beam. If the light is not linearly polarized, it will not be transmitted through the sheet when the polaroid is turned in the direction of polarization. The transmission varies only slightly throughout when the axis of the polaroid sheet is rotated 1 through 90°. If the transmitted intensity is independent of the orientation of the system, the beam is not linearly polarized.

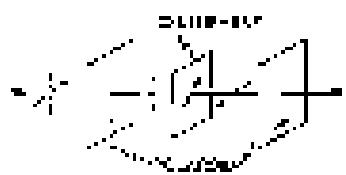


Fig. 30-3. An experimental observation of the birefringence of cellulose. The electric vectors in the light are indicated by the dotted lines. The axes of the polaroid sheets and their axes of the cellulose are indicated by arrows. The incident beam is unpolarized.

To demonstrate the birefringence of cellulose, we use two sheets of polaroid, as shown in Fig. 30-3. The first gives us a linearly polarized beam which we pass through the cellulose and then through the second polaroid sheet, which serves to detect any effect the cellulose may have had on the polarized light passing through it. If we first set the axes of the two polaroid sheets perpendicular to each other and rotate the cellulose sheet, we observe that in general the cellulose makes it possible for some light to pass through the second polaroid. However, there are two orientations of the cellulose sheet, at 45° angles to each other, which permit no light to pass through the second polaroid. These amounts given in which linearly polarized light is transmitted through the cellulose with no effect on the direction of polarization are those for the directions parallel and perpendicular to the optic axis of the cellulose sheet.

We suppose that the light passes through the cellulose with two different velocities to cause two different retardances, but this is not true without changing the directional retardation. When the cellulose is twisted halfway between these two orientations, as shown in Fig. 30-4, we see that the light transmitted through the second polaroid is bright.

It just happens that ordinary cellulose used in commercial packaging is very close to a half-wave thickness for most of the colors in white light. Such a sheet will turn the axis of linearly polarized light through 90° if the initial linear polarization is perpendicular to an angle of 45° with the optic axis, so that the beam emerging from the cellulose is propagating in the right direction to pass through the second polaroid sheet.

If we use white light in this form, rather than ordinary white, the cellulose sheet will be of the proper birefringent thickness only for a particular component of the white light, and the transmitted beam will have the color of this component. The color transmission depends on the thickness of the cellulose sheet, and we can vary the effective thickness of the cellulose by stretching it so that the light propagates through the cellulose at roughly one-and-a-half times longer path in the cellulose. As the sheet is stretched the transmissive color changes. With a thickness of 1.1 mm, the thickness can transmit blue light, still transmit yellow colors, the two polaroid sheets have their axes perpendicular, and the components of the light waves of the two polaroid sheets are parallel.

Another interesting application of aligned molecules is optical plastic. Certain plastics are composed of very long and complicated molecules which are joined together. When the plastic is crystallized, usually the molecules are all lined up in a more or less regular arrangement, many aligned in one direction or another, and so the plastic is not particularly transparent. Usually there are strains and stresses induced, and when the material is solidified, as the material is at first, there are

porous. However, if we apply tension to a piece of this plastic material, it is as if we were pulling a molecule of stress, and there will be more strings preferentially aligned parallel to the tension than in any other direction. So when a stress is applied to each plastic, they become birefringent, and one can see the effect of the birefringence by passing polarized light through the plastic. If we examine the transmitted light through a polaroid sheet, patterns of light and dark fringes will be observed (similar, if white light is used). The pattern of these fringes is applied to the sample, and by counting the fringes and seeing where most of them are, one can determine what the stress is. Engineers use this phenomenon as a way of finding the stresses in cylindrical pieces that are difficult to examine.

Another interesting example of a way of detecting birefringence is by means of a liquid substance. Consider a liquid composed of long, asymmetric molecules which carry a plus or minus charge near the ends of the molecule, so that the molecule is an electric dipole. In the collisions in the liquid the molecules will randomly bump into each other, with no net electric potential in one direction as a result. If we apply an electric field the molecules will tend to line up, and the moment they lose the dipole becomes extinguished. With two polaroid sheets and a trough containing such a polar liquid, we can devise an arrangement with the property that light is transmitted only when the electric field is applied. So we have an electric switch for light, which is called a Kerr cell. This effect, i.e., an electric field can produce birefringence in certain liquids, is called the Kerr effect.

13.4 Polarizers

So far we have considered substances in which the refractive index is different for light polarized in different directions. Of very great interest are the crystals and other substances in which not only the index, but also the coefficient of absorption, is different for light polarized in different directions. By the same arguments which I presented before of birefringence, it is understandable that a substance can vary with the direction in which the charges are forced to vibrate in an anisotropic substance. That's fine, so what example and example is another. Polaroid consists of a thin layer of small crystals of hemaphthite, or salt of iodine and quinone, all aligned with their axes parallel. These crystals absorb light when the oscillations are in one direction, and they do not absorb appreciably when the oscillations are in the other direction.

Suppose that we send light into a polaroid sheet polarized linearly at an angle θ to the passing direction. What intensity will come through? This incident light can be resolved into a component perpendicular to the pass direction which is proportional to $\sin \theta$, and a component along the pass direction which is proportional to $\cos \theta$. The amplitude which comes out of the polaroid is only the cosine of θ , since the sin θ component is absorbed. The amplitude which passes through the polaroid is smaller than the amplitude which entered, by a factor cos θ . The energy which passes through the polaroid, i.e., the intensity of the light is proportional to the square of cos θ . Use I_0 there, as the intensity transmitted when the light enters polarized at an angle θ to the pass direction. The absorbed intensity, of course, is $\sin^2 \theta$.

An interesting situation is presented by the following situation. We know that it is not possible to send a beam of light through two polaroid sheets with their axes crossed at right angles. But if we place a third polaroid sheet between the first two, with its axis tilted at 45° to the crossed axes, some light is transmitted. We know that polaroid reflects light, it does not do anything. Nevertheless, the addition of a third polaroid at the surface makes light to get through. The analysis of this phenomenon is left as an exercise for the student.

One of the most interesting examples of polarization is not in complicated crystals or difficult substances, but is one of the simplest and most familiar of situations: the reflection of light from a surface. Before it is lost, when light is reflected from a glass surface, it may be polarized, and the physical explanation of this is very simple. It was discovered empirically by Brewster that light reflected

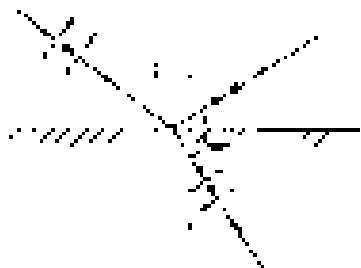


Fig. 33-4. Reduction of linearly polarized light at Brewster's angle. The polarization direction is indicated by dashed arrows; solid dots indicate polarization normal to the paper.

If a beam lies in a plane of polarization of the reflected beam, and the beam refracted into the mirror, form a right angle. The situation is illustrated in Fig. 33-4. If the incident beam is polarized in the plane of incidence, the beam will be reflected at all. Only if the incident beam is polarized normal to the plane of incidence will it be reflected. This reason is very easy to understand. If the reflecting material the light is polarized transversely, and we know that it is the motion of the charges in the material which generates the emergent beam, which we call the reflected beam. The source of this so-called reflected light is not simply that the incident beam is reflected; our deeper understanding of this phenomenon tells us that the incident beam drives an oscillation of the charges in the material, which in turn generates the reflected beam. From Fig. 33-4 it is clear that only the beam normal to the paper, i.e., radiating in the direction of reflection, and consequently the reflected beam will be polarized normal to the plane of incidence. If the incident beam is polarized in the plane of incidence, there will be no reflected light.

This phenomenon is readily demonstrated by reflecting a linearly polarized beam from a pane of glass. If the glass is turned to present different angles of incidence to the reflected beam, slight variation of the reflected intensity is observed when the angle of incidence passes through Brewster's angle. This variation is observed only if the plane of polarization lies in the plane of incidence. If the plane of polarization is normal to the plane of incidence, constant reflected intensity is observed at all angles.

33-5 Optical activity

Another most remarkable effect of polarization is observed in materials composed of molecules which do not have reflection symmetry: molecules shaped something like a dumbbell, or like a gloved hand, or any shape which, if viewed through a mirror, would be reversed in its sense (as, e.g., a left hand viewed inside as a right hand glove). Suppose all of the molecules in a substance are the same, i.e., none is a mirror image of any other. Such a substance may show an interesting effect called optical activity, whereby an initially polarized light passes through the substance, the direction of polarization rotates around the beam axis.

To inquire into the phenomenon of optical activity requires some calculation, but we can see qualitatively how it arises. Let's take a look, without actually carrying out the calculations. Consider an asymmetric molecule in the shape of a spiral, as shown in Fig. 33-5. Molecules need not actually be shaped like a corkscrew in order to exhibit optical activity, but this is a simple shape which we shall take as a typical example of those that do not have reflection symmetry. When a light beam linearly polarized along the y -direction falls on this molecule the electric field will drive charges up and down the helix, thereby generating a current in the helix; i.e., emitting an electric field, E , polarized in the y -direction. However, if the electrons are concentrated to move along the spiral, the spiral electrons will move in the x -direction to the x -axis, driving upward. When a current is flowing up the spiral, it is also flowing into the paper at $x = x_1$ and out of the paper at $x = x_2$. Let A be $\pi d^2/4$ the diameter of our molecular spiral. One might suppose that the current in the x -direction would produce no net field, since the currents are in opposite directions on opposite sides of the spiral. However, if we consider the contributions of the electric field to $x = x_1$ and $x = x_2$, we find that the field generated by the current at $x = x_1$ and $x = x_2$ are the field generated close to $x = x_1$ at time t and at $x = x_2$ separated in time by the distance d , and thus $\pi/2$ out of phase by $\pi/2$ radians. Since the phase difference is $\pi/2$, exactly $\pi/2$ between the fields due to the two ends, and because of the small x -component in the electric field generated by the motion of the electrons in the molecule, whereas the driving electric field has only a y -component. This small x -component, added to the large y -component, produces a resultant field that is tilted slightly with respect to the y -axis, the original direction of polarization. As the light moves through the molecule, the direction of polarization rotates about the beam axis. By carrying a few examples and calculating the correct effect will be set in motion by an accident.

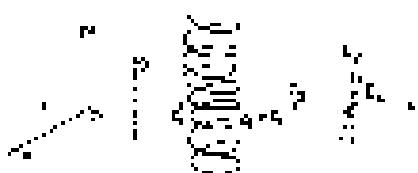


Fig. 33-5. A molecule with a shape that is not symmetric when reflected in a mirror. A beam of light, linearly polarized in the y -direction, falls on the molecule.

electric field, one can convince himself that the existence of optical activity and the sign of the rotation are independent of the orientation of the molecules.

Corn syrup is a common substance which possesses optical activity. The phenomenon is easily demonstrated with a polaroid sheet to produce a linearly polarized beam of transmission with an amplitude A , and a second polaroid used to detect the rotation of the direction of polarization as the light passes through a corn syrup.

31-6 The intensity of reflected light

Let us now consider quantitatively the reflection coefficient as a function of angle. Figure 31-6 illustrates a beam of light striking a glass surface, where it is partly reflected and partly refracted into the glass. Let us suppose that the incident wave has unit amplitude, is linearly polarized normal to the plane of the paper. Then we call the amplitude of the reflected wave b , and the amplitude of the refracted wave a . The refracted and reflected waves will, of course, be linearly polarized, and the electric field vectors of the incident, refracted, and reflected waves are all parallel to each other. Figure 31-6(a) shows the same situation but now we suppose that the incident wave, of unit amplitude, is polarized in the plane of the paper. Now let us call the amplitude of the reflected and refracted waves b and a , respectively.

We wish to calculate how strong the reflection is in the two situations illustrated in Fig. 31-6(a) and 31-6(b). We already know that when the angle θ is zero the reflected beam and refracted beam is a right angle. There will be no reflected wave in Fig. 31-6(a), but let us see if we cannot get a quantitative answer. Recall Eq. 30-1 for R which is a function of the angle of incidence, θ .

The principle that we must understand is as follows. The currents that are produced in the glass produce two waves. First, they produce the reflected wave. Moreover, we know that if there were no currents generated in the glass, the incident wave would continue straight in a straight line. Remember that all the sources in the world make the net field. The source of the incident light beam produces a field of unit amplitude which would move it to the glass along the dotted line in the figure. This field is not opposed, and therefore the currents generated in the glass must produce a field of amplitude -1 which moves along the dotted line. Using this fact, we will calculate the amplitudes of the reflected waves, b and a .

In Fig. 31-6(a) we see that the field of amplitude b is produced by the motion of the charges in the glass which are responding to a field coming in the glass, and that therefore b is proportional to v . We might suppose that since our two figures are exactly the same, except for the direction of polarization, the ratio b/a would be the same in both cases. This is not quite true, however, because in Fig. 31-6(a) the polarization directions are not all parallel to each other, as they are in Fig. 31-6(b). It is only the component of a which is perpendicular to B , A and i that is effective in producing B . The correct expression for the proportionality is then

$$\frac{b}{a} = \frac{\theta}{4 \pi n^2 (i + r)} \quad (31-1)$$

Now we can argue. We know that in both (a) and (b) of Fig. 31-6 the electric field in the glass must produce oscillations of the charges which generate a field of amplitude -1 , polarized parallel to the incident beam, and moving in the direction of the incident beam. But we see from part (b) of the figure that only one component of a that is normal to the original line has the right polarization to produce this field, whereas in Fig. 31-6(a) the amplitude is given in terms of the polarization of wave a to pass into the polarization of the wave of amplitude -1 . Therefore we can write

$$A \cos(i + r) = - \frac{a}{b} \quad (31-2)$$

since the two amplitudes on the left side of Eq. (31-2) each produce the wave of amplitude -1 .



Fig. 31-6 An incident wave of unit amplitude is reflected and refracted at a glass surface. In (a) the incident wave is linearly polarized normal to the plane of the paper. In (b) the incident wave is linearly polarized in the direction shown by the dotted electric vector.

Dividing Eq. (G1.1) by Eq. (G1.2) we obtain

$$\frac{S}{R} = \frac{\sin(\beta + \alpha)}{\sin(\beta - \alpha)}. \quad (G1.4)$$

a result which we can check against what we already know. If we take up Eq. (F2.2) it gives $S = 0$, as R never says it should be, so our results so far are at least not obviously wrong.

We have examined our amplitudes for the incident waves, so that $|S|^2/|R|^2$ is the reflection coefficient for waves polarized in the plane of incidence, and $|R|^2/|I|^2$ is the reflection coefficient for waves polarized normal to the plane of incidence.

The ratio of these two reflection coefficients is determined by Eq. (G1.4).

Now we can't just multiply our amplitude ratios together, but each oscillates $|S|^2/|R|^2$ and $|R|^2/|I|^2$ individually! We know from the conservation of energy that the energy in the reflected wave must be equal to the incident energy minus the energy in the reflected wave, $|I - R|^2$ in one case, $|S - R|^2$ in the other. Furthermore, the energy which passes into the glass in Fig. 32.6(b) is in the energy which goes into the glass in Fig. 32.6(a) as the sum of the square of the incident amplitudes, $|I|^2/|a|^2$. One might ask whether we really know how to compute the energy inside the glass, because, after all, there are energies of motion of the atoms in addition to the energy in the electric field. But $|I|^2$ is proportional to the various contributions to the total energy will be proportional to the square of the amplitude of the electric field. Therefore we can write

$$1 - |R|^2 = \frac{|I|^2}{|a|^2} = \frac{|S|^2}{|R|^2}. \quad (G1.5)$$

We now substitute Eq. (G1.5) into equation (G1.4) from the expression above, and express δ in terms of θ by means of Eq. (G2.3).

$$\frac{1 - \frac{|S|^2}{|R|^2}(1 + r)}{1 - |R|^2} = \frac{1}{\cos^2(\beta - \alpha)}. \quad (G1.6)$$

This equation contains the only one unknown amplitude, δ . Solving for δ^2 , we obtain

$$\delta^2 = \frac{\sin^2(\beta - \alpha)}{\sin^2(\beta + \alpha)} \quad (G1.7)$$

and, with the aid of (G1.1)

$$|R|^2 = \frac{\sin^2(\beta - \alpha)}{\sin^2(\beta + \alpha)}. \quad (G1.8)$$

so we have found the reflection coefficient $|R|^2$ for an incident wave polarized perpendicular to the plane of incidence, and so the reflection coefficient $|S|^2$ for an incident wave polarized in the plane of incidence!

It is possible to go on with arguments of this nature and deduce that R is real. To prove this, one must consider a case where light is coming from both sides of the glass at once in the same time, a situation not easy to arrange experimentally, but fun to analyze theoretically. I've never tried it in general case, so I can prove that R must be real, and therefore, in fact, that $\delta = \pm \sin(\beta - \alpha)/\sin(\beta + \alpha)$. It is even possible to determine the sign by considering the case of a very, very thin layer in which there is reflection from the front face from the back surface, and taking along how much light is reflected. The *sign* how much light should be reflected by that layer, because we know how much intensity is present, and we have been working out the fields produced by such structures.

One can show by these arguments that

$$R = \frac{\sin(\beta - \alpha)}{\sin(\beta + \alpha)}, \quad R = -\frac{\tan(\beta - \alpha)}{\tan(\beta + \alpha)}. \quad (G1.9)$$

These expressions for the reflection coefficients as a function of the angles of incidence and reflection are called Fresnel's reflection formulae.

If we consider the limit as β approaches $\pi/2$ and α goes to zero, we have, for the case of normal incidence, that $|S|^2 = |R|^2 = 1 - r^2/2(1 + r)^2$ for both polarizations.

since the angles are practically equal to the angles θ and ϕ in the diagram. But we know that $\sin \phi = \sin \theta$ when the angle of refraction $\phi = \theta = \pi$. It is thus clear that the coefficient of reflection for normal incidence is

$$R^2 = S^2 = \frac{1}{n} = \frac{D^2}{D^2 + D_0^2}.$$

It is interesting to find out how much light is reflected or normal incidence from the surface of water, for example. For water, $n = 1.33$, so that the reflection coefficient is $(1/1.33)^2 = 7\%$. A normal incidence only two percent of the light is reflected from the surface of water.

33-7 Anomalous refraction

The last polarization effect we shall consider was actually one of the first to be discovered: anomalous refraction. When viewing Iceland Spar back to the open sky (Fig. 33-7) which had the surprising property of not allowing anything over through the crystal appear divided, i.e., as two images. This came to the attention of Hippocrates, and played a important role in the discovery of polarization. As before, we note, the polarizations which are discovered first are the harder, if hardly, to analyze. It is only after we understand a physical concept long enough that we can carefully select those parameters which most easily and simply determine that concept.

Anomalous refraction is a particular case of the same kind of effect we considered earlier. Anomalous refraction occurs when the ray is optic, i.e., long axis of an optically uniaxial, or nonparallel to the surface of the crystal. In Fig. 33-7 are shown two pieces of birefringent material, with the optic axis as shown. In the upper figure, a incident beam falling on the material is refracted parallel to a direction perpendicular to the optic axis of the material. When this beam strikes the surface of the material, each point on the surface becomes source of a wave which travels along the crystal with velocity v_1 , the velocity of light in the crystal when the plane of polarization is normal to the optic axis. The wave front is not the envelope of dots of waves little spirals wavy, and the wave front moves straight through the crystal and out the other side. This is just the ordinary behavior we would expect, and the ray is called the ordinary ray.

In the lower figure, a linearly polarized light falling on the crystal has its direction of polarization passed through 90° , and at the surface lies in the plane of polarization. When we now consider the wave propagation at any point on the surface of the crystal, we see that they do not spread out in spherical waves but are travelling along the surface with velocity v_1 because the polarization is perpendicular to the optic axis, whereas the light travelling perpendicular to the optic axis travels with velocity v_2 because the polarization is parallel to the optic axis. This requires $v_2 < v_1$, and in the figure $v_2 < v_1$. A more complete analysis will show that the waves spread out on the surface of an ellipsoid, with foci on the major axis of the ellipsoid. The envelope of all the elliptical waves is the wavefront which spreads out roughly in a parabola in the direction shown. Again, at the back surface the beam will be detected just as it was at the front surface, as that the optic axis is parallel to the incident beam, but separated from it. Clearly, the beam does not scatter greatly here, but goes in the extraordinary direction. It is therefore called the extraordinary ray.

When an unpolarized beam strikes an uniaxially birefringent crystal, it is separated into an ordinary ray, which travels every through the material with the same velocity which is displaced as it passes through the crystal. These two emergent rays are linearly polarized at right angles to each other. This is true for linear polarization with a state of polarization to analyze the polarization of the emergent rays. We can also demonstrate by our interpretation of this phenomenon is correct by sending linearly polarized light into the crystal. By properly orienting the direction of polarization of the incident beam, we can make the light pass right through the crystal, or we can make it go through without splitting our wave a displacement.

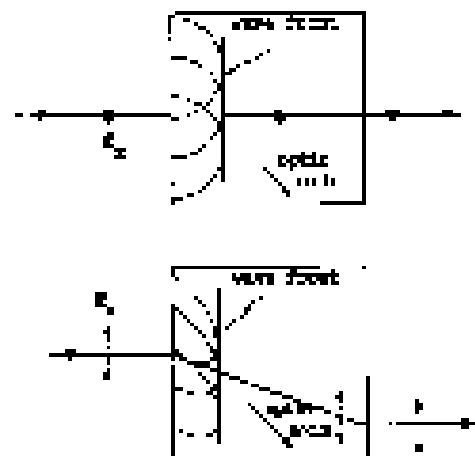


Fig. 33-7. Top diagram shows the path of the ordinary ray through a uniaxially birefringent crystal. The extraordinary ray is shown in the lower diagram. The optic axis lies in the plane of the paper.

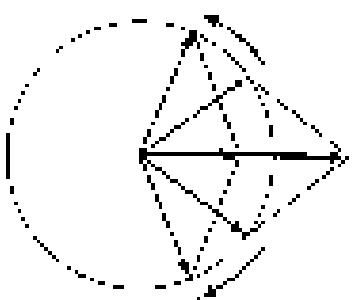


Fig. 33-5. Two oppositely rotating vectors of equal amplitude add to produce a vector in a fixed direction, but with an oscillating amplitude.

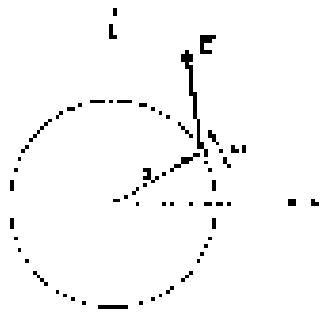


Fig. 33-6. A charge moving in a circle in response to circularly polarized light.

We have represented all the various extinction cases in Figs. 33-1 and 33-2 as superpositions of two fixed polarization states, namely x and y , in various amounts and phases. Other polarizations may we have been used. Polarization along any plane perpendicular to x and y is known to x and y as well as well [2]. For example, any polarization can be made up of superpositions of cases (a) or (c) of Fig. 33-2. It is interesting, however, that this can also occur for other cases [3]. For one spike, any linear polarization can be made up by superposing two while allowing suitable phases of light and left circular polarization (case (d)) are [3] (cf. Fig. 33-2'), since two equal waves collinear in opposite directions tend to give a single vector oscillating in a straight line (Fig. 33-3). If the phase of one is shifted relative to the other, the "line is helical." Thus all the pictures of Fig. 33-2 could be labeled "the superposition of equal amounts of right and left circular polarized light in various relative phases." As the left spike behind the right in phase, the direction of the linear path is reversed. The other possibly unique materials are in a sense, birefringent. Their nature has yet to be described by saying that they have different indexes for right- and left-hand circularly polarized light. Superposition of right and left circularly polarized light outside our interests produces elliptically polarized light.

Circular polarized light has another interesting property—it carries angular momentum (about the direction of propagation). To illustrate this, suppose that such light falls on an atom represented by a harmonic oscillator that can be displaced rapidly in any direction. In the plane of "Fig. 33-2" the x displacement of the electron will respond to the E_x component of the field, while the y -component responds, equally, to the zero- E_y component of the field but 90° behind it in phase. That is, the responding electron goes around in a circle with angular velocity ω in response to the rotating electric field of the light (Fig. 33-9). Depending on the direction and orientation of the response of the oscillator, the direction of the displacement of the electron, and the direction of the force q_E it need not be the same but they relate as follows: If q_E may have a component in right angles to E , the work done on the system and change in θ is zero. The work done per second is q_E . Over a period of time T the energy absorbed is $q_E T$, and $q_E T$ is the angular momentum delivered to the mass about the E axis. We see the effect that a beam of right circularly polarized light communicates a total energy $q_E T$ and angular momentum (work) $q_E T$ directed along the direction of propagation. For whom this beam is absorbed the angular momentum is delivered to the absorber. Left-hand circular light carries angular momentum of the opposite sign, $-q_E T$.

Relativistic Effects in Radiation

34-1 Moving sources

In the present chapter we shall describe a number of interesting effects in connection with radiation, and then we shall be finished with the classical theory of light propagation. In our analysis of light we have concentrated in great detail on only phenomena of very low speed associated with electro-magnetic radiation that we have not discussed is what happens if the speeds are increased to a few $\times 10^8$ cm/sec. Effectively, the size of the band being comparable to a wavelength — one to ten times slower than light itself. The phenomena of so-called radio waves and microwaves we shall discuss later; we shall first use another physical example: sound, and the next chapter will be on ultrasound. Except for this, the present chapter is a brief consideration of the classical theory of light.

Very soon we shall see the effects that we shall now discuss by performing test they have to do with the effects of moving sources. We no longer assume that the source is localized, with all its motion being at a relatively low speed near a fixed point.

We know that the fundamental laws of electro-dynamics say that, at large distances from a moving charge, the electric field is given by the formula

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 c^2} \frac{\mathbf{v}_S}{v_S^2 + c^2} \quad (34.1)$$

The second derivation of this first vector eqn. was given in the appendices to the notes, > the lecture notes on the electric field. This last vector does not point toward the present position of the charge, of course, but rather in the direction that the charge would seem to be, if the information traveled only at the finite speed c from the charge to the observer.

Associated with the electric field is a magnetic field, always at right angles to the electric field and at right angles to the apparent direction of the source, given by the formula

$$\mathbf{B} = -\mathbf{v}_S \times \mathbf{E} \quad (34.2)$$

Until now we have considered only the case in which motion is one-dimensional, at speed, so that there is no significant change in the direction of the source to be considered. Now we shall be more general and study the case where the motion is in x - y - z very arbitrary, and ask what effect this may be expected in those circumstances. We shall let the motion be an arbitrary speed, but of course we shall let it assume that the detector is very far from the source.

We already know from our discussion in Chapter 26 that the only thing that counts in $d\mathbf{v}_S/dt$ are the changes in the direction of \mathbf{v}_S . Let the position vector of the charge, $\mathbf{r}_S(t, x, y, z)$, with a measured along the direction of motion (Fig. 34-1). At a given moment in time say, $t = t_0$, the three components of the position are x_0 , y_0 , and z_0 . The distance R is very nearly equal to $(x_0^2 + y_0^2 + z_0^2)^{1/2}$. Now the direction of the vector \mathbf{v}_S depends mainly on x_0/R , but how you y_0/R and z_0/R . Now the three components of the unit vector are x/R and y/R , and when we differentiate these components we get things like x^2/R^2 in the denominators:

$$\frac{dx(R)}{dt} = \frac{dx/dt}{R} = \frac{dx}{dt} \frac{x}{R} \quad$$

34-2 Moving sources

34-2.1 Finding the "apparent" motion

34-2.2 Synchrotron radiation

34-2.3 Cherenkov radiation

34-2.4 Ultrasonics

34-2.5 The Doppler effect

34-2.6 The x, y, z four-vector

34-2.7 Aberration

34-2.8 The connection of light

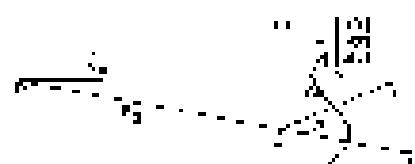


Fig. 34-1. The path of a moving source. The true position of the source is at t , but the retarded position, at $t + dt$, is used to find the apparent motion.

So when we are far enough away the only terms we have to worry about are the derivatives of x and y . Thus we take out the factors R_t and \dot{R}_t .

$$\begin{aligned} E_x &= -\frac{q}{4\pi\epsilon_0 R_t^2} \frac{\partial^2 x}{\partial t^2} \\ E_y &= -\frac{q}{4\pi\epsilon_0 R_t^2} \frac{\partial^2 y}{\partial t^2}. \end{aligned} \quad (34.8)$$

where R_t is the distance, measured to the origin of the coordinates (x, y, z) . Thus the electric field has a component anti-parallel to a very simple thing, the second derivatives of the x and y coordinates. (We could put it more mathematically by calling x and y the transverse components of the position vector r of the charge, but this would not add to the clarity.)

Of course we realize that the coordinates must be measured at the recorded time. Does we find that $\ddot{x}(t)$ does affect the calculation. Why? isn't the potential $V(t)$? If the time of observation is t , (the time of \ddot{x}) then the time τ to which this corresponds is $\tau = t - \Delta t$, but is delayed by the usual distance due the light having to go, divided by the speed of light, c , before approximation. This delay is R_t/c , a constant for uninteresting features, but in the next approximation we must include the effects of the position in the calculation of the time τ . Because it's a little further back, there is a little more retardation. This is an effect that we have neglected before, and it is the only change needed in order to make our results valid for all speeds.

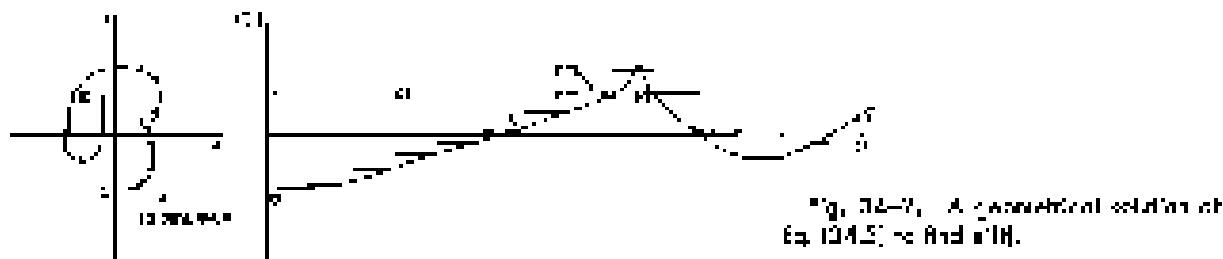
What we must now do is to choose a certain value of τ and calculate the value of \ddot{x} from it, and then find out where x and y are at τ . That is, given the recorded x and y , which we call x' and y' , where several derivatives determine the "tilt". This τ is determined by

$$\tau = t + \frac{R_t}{c} \div \frac{\dot{x}(t)}{c}$$

and

$$x'(t) = x(\tau), \quad y'(t) = y(\tau). \quad (34.9)$$

Now these are complicated equations, but it is easy enough to make a geometrical picture to describe their solution. This picture will give us a good qualitative feeling for how things work, but it still takes a lot of detailed mathematics to reduce the physical analysis of a complicated problem.



34-2 Finding the "apparent" motion

The above equation has an interesting simplification. If we disregard the time-varying component delay R_t/c , what we really did is just change the origin of τ by a constant. Then it says that

$$\ddot{x} = \alpha + \beta(t), \quad x' = x(t), \quad y' = y(t) \quad (34.10)$$

Now we need to find α and β as functions of t and τ , and we can do this in the following way: Eq. (34.9) says that we should take the recorded time and add a constant (the speed of light) times τ . What this comes out to mean is shown in Fig. 34-2. We take the actual motion of the charge (shown at 't') and imagine that it is going around it is being swept away from the point t at the speed c (there are no corrections for relativity or anything like that; this is just a mathematical addition of the τ). In this way we get a new motion, in which the line

of sight coordinate x , as shown at the right. (The figure shows the result for a rather complicated motion at a place, but of course the motion may not be to one point—it may be over some complicated curve motion in a plane.) The point is that the horizontal distance between the new x is no longer the old x , but is $x - ct$. And therefore is \vec{r} . Thus we have found a picture of the curve. A good argument! All we have to do is find the field is to look at the acceleration of the curve, i.e., to differentiate twice. So the final answer is: in order to find the electric field for a moving charge, take motion of the charge and calculate a book which speeds up "open it up." On the curve, so drawn, is a curve of the x and y positions of the function of t . The acceleration of this curve gives the electric field as a function of t . Or, if we like, we can say imagine that this whole "book" drove away forward at the speed c . I wish the plane of sight, so that the point of intersection with the plane of sight has coordinates x and y . The acceleration of this point makes the electric field. This solution is just as correct as the formula we started with—simply a geometrical representation.

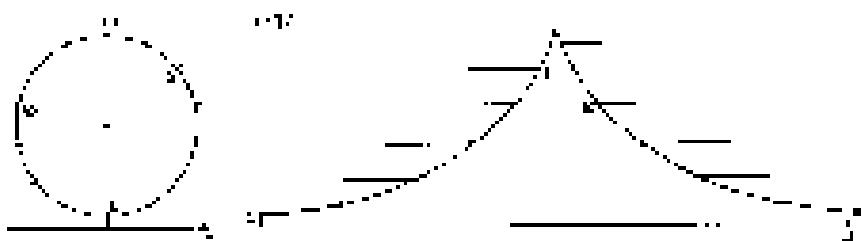


Fig. 24-3. The field curve for a particle moving at constant speed $v = 0.9cc$, a circle.

If the motion is relatively slow, so, for instance if we were in reality as just going up and down slowly, then when we went that much away at the speed of light, we would just of course, a simple engine curve, and that gives a formula we have been making out for a long time, it was the field produced by an accelerating charge. A more interesting example is an electron moving, exactly, very nearly at the speed of light, i.e., a circle. If we look in the plane of the circle, the retarded field appears as shown in Fig. 24-3. What is this curve? It is the y -axis given a radius vector from the center of the circle to the charge, and if we extend this vector by a little bit past the charge, just a shade if it is going fast, then we come to a point on the line that goes to the speed of light. The line, which we consider the motion back to the speed of light, can correspond to having a "wind" with a charge on it rolling backward (without slipping) at the speed c ; this is not a curve which is very close to a cycloid—it is called a hyperboloid. If the charge is going very nearly at the speed of light the "wings" are very sharp indeed; if it were actually the speed of light they would be almost exactly infinitely sharp. "Ininitely sharp" is interesting; it means that near c the slope of the curve is enormous. That is, in each cycle there is a sharp pulse of electric field. This is not at all what we would get from a nonrelativistic motion, where each time the charge goes around there is a uniform field at all other times ("strength") E the time. Instead, there are very sharp pulses of electric field spaced at time intervals $1/T$, equal where T is the period of revolution. The strong electric fields are caused in a sense due to the direction of motion of the charge. When the charge is moving away from us, there is very little resistance and there is very little radiation field in the direction of P .

24-3 Radiation radiation

We have very frequently seen moving in circular paths in the synchrotron; they move exactly at very nearly the speed c , and the point of interest above occurred as you might expect. Let's discuss this in more detail.

In the synchrotron we have electrons which go round in circles in a uniform magnetic field. First let us see why they go in circles. From Eq. (12.10), we know that the force on a particle in a magnetic field is given by

$$\vec{F} = q\vec{v} \times \vec{B}, \quad (24.1)$$

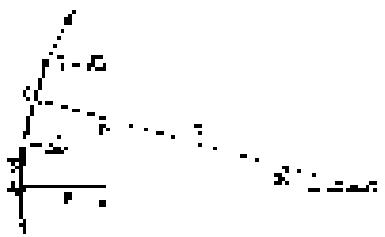


Fig. 34-6. A charged particle moves in a circle by radial path in a uniform magnetic field.

and v is at right angles both to the field and to the velocity. As usual, the force is equal to the rate of change of momentum with time. If the field is directed upward out of the paper, the weight $m g$ of the particle and the force on it are as shown in Fig. 34-4. Since F is perpendicular to the velocity, the kinetic energy and therefore the speed remain constant. All the magnetic field does is to change the direction of motion. In a short time Δt , the momentum vector changes at right angles to itself by an amount $B p = B m v$, and therefore p turns through an angle $\Delta\theta = B p \Delta t / m v = q B \Delta t / m$, since $p = m v$. But in this same time the particle has gone a distance $d = v \Delta t$. But only the two lines 1.8 and 2.0 mi intersect at a point O such that $O A = O B = R$ where $R = v \Delta t$. Comparing this with the previous expression, we find $R = v \Delta t = q B \Delta t / m$, from which we find

$$v = q B R \quad (34-7)$$

and

$$R = q B / m. \quad (34-8)$$

Since this argument can be applied during the next instant, the next, and so on, we conclude that the particle must move, in a cycle of radius R , with angular velocity ω .

The result that the momentum of the particle is equal to $q B R$ times the radius times the magnetic field is a very important law that is used a great deal. It is important for practical purposes because if we have extremely particles which all have the same charge, and we observe them in a magnetic field, we can measure the ratio of currents of their orbits and, knowing the magnetic field, thus determine the momenta of the particles. If we multiply both sides of Eq. (34-7) by q , and express $q m$ in terms of the electron charge, we can measure the momentum in units of the electron volt. In these units our formula is

$$p e(v) = 1.8 \times 10^9 (q/m) B R, \quad (34-9)$$

where B , R , and the speed of light are all expressed in C.G.S. system, the latter being 1.8×10^9 , numerically.

The value of magnetic field is called a gauss per square meter. This is an older unit which is still in use today, called a gauss. One gauss m^{-2} is equal to 10^4 oersted. To give an idea of how big a magnetic field is, one of the strongest magnets made thus far can usually make an intensity of about 1.1×10^4 gauss, beyond that the advantage of using iron disappears. Today, electric currents around wire superconducting at the melting temperature produce fields of over 10^6 gauss strength, that is, 10 kilogauss. The field of the sun is a few tenths of a gauss at the equator.

Returning to Eq. (34-9), we can imagine the synchrotron running a billion electron volts beam with 10^9 per billion electron volt. (We shall come back to the energy in just a moment.) Then, if we had a B corresponding to one 1000 gauss, which is a 10^4 oersted field, our R would be 10^9 cm, or 10 km. It would have to be 2.3 meters. The polar radius of the California synchrotron is 2.7 miles, the field is a little bigger, and the energy is 1.5 billion, but it is the same idea. So now we know a feeling for why the cyclotron has to stay at low E .

We have calculated the momentum. We also know that the total energy including the rest energy, is given by $E = \sqrt{p^2 c^2 + m^2 c^4}$, and for an electron the rest energy corresponding to $m c^2$ is 0.511×10^6 ev, or when p is 10^9 ev we neglect $m c^2$, and as for all practical purposes the p is small, the speed is relativistic. It is practically the same as say the energy of an electron in a billion electron volts as in eq. 34-9, numerically (that is, a billion-electron-volt $E^2/2P = 10^9$ ev), it is easy to show that the speed C is less than the speed of light by less than one millionth.

We shall now let the particle emitted by such a particle. A particle having 20.0 miles of curves 1.5 meters in 20 meters circumference, since around must be roughly the time it takes light to go 20 meters. So the wavelength λ would be emitted by such a particle would be 20 miles in the x-ray wave region. But because of the filling up effect that we have been calculating (Fig. 34-5), and because the distance by which we may extend the radius to reduce the speed is 24-4

only one part of each Miller of the radius. The steps of the hyperbolas are conveniently sharp compared with the distance between them. The acceleration, which increases toward the positive side, is equal to $(m/e)E$, where E is the voltage, for $m/e = 2 \times 10^8$ becomes, on the scale, a value of eight million times the acceleration of the electrons. Thus we might say that effective space would be made shorter, to the extent of 64 times by 10^8 if the electrons had been confined to the γ -ray region. Actually, the step size is not the only determining factor, since there also influence is the region where the step. The change in frequency is determined of the source, but it is also a consequence of the motion. Thus, even though a slowly moving electron would have more time between radiations, the frequency of the wave it is emitting is so much that we can ignore it. Now, the light should be accelerated with constant frequency parallel to the direction of motion.

It further appears that we would also be supposed that we could make such light no sharply enough because the source is so far apart from us, we can just take one reflector and direct it under a different angle, so that it is at 90° to scattering wires. After the pulse leaves away from the grating, what does it do? (We should see it again, but in glass of course, it would not light at all.) Well, as we see? The pulses strike the glass surface and all the oscillations of the grating, being here no electric field, need not add their rays shown again, but cancel. They have source shown in ω form, ω is shown in Fig. 34-5. Let the point P be at the end of the grating arm in the ω direction, ω and ω' . The field is held away from wind A , now from B , and so on; finally, the pulse from reflector surface. It goes to one of the reflectors, all the successive waves cancel out. In Fig. 34-5(a), the electric field with its successive pulses and the very like a long wave whose wavelength is the distance between the pulses just as would be the wave going out of the slit along the ω axis. So we get shorter light all right. But, ω in some amount will be too great. Then another kind of a "pulse" like supposed to be made up of such two other waves or wave and of the scattered waves together generated by a small angle between them. This is seen. Then we see that the field would not shorten it to a very short length, because each pulse does not carry much of the wave function, scattered pulses.

The electric magnetic radiation emitted by a radio is called radio waves. Induction in a magnetic field is called magnetism. It is to name the activities of man, but, almost, nothing specifically to synthetic ones, or even man's natural properties. This is the main reason that it also occurs in nature.

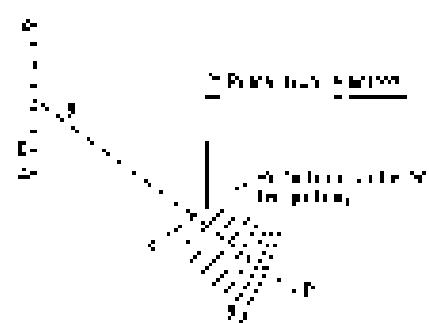


Fig. 34-5. The light after striking a grating at a large angle ω from ω' will have a certain reflection angle at different colors.



Fig. 34-6. The electric field E is the same as in Fig. 34-5, but the radio wave has a much larger amplitude.

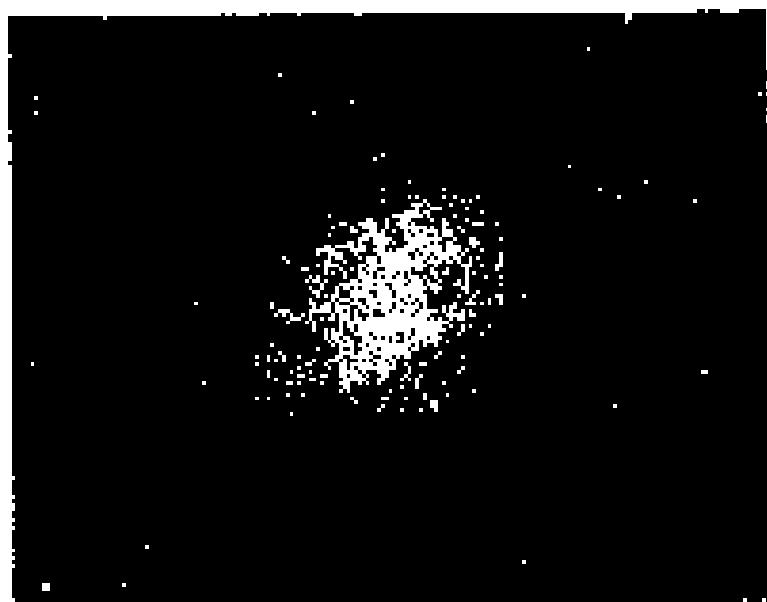


Fig. 34-7. The multi-colored pattern on a color filter.

34-4 Cosmic gamma-ray radiation

In the year 1054 the Chinese and Japanese citizens were watching the most brilliant display of fireworks in the entire universe, and they recorded, more or less graphically, an explosive bright star that year. (It is interesting to note that the European monks, writing in the books of the middle ages, even bothered to write in Latin especially for the sky, but they did not.) Today we may take a picture of that star, and what we see is shown in Fig. 34-7. On the outside is a big mass of red filaments which are produced by the stars of the hot gas "boiling" at the surface. In practice this means a bright line spectrum with different frequencies in it. The red component is the one to be due to nitrogen. On the other hand, in the central region is a magnetized, very point of light in a continuous distribution of frequency, i.e., there are no special frequencies associated with particular resonance. This is not dust "blown up" by nuclear stars, which is one way by which one can get a continuous spectrum. We can see stars through it, so it is transparent, but it is glowing light.

In Fig. 34-8 we look at the same object, using light in a region of the spectrum which has no bright spectral lines, so that we examine the central region. When the two solar polarizers have been put on the telescope, and the two views correspond to two orientations SIP and T. We see that the pictures are different. This is due to the light polarization. The reason, presumably, is that there is a large magnetic field, and many very energetic electrons are going around in that magnetic field.

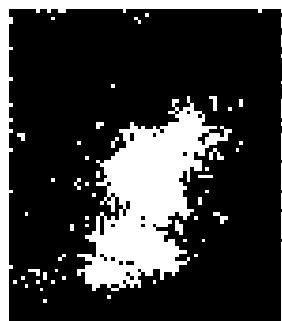
We have just illustrated that the electrons could go around the field in a circle. We'd add to this statement say uniform motion in the direction of the field, since the force $(qv) \times B$, has no component in this direction and, as we have already realized, the stimulated radiation is evidently polarized in a direction at right angles to the projection of the magnetic field onto the plane of sight.

Putting these two facts together, we see that, in a region where one picture is bright and the other one is black, the light in the first is clearly field-compared polarized in one direction. This means that there is a magnetic field at right angles to this direction, while in other regions where there is no strong emission in the center picture, the magnetic field runs in the other way. If we look carefully at Fig. 34-4, we may notice that there is, roughly speaking, a general set of "lines" but one way in one picture and at right angles to this in the other. The pictures correspond to those directions. Presumably the magnetic field lines will tend to extend indefinitely long distances in their own direction, and so, presumably, the two long regions of magnetic field will all be running spiraling one way, while in another region the field is the other way and the electrons are also spiraling that way.

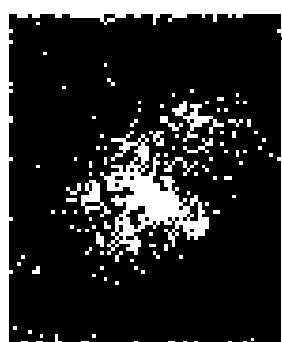
What keeps the electron energy so large for so long a time? At least it is 1000 years since the explosion. How can they keep going so fast? How they maintain their energy and how this whole thing keeps going is still not thoroughly understood.

34-5 Decelerating

We can now remark briefly on one other interesting effect of a very fast-moving particle that requires energy. The idea is very similar to the one we have just discussed. Suppose that there is an uncharged particle in a piece of paper and a very fast electron, say coming by (Fig. 34-9). Then, because of the electric field around the electric charges, the electron is pulled, accelerated, so that the curve of its motion has a slight kink or bend in it. If the electron is travelling at very nearly the speed of light, what is the electric field caused in the direction of? Remember our rule: we take the electric field from translation (not velocity) if speed v , and that's exactly where things measure the electric field. It was moving toward us at the speed v , so the field must run with the whole picture compressed into a small distance in proportion as $v = c$ is smaller than c . So, $E = \infty$? There is a very sharp and rapid decrease of E , and when we take the second derivative of that we get a very high field in the direction of the motion. So when very energetic electrons move through matter they spit radiation in a longitudinal direction. This is called decelerating. As a matter of fact, they don't



[a]



[b]

Fig. 34-7. The same object is seen through a blue filter and a polaroid. [a] Magnetic vector vertical. [b] Electric vector horizontal.



Fig. 34-9. A fast electron moving near a massive nucleus energy in the direction of its motion.

we would have to go to higher-energy electrons (actually if we could get them out of the machine more conveniently we could just say this) so to make very energetic photons—gamma rays. By passing the energetic electrons through a field larger than $\frac{1}{2} m_e c^2$ one letting them radiate, photons from this bremsstrahlung effect.

34-6 The Doppler effect

Now we go on to consider some other examples of the effects of moving sources. Let us suppose that the source is a stationary atom which is oscillating at one of its natural frequencies ω_0 . Then we know that the frequency of the light we would observe is ω_0 . But now let us take another example, at which we have a simple oscillator with length a with a frequency ω_0 , and at the same time the source moves, the whole oscillator, is moving along it's direction toward the observer at velocity v . Then the actual motion of source is as shown in Fig. 34-10(a). Now we play our usual game we add v ; that is to say, we calculate the wave curve between and beyond the source, as in Fig. 34-10(b). In a given amount of time t , when the oscillator would have gone a distance a , on the x - y coordinate it has a displacement $= -vt$. So all the oscillations of frequency ω_0 in the last $2a$ are now found in the interval $2a - (-vt)$ and $2a$; they are equivalent to $\omega_0 t$, and appear to us since $v \gg a$ speed c , now with a light of a higher frequency, higher by just the compression factor $(1 - v/c)$. This we observe:

$$\omega = \frac{\omega_0}{\sqrt{1 - v/c}}. \quad (34-10)$$

We can, of course, analyze it's situation in various ways. Suppose that the source were emitting instead of continuous, a series of pulses, pip-pip-pip-pip at a certain frequency ω_0 . At what frequency would they be received by us? The first one that arrives would be in full, but the next one is delayed because it's traveling the extra distance due to the receiver. Therefore, the time between successive pulses, instead of the interval t was analysis the geometry of the situation, we find that the frequency of the pipe is increased by the factor $1/\sqrt{1 - v/c}$.

If $v = \omega_0/c(1 - v/c)$, then, the frequency that would be observed if we took an ordinary atom, which had a natural frequency ω_0 , and caused it to move the source at speed v , the source with $v/c < 1$, the altered frequency ω_1 of a uniform atom is not the same as that measured when $v \gg a$ and $v/c \ll 1$, because of the relativistic addition in the rate of passage of time. That is, we're the $1/(1 - v/c)^{1/2}$ frequency, i.e., the endothermic frequency, would be

$$\omega_1 = \omega_0 \sqrt{1 - v/c}. \quad (34-11)$$

Therefore the observed frequency is

$$\omega = \frac{\omega_0 \sqrt{1 - v/c}}{1 - v/c}. \quad (34-12)$$

The shift in frequency observed in the above situation is called the Doppler effect. If something moves toward the light it emits appears more violet, and if it moves away it appears more red.

We shall now give an analysis of this same interesting and important result. Suppose, now, that the source is also doing still another motion, at frequency ω_0 , while the observer is moving with speed v toward the source. After a total period of time t , the source will have moved to a new position, a distance $v t$ from where it was at $t = 0$. How many radians of phase $\omega_0 t$ is there now? At a certain number, say, went past any given point, and in addition to that it has kept past some more by his own motion, carrying a number $v t$, the number of cycles is $\omega_0 t + v t$ times the cycles per sec. So the total number of radians in the time t , or the observed frequency, would be $\omega_0 + v \omega_0$. We have made this analysis from the point of view of a inertial rest. We would like to know how it would look to the man who is moving. Then we have to worry again about the difference in clock rate for the two observers, and this time the results that we have to multiply v/c by ω_0 . So if ω_0 is the source's number of cycles

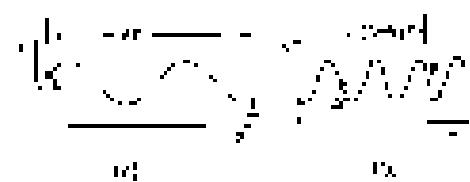


Fig. 34-10. The $x = x$ and $x' = t$ curves of a moving oscillator.

the vector is the direction of motion, and ω_0 is the frequency, then the observed frequency for a moving man is

$$\nu = \frac{\omega_0}{\sqrt{1 - v^2/c^2}} \quad (34.13)$$

For the case of light, we know each $\lambda_0 = \lambda/c$, so, in this particular problem, the equation $\nu = \lambda/c$ read

$$\nu = \frac{\omega_0(1 - v/c)}{\sqrt{1 - v^2/c^2}}, \quad (34.14)$$

which looks completely unlike formula (34.12)! The frequency that we would observe if we never moved a source different than the frequency that we would see if the source moved toward us? Of course not! The theory of relativity says that v/c must be zero by definition. If we don't want strange mathematical results we would naturally recognize this. These two measurement frequencies are exactly equal. In fact, the *observed* equality of the two frequencies is one of the ways by which some people like to demonstrate the *realistic* requirement of time dilation, because if v/c did not fit these square-root factors in, they would no longer be equal.

Since we know about relativity, let us analyze it in full & think why, while they appear to differ, there *are* general. It is really the same thing since it makes no difference how we do it! According to the relativity theory there is a relationship between position and time as observed by one man and another, and time as seen by another who is moving relative to him. We will show these relationships long ago (Chapter 18). This is the Lorentz transformation and its inverse:

$$\begin{aligned} t' &= \frac{t - vx}{\sqrt{1 - v^2/c^2}}, & r' &= \frac{r - vc}{\sqrt{1 - v^2/c^2}}, \\ r' &= \frac{1 + v/c}{\sqrt{1 - v^2/c^2}}, & t &= \frac{t' + v/c}{\sqrt{1 - v^2/c^2}}. \end{aligned} \quad (34.15)$$

If we were standing still on the ground, the term r/v would be constant. And so if the reader and author and audience would follow that term. But what can a man do now, especially the non-physical ones, well? Well, if the field is zero, the positions of all the readers are the same (within the field is zero, everyone measures the field is zero); that is a relativistic invariant. So the term is the sum of the x 's, and we expect that we will transform it into the form of reference

$$\cos(\omega t - kx) = \cos\left[\omega \frac{t - v/c}{\sqrt{1 - v^2/c^2}} - k \frac{r - vc}{\sqrt{1 - v^2/c^2}}\right].$$

If we expand the terms inside the brackets, we get

$$\begin{aligned} \cos(\omega t - kx) &= \cos\left[\frac{\omega - k\omega}{\sqrt{1 - v^2/c^2}} t - \frac{v - \omega/c^2}{\sqrt{1 - v^2/c^2}} x\right] \\ &= \cos\left[\omega t - \omega - \frac{v - \omega/c^2}{\sqrt{1 - v^2/c^2}} x\right]. \end{aligned} \quad (34.16)$$

This is $\omega' + \omega/c^2$, a wave, in which there is a certain frequency ω' , a wave number k' , and some other constant, ω/c^2 , multiplying ω . We call ω' the wave number or the number of waves per meter, for the other man. Therefore the other man will see a new frequency and a new wave number given by

$$\omega' = \frac{\omega + kv}{\sqrt{1 - v^2/c^2}}, \quad (34.17)$$

$$k' = \frac{k - \omega/c^2}{\sqrt{1 - v^2/c^2}}. \quad (34.18)$$

If we look at (34.14), we see that ν is the same for rule (34.18), then we obtained by a more physical argument

34-7 The ω , k four-vector

The relationships indicated in Eqs. (34.17) and (34.18) are very interesting, because they say that the new frequency ω' is a combination of the old frequency ω and the old wave number k , and that the new wave number is a combination of the old wave number and frequency. Now the wave number is the rate of change of phase with distance, and the frequency is the rate of change of phase with time, and in these two equations we see a close analogy with the Lorentz transformation, in which the time coordinate t is divided by c and the space coordinate x is divided by c . Then the new ω' will be like ω , and the new k' will be like k/c^2 . That is to say, because the Fourier transformation would transform the wave number k in exactly the same way as it transforms the frequency ω , when a quantity has four components transforming like time and space, it is a four-vector. Everything about ω is right. But, except for one little thing: ω and k as four-vectors have to have four components: where are the other two components? We have seen that wave k in the time and space is the \hat{x} -axis direction, but ω is in all directions, and so we must solve fully the problem of the propagation of light in three space dimensions, not just in one dimension, or we have been doing up idealism.

Suppose that we have a coordinate system $x, y, z, \text{and } t$ with x is travelling along and whose variation axes are shown in Fig. 34-1. The wavelength of the wave is λ , but the direction of motion of the wave does not happen to be \pm the direction of one of the axes. What is the formula for such a wave? The wavevector already has $(\omega - kx)$, where $k = k_x \hat{x}$ and x is my distance along the direction of motion of the wave—the component of the spatial position in the direction of motion. Let us put it this way: if $x > 0$ the vector position of a point in space, \mathbf{r} , is $\mathbf{r} = \hat{x}x$, where \hat{x} is a unit vector in the direction of motion. That is, x is just $\cos(\theta, \hat{x})$, the component of \mathbf{r} in the direction of motion. Then the wavevector is $\omega - k_x \hat{x} - k_y \hat{y} - k_z \hat{z}$.

Now it turns out to be very convenient to define a vector \mathbf{k} , which is called the ω , k vector, which has a magnitude equal to the wave number, Eq. 34.16, and is pointed in the direction of propagation of the wave:

$$\mathbf{k} = 2\pi\mathbf{v}/\lambda = \mathbf{v}\hat{\omega} \quad (34.19)$$

Using this vector, our wave can be written as $\cos(\omega t - \mathbf{k} \cdot \mathbf{r})$, or as $\cos(\omega t + \hat{\omega}_x x - \hat{\omega}_y y - \hat{\omega}_z z)$. What is the significance of a component $\hat{\omega}_x$? $\hat{\omega}_x$ is $\omega \sin \alpha$. Clearly, $\hat{\omega}$ is the rate of change of phase with respect to x . Referring to Fig. 34-11, we see that the phase changes as we change x , just as if there were a wave along a fan of a large-wavelength. The “wavefronts” in the x -direction“ \rightarrow longer than a natural, free wavelength by the secant of the angle α between the actual direction of propagation and the x -axis.

$$\hat{\omega}_x = \omega/\cos \alpha \quad (34.20)$$

Because the rate of change of phase, which is proportional to the reciprocal of λ_ω , is smaller by the factor $\cos \alpha$, that is just how k_x would vary if we took the magnitude of \mathbf{k} times the cosine of the angle between \mathbf{k} and the x -axis!

There, then, is the nature of the wave vector (and we use to represent a wave in three dimensions). The four quantities ω , $\hat{\omega}_x$, k_x , $\hat{\omega}_y$, k_y , $\hat{\omega}_z$, k_z transform in analogy as a four-vector, where ω corresponds to the time, and $\hat{\omega}_x$, $\hat{\omega}_y$, $\hat{\omega}_z$, k_x , k_y , k_z correspond to the x , y , z , and t components of the four-vector.

In our previous discussion of gravitational relativity (Chapter 17), we learned that there are ways of taking dot-products with four-vectors. If we take the position vector \mathbf{r}_ω , where ω stands for the four components (time and three space ones), and if we call the wave vector \mathbf{k}_ω , with ω the time, \mathbf{k}_ω has four space components, time and three space ones, then the dot product of \mathbf{r}_ω and \mathbf{k}_ω is written $\Sigma k_\omega \epsilon_{\omega i}$ (see Chapter 17). This last relation is an invariant, independent of the coordinate system: what is it equal to? By the definition of a dot product for four-vectors, it is

$$\Sigma k_\omega \epsilon_{\omega i} = \omega - \hat{\omega}_x x - \hat{\omega}_y y - \hat{\omega}_z z \quad (34.21)$$

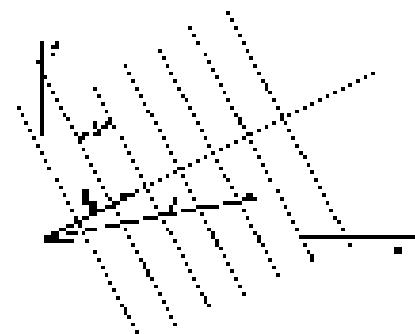


Fig. 34-11. A plane wave traveling in one direction.

We know from our study of vectors that $\nabla \cdot \vec{v}_r$ is invariant under the Lorentz transformation, since \vec{v}_r is a four-vector. But this quantity is precisely what appears inside the braces for a point vector, and it ought to be invariant under a Lorentz transformation. We connect this to a formula with something like a "charge inside the braces," since we know that the place of having connected charge with we change the coordinate system.

24-8 aberration

In deriving Eqs. (24.7) and (24.13), we have taken a simple example where it happened to be a direction of motion, but, of course, we can generalize to other cases also. For example, suppose there is a source sending out light in a general direction but at a point of view of $\vec{v}_r = 0$, at rest, but we are moving along on the earth, as in Fig. 24-12. From which direction does the light appear to come? To find this, we will need to write down the four components of \vec{v}_r and apply the Lorentz transformation. The answer, however, can be found by the following argument: we have to point our telescope at an angle to see the light. Why? Because the light is coming down at the speed c , and we are moving sideways at the speed v , so the telescope has to be tilted forward so that the light comes down at an angle, slower than the rate. It is very easy to see that the horizontal distance is given when the vertical distance y is given; therefore, x is the hypotenuse, and $\theta = \pi/2 - \phi$. However! How nice, indeed, even for such a thing! If θ is the angle at which we would have seen the telescope relative to the earth, because we made our telescope. The point of view is a "fixed" observer. When we stand the horizontal distance is up, the angle of the earth would have been a different θ . Determining θ measured with a "spinning" robot, informs us that because of that centrifugal effect,

$$\tan \theta = \frac{v/c}{\sqrt{1 - v^2/c^2}}$$
 (24.14)

which is equivalent to

$$\sin \theta = \gamma \phi. \quad (24.15)$$

It will be the reader's for the student to derive the result using the Lorentz transformation.

This effect, the telescope has to be tilted, is called aberration, and it has been observed. Now can we observe it? What may occur where a given star should be? Suppose we do have a fault in our wrong calculation in such a way, however, we know it is a wrong correction? Decrease the earth's speed all the same. Today we have to point the telescope one way; six months later we have to tilt the telescope the other way. But is it so we can tell that there is such an effect?

24-9 The aberration of light

Now we turn to a different topic. We have never, in all our discussion of the previous chapters, said anything about the effects of the magnetic field and its associated with light. Ordinarily, the effects of the magnetic field are very small, but there is one interesting and important effect which is a consequence of the magnetic field. Suppose that \vec{v}_r is $v \hat{i}$, and there is a current i passing all a charge, and driving that charge q , and down. We will suppose that the charge moves in the x direction, i.e., the motion of the charge is also in the x direction; it has a position r and a velocity \vec{v}_r , as shown, in Fig. 24-13. The magnetic field is at right angles to the electric field. Now as the electric field acts on the charge and makes it go up and down, what does the magnetic field do? The magnetic field acts on the charge (say an electron) only if the it is moving in the electric field direction, it is driven by the electric field, so the two of them work together. While the thing is going to the left, v has a velocity $v \hat{i}$, there is a force on it. It has a $v \times \vec{B}$; but in which direction is this force? It has the direction of the polarization of light. Therefore, when $v \hat{i}$, v is going at a charge and it is oscillating in a perpendicular \hat{j} -direction.

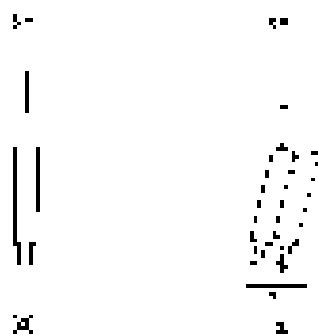


Fig. 24-12. A telescope on Earth is viewed by (a) stationary telescope and (b) a leisurely moving telescope.

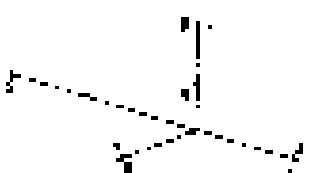


Fig. 24-13. Electromagnetic forces on a charge which is driven by the electric field and by the direction of the light beam.

charge there is a driving force in the direction of the light beam. This is called radiation pressure or light pressure.

Let us determine how strong the radiation pressure is. Once again it is $F = qE$, where q is the charge. If everything is oscillating, it is the time average of this. From (24.2), the energy E of the magnetic field is the same as the sum of the electric field squared by c , so we need to find the average of the electric field, times wave velocity, times the charge. This is $1/c$. ($F = qE/c^2$). But the charge q times the field E is the electric current on a charge, and this sum over the charge gives the radiation pressure for a single atom on the charge! Therefore the force, the "pushing momentum," that is delivered per second by the light is equal to $1/c$ times the energy absorbed from the light per second! That is a general rule, since we did not say how strong the oscillator was, or whether some of the charges varied more. In our circumstances stars light is being diminished above a percent. The momentum that the light delivers is always equal to the energy that is absorbed, divided by c :

$$F = \frac{dE/dt}{c}. \quad (24.24)$$

That light carries energy we already know. We now understand that it also carries momentum and hence, that the momentum is equal to $1/c$ times the energy.

When light is emitted from a source there is a recoil effect. The something is reverse. It can alone is carrying an energy E in some direction, then there is a recoil momentum $p = E/c$. If light is reflected normally from a mirror we get twice the force.

This is as far as we shall go using the classical theory of light. Of course we know that there is a quantum theory, and that in it light represents light acts like a particle. The energy of a light-particle is a constant times the frequency:

$$E = h\nu = pc. \quad (24.25)$$

We now express the fact that light also carries a momentum equal to the energy divided by c in another way. Let the effect be γ per photon. Then γ is momentum

$$\gamma = E/c = h\nu/c = p. \quad (24.26)$$

The direction of the momentum is, of course, the direction of propagation of the photon. γ is just p in vector form,

$$\gamma = \gamma_x \hat{i} + \gamma_y \hat{j} + \gamma_z \hat{k}. \quad (24.27)$$

We now show, of course, that the energy and momentum of a particle that forms a wavelet. We have just observed that ω and k form our vector. Therefore it is a good thing that (24.27) has the same form in both cases; it means that the quantum theory and the theory of relativity are morally consistent.

Equation (24.27) can be written more elegantly as $\gamma = \hbar k$, if we replace ν by ω , for a particle associated with a wave. Although we have discussed this only for photons, for which \hbar (the magnitude of \hbar) equals one and $\nu = \omega/c$, the relation is much more general. In quantum mechanics all particles, not only photons, exhibit wavelike properties, but the frequency and wave number of the wave is related to the energy and momentum of particle by (24.27) (called the deBroglie relation) even when \hbar is not equal to \hbar_{Planck} .

In the last chapter we saw that a beam of right or left circularly polarized light can carry angular momentum in an amount proportional to the energy E of the beam. In the (ν, ω, k) picture, a beam of circularly polarized light is represented as a screw of photons, each carrying an angular momentum $\pm \hbar$ along the direction of propagation. That is what the axes of polarization in the complex plane of ν —the previous ω —carry angular momentum, as spinning electrons do. But this "bullet" picture is really incomplete in the "wave" picture, and we shall have to discuss these ideas more fully in a later chapter on Quantum Relativity.

Color Vision

35-1 The Human Eye

The perception of colors depends partly on the physical world. We discuss the colors of light and what is being produced by interference. But then, obviously, it depends on the eye, or what happens behind the eye, in the brain. These characteristics of light that enter the eye, but also that our sensations are the result of physiological and psychological processes.

There are many interesting phenomena associated with vision which involve a mixture of physical phenomena and physiological processes, and the full appreciation of natural phenomena as we see them must go beyond physics to the social sciences. We make no apologies for making these excursions into other fields, since the separation of fields, as we have anticipated, is merely a human convenience, and an unnatural thing. Nature is not interested in our convenience; hence, and many of the interesting phenomena bridge the gaps between fields.

In Chapter 1 we have already discussed the relation of physics to the other sciences in general terms, but now we are going to look in considerably more specific field in which physics and other sciences are very, very closely interrelated. That one is vision. In geometry, "visual" effects refer to the present of light. At the present of light, we shall discuss mainly the observable phenomena of human vision, and in the next chapter we shall consider the physiological bases of vision, both in man and in other animals.

It all begins with the eye; so, in order to understand what phenomena we see, some knowledge of the eye is required. In the next chapter we shall discuss in some detail how the various parts of the eye work, and how they relate to the connection with the nervous system. For the present, we shall describe only briefly how the eye functions (Fig. 35-1).

Light enters the eye through the cornea; you have already learned how it is bent and is focused on the retina, called the "bottom" of the lens of the eye, so that different parts of the retina receive light from different parts of the visual field outside. The result is not necessarily uniform; there is a "fovea," a spot, in the center of our field of view which we can see. You can prove this yourself carefully, and of which we have the greatest clarity of vision if we call it the *foveal macula*. This is deep in the eye, so we can immediately appreciate from our experience in looking at things, why we are able to see best in seeing. That is in the center of the eye. There is also a spot in the fovea where the nerves are trying all the information out front; that is a blind spot. There is no sensitive part of the retina here, and it is possible to demonstrate this. You take a key, hold it up and look straight at something, and then move a finger or another small object slowly out of the field of vision, suddenly disappears, vanishes. This is just what would happen if that part of the eye were like a photographic negative, because quite a fellow, in the court of a king of France by painting his coat of arms, in the boring sessions that he was with his courtiers, the king could amuse himself by "cutting off their heads" by breaking at one and separating another head disappear.

Figure 35-2 shows a magnified view of the inside of the retina in somewhat schematic form. In different parts of the retina, the nerve fibers and optic nerves. The objects that occur more densely near the periphery of the retina are rod cells. Close to the fovea, we find, besides these rod cells, also cone cells. They are, I suppose, the receptors of the rods. As we get closer to the fovea, the number of cones increases, etc. In the fovea itself there are (in fact, nothing but) cone cells, packed very tightly, so (which fits), the cone cells are substituted, a narrower

35-2 The human eye

35-3 Color depends on frequency

35-4 Measuring the color spectrum

35-5 The chromaticity diagram

35-6 The mechanics of color vision

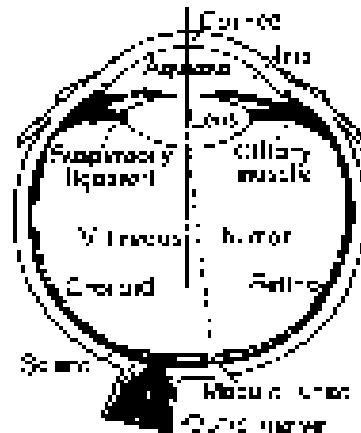


Fig. 35-1. The eye.

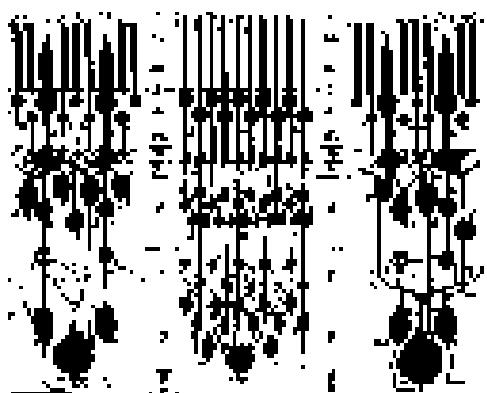


Fig. 35-2. The structure of the retina (light rays have been added).

Now there is something else. You must appreciate that we see with the cones right in the middle of the field of view, but as we go to the periphery we have: the rod cells, the rods. Now the interesting thing is: That is the part's each of the cells which is sensitive to light is not connected by a fiber directly to the optic nerve, but is connected to many other cells, which are the other members of the cone system. There are several kinds of cells; there are rods that carry the information toward the optic nerve. In other words, there are rods that are actually interconnected. "Locally." There are essentially two kinds of cells. The overall organization details are: The main thing we emphasize is: the "the light signal is already being 'thought about.'" To put it simply, the information from the various rods does not immediately go to the optic, spot for spot, but it is carried as a certain amount of the information has already been "thought" by a combining of the information from several visual receptors. It is important to understand that some brain-function processes occur in the eye itself.

35-2 Color depends on luminosity

One of the most striking phenomena of vision is the dark adaptation of the eye. If we pass into the dark from a brightly lit room, we cannot see very well. We're white, but gradually things become more and more discernible, and eventually we can see something where we could see nothing before. If the intensity of the light is very low, the things that we see have no color. It is known that this dark-adapted vision is almost entirely due to the rods, while the vision in bright light is due to the cones. As a result, there are a number of phenomena that we can easily appreciate because of this transfer of attention from the cones and rods that first engaged our eye.

There are many situations in which, if the light intensity were stronger, we would see colors, and we would find these things quite beautiful. One example is the through a telescope we usually always see "black and white" images of these nebulae, but W. C. Miller of the Mt. Wilson and Palomar Observatories had the permission to do color pictures of some of these objects. Naturally he could really see these colors with the eye, but they cannot be taken because it is merely that the light intensity is not strong enough for the cones in our eye to see them. As long as the main spectrum of such objects are the same nebula and the Cepheids. The former shows a definite blue color at night, bright red color during the day, and red-orange during the day, punctuated by bright red-orange flares.

In the bright light, apparently, the rods are at very low sensitivity - but in the dark, as time goes on they pick up their ability to see light. The variations in light sensitivity here which one can observe is over a million to one. Nature does not do all this with just one kind of cell, but she gives her job from bright-light-sensing cells, the cone-sensing cells, the cones, to low-intensity, dark-adapted cells, the rods. Again, the interesting consequence of this shift is this: first, that there is inversion, and second, that there is a difference in the relative brightness of sufficiently colored objects. It is to note that the rods are never trying the sun than the cones do, and the inversion can, for example, keep light while the rods find the color fully impossible to see. So what's black at first is the color now normal. For two pieces of colored paper, say this and that, in which the red might be even brighter than the human eye, will, in the dark, appear completely reversed. It is a very striking effect. If we are in a dark room and we find a magazine or something that has only a red, before we know for sure what the colors are, we judge the place and darker areas, and if we then take the magazine into the light, we may see this very remarkable shift in color which was the brightest color and which was red. This phenomenon is called the Purkinje effect.

In Fig. 35-3, the dashed curve represents the sensitivity of the eye in the dark, i.e., using the rods while the solid curve represents it in the light. We see that the peak sensitivity of the rods is in the green region, and that of the cones is more in the yellow region. If there is a red-colored page, (say a red nothing) we can see it if it is brightly lighted, but in the dark it is almost invisible.

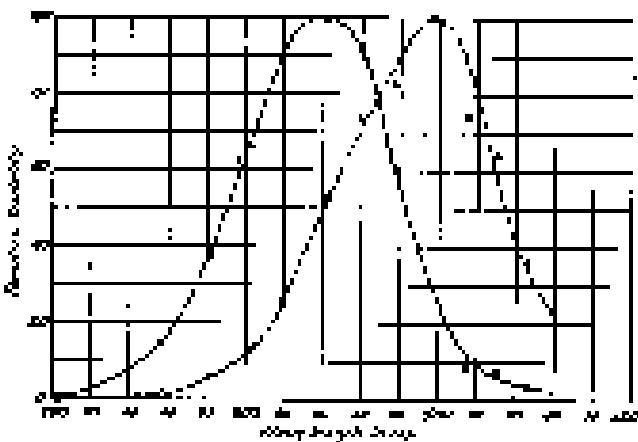


Fig. 35-9. The spectral sensitivity of the eye. Dashed curve, received & review.

Another effect of the Sun that only lasts over at the dusk, and then there are no clouds in the future, is that when we look straight at something in the dark, our vision is not quite as acute as when we look to one side. A faint star or nebula can sometimes be seen better by turning a little to one side than directly at it. However we do not have sensitive rods at the middle of the fovea.

Another interesting effect of the Sun is that the number of cones decreases as we go further out. The edge of the field of view is the last, even in a bright light, cover disappears as the object goes far enough away. The way to test that is to look in some particular field direction. Let a friend walk in from one side with colored cards, and try to decide what color they are before they are right in front of you. One hand can be seen out at the earliest distance, long before he can determine the color. When doing this, it is sufficient to come up from the side again, a few steps, since as it is converging rather easily to almost see the color, then does not see anything, then to see the color again.

Another interesting phenomenon is that the periphery of the retina is very sensitive to motion. Although we cannot see very well from the center of our eye, if a little bug comes and we do not expect anything to be moving over there we are immediately sensitive to it. We are all "wired up" to look for something jiggling to the side of the field.

35-3 Measuring the color stimulus

Now we pass from the more refined, to the brighter vision, and we come to the question which is most characteristic of wave vision, and the *color*. As we know, white light can be split by a prism into a whole spectrum of wavelengths which appear to us in many different colors; that is what colors are, of course, appearances. Any source of light can be analyzed by a grating or a prism, and one can determine the spectral distribution, that is, the "amount" of each wavelength. A certain light may have a lot of blue, considerable red, very little yellow, and so on. That is all very precise at the sense of vision, but the question is, what color will it appear to us? It is evident that the different colors depend somehow upon the spectral distribution of the light, but the problem is to find what characteristic of the spectral distribution produces the various colors. For example, what do we have to do to get a green color? We all know that we can simply take a glass of incandescent light which is green. But is that the only way to get green, or orange, or any other color?

In the course I am one spectral distribution which produces the same apparent visual effect. I hope answers a, definitely yes. There is a very limited number of your effects, in fact, just a three-dimensional manifold of them, as we shall shortly see, but there is an infinite number of different curves that we can draw to the light the colors from different sources. Now the question we have to discuss is, under what conditions the distributions of light appear as exactly the same color to the eye?

The true power of psychophysical technique in color judgment is to use the eye as a *unit instrument*. And so we do not try to find in which conditions a given sensation is to measure its whatness or what it gives us back because it turns out that this is not easily accomplishable. Instead, we study the conditions under which two stimuli are indistinguishable. Then we do not have to decide whether two people see the same coloration in different circumstances; h. only whether, if two people two sensations are the same. They can also decide for another. We do not have to decide whether, when one sees something green, what it feels like outside is the same as when he looks like inside someone else when he sees something green, we do not know anything about that.

To illustrate the possibilities, we may use a series of four projectors upon which have filters on them, and whose brightness are continuously adjustable over a wide range: one has a red filter and makes a spot of red light on the screen, the next one has a green filter and makes a green spot, the third one has a blue filter, and the fourth one is a white circle with a black spot at the middle of it. Now if we look on colored light, and not on a pure monochrom, we see that it is the case of overlap it produces a sensation which is not what we call reddish green, but a new color, yellow in this particular case. By changing the proportions of the red and the green, we can go through various shades of reddish green. But if we have an R, G, B, and a yellow, we can also obtain that same yellow not by mixing these two colors but by mixing some other ones, perhaps a yellow filter with some "green," or something like that, to get the same sensation. In this way, it is possible to make various colors by using the same way by mixing the lights from various filters.

What we have just discovered may be expressed symbolically as follows. A particular color, say, for example, can be represented by a certain symbol R which is the "sum" of certain amounts of red, green, light filtered, and colored light (G). By using determinate values and G , to describe how bright the R, and G are, we can write a formula for this yellow:

$$Y = cR + cG \quad (1)$$

The question is, can we make all the different colors by adding together two or three lights of different fixed colors? Let us see what can be done in real connection. We certainly cannot get all the different colors by mixing only red and green, because, for instance, the orange appears in such a mixture. However, if you try to make blue the central region, where all the lights overlap, may be made to appear a very fairly "ice white." By mixing the various colors and keeping at this center region, we find the *whole* possible visible range of colors in this region by changing the proportions, and so it is not impossible that all the colors can be made by mixing these three colored lights. We shall discuss this point. This is true; it is in fact, very likely correct and we shall shortly see how to confirm the proposition better.

In order to illustrate our point, we move the spots on the screen so that they all fall on top of each other and then we try to make a particular color which appears in the overlapping shade by the fourth lamp. What we find though, is that "orange" coming from the first, it always appears yellowish. One may try to match this by adjusting the red and green one. But as far as we can see a kind of error and error, and we find that we can approach rather closely this particular shade of "orange" color. So it is very hard to believe that we can't make the color. We can try to make yellow in a moment, but before we do that, there is no doubt that orange is very hard to make. People who give lectures on colorimetry, the "colorist" unless, but they never make brown, and it is hard to recall ever having seen brown light. As a matter of fact, this color is never used for any stage effects, nor never seen a yellow with brown light, so we think it might be impossible to make brown. In order to find out whether it is possible to make brown, we point out that brown light is mostly something that we are not used to seeing without its background. As a matter of fact, we can make it by mixing red and yellow. To prove that we are looking at brown light, we merely increase the brightness of the peculiar

superposed against which we see the very same light, and we see that that b. is full white, we call it yellow! Brown is always a dark color due to a lighter background. We can easily change the coloring of the brown— for example, if we have some green, and we get a reddish brown, apparently a chocolate reddish brown, and if we put some green to it, in proportion, we get the beautiful color which all the uniforms of the Army are made of, for the light from that color is not so sensible by itself as the yellowish green, but seen against a light background.

Now we put a yellow glass in front of the sun, to filter and try to see whether the intensity must be exactly as before the range of the various lights; we cannot make something which is too bright, because we do not have enough power to hold up. But we can make the yellow; without a green and no mixture, and you add a touch of blue to make it even more perfect. Perhaps we are ready to believe that, under good conditions, we can make a perfect mixture of any given colors.

Now let us discuss the laws of color mixing. In the first place, we find that different spectral distributions etc., produce the same color. Next, we saw that "why" could not be made by taking twelve these spectral colors, not if we had just one. The most interesting feature of color mixing is that if we have a certain light, which we may call X , and if it appears indistinguishable from Z to the eye (it may be a different spectral distribution, but it agrees with probabilities), we call them colors "equal." In the sense that the eye sees them as equal, we write

$$X = Y \quad (35.2)$$

Here is one of the great laws of color. If two spectral distributions are indistinguishable, and we add them to a certain light, say Z , of intensity $X - Z$, this means that we shine both lights on the same patch, and then we take X and add the same amount of the tinted or light Z , the two colors are also indistinguishable.

$$X + Z = X + Z \quad (35.3)$$

We have not mentioned the yellow; if we now shine pink light on the white China, it will still match. Shining any other light to the pink or light, we're a match. In other words, we can summarize all these exact relationships by saying, and once we have said it, the same two colored lights, say X and Z to each other in the same circumstances, then this makes X red, and one light can be substituted for the other light in any other color mixing situation. In fact, it turns out, and it is very important, and interesting, that this mixing of the colors of lights is not dependent upon the characteristics of the eye at the moment of observation; we know that, if we look at a bright red surface in a bright red light, and then look at a white paper, it looks grayish, and other colors are also changed by our seeing looked so long at the bright red. If we now have a match between, say, two surfaces, and we look at them and make the match, then we look at a bright red surface for a long time, and then turn back to the yellow, it may not look very red; but when what color it will look, and it will not look yellow. Now unless the surface and wall look washed, and we, as the eye, adapt to various ranges of intensity, the color match X works with the absolute exception of when we go into the region where the intensity of the light goes down, then we have shifted from one to the other, then the color match is no longer a color match, because we are using a different system.

The second principle of color mixing, of lights is this, any color of all can be made from three different colors, the primary, red, green and blue lights. By suitable mixing the three regresses we can create anything at all, as we demonstrated with our own examples. Furthermore, there are very interesting mutual relations. Let's see what are interested in the two halves of one thing, it turns out as follows. Suppose that we have our three colors which were red, green, and blue, but label them A , B , and C , and call them our primary colors. Then any color can be made by certain amounts of these three; say an amount x of color A an amount y of color B , and z amount of color C makes X .

$$X = ax + by + cz \quad (35.4)$$

Now suppose we have color Z & make from the same three colors:

$$Z = \sigma X + \tau Y + \rho Z' \quad (35.2)$$

This is just $\text{out } Z'$, the mixture of the original X, Y & Z & one of the components of the base that we have already previously obtained by taking the sum of the components of X, Y & Z :

$$Z = X + Y - (1 - \sigma)X - (\bar{\sigma} - \delta)Y + (\rho + \varepsilon)Z' \quad (35.3)$$

It is just like the mechanics of the addition of vectors when (α, β, γ) are the components of one vector and (x, y, z) are those of another vector, and the new light Z is just the "sum" of the vectors. This subject has always appealed to physiologists and mathematicians. In fact, Spinoza's wrote a wonderful paper on color vision¹ which he developed his theory of vector analysis as applied to the mixing of colors.

Now a question is, what are the correct primary colors in \mathbb{R}^3 ? There is no such thing as "the" normal primary colors for the mixing of lights. There may be, for practical purposes, some prims that are more useful than others. By putting a palette of colored pigments, but we can see differentiating them now day from differently colored lights & moreover can always be mixed in the correct proportion to produce any other color at it. Can we assume we have found our four? Instead of using red, green, and blue, let us use red, blue, and yellow in our projector. Can we use red, blue, and yellow to make, say, green?

By mixing these three colors in various proportions, we get quite an array of different colors, ranging from white to gray. But as a matter of course, after a lot of trial and error, we find that certain sets look like green. The question is, are we *sure* about it? The answer is, *now*², by projecting now red & blue the green, then we can make a match with a certain mixture of yellow and blue! So we have *two* blues there, except that we had to cheat by putting the red on the other side. But since we have two in the red-yellow location, we can approach it this way. We *only* showed was not that *any* colors always be made, say, of red, blue, and yellow, but by putting colored on the other side we found that red plus X could be made out of blue & yellow. That is, without loss of generality, we can interpret X as a negative number, so it *can* still allow that the coefficients in equation (35.4) can be both positive and negative, and if we interpret negative amounts to mean the colors would have to subtract out, then any color can be produced by any mix, and there is no such thing as "the" fundamental primaries.

We must ask whether there are three colors, i.e., some only with positive amounts for all its digits. The answer is no. Every set of three primaries requires negative amounts for some colors, and therefore there is no such as red, green, & blue a primary. In elementary books they insist it is red, green, and blue, but this is merely because with these a wider range of colors is available without minus signs for some of the coefficients.

35-4 The chromaticity diagram

Now let us discuss the combinations of colors or, in mathematical language, geometrical group addition. If any one vector is represented by Eq. (35.4), we can plot it as a vector in space by starting from the origin O , in a line a , and then a certain vector is a point. If a color vector is (σ, τ, ρ) , the color is σ red, τ green, ρ blue. In general however, as we know, the color which comes from adding these as vectors. We can simplify the diagram and reduce everything to a plane by the following observation: If we had a beam of color light, and merely multiplied it and to itself, then, if we make them all stronger in the same ratio, it is the same color, but brighter. So, if we project everything to the same light intensity, then we can project every thing into a plane and the we have there in Fig. 35-1, it follows that any color obtained by adding a given red & some pro-

¹ Except, of course, if one of the bases can be matched by mixing the other two.

position λ . This is where the mixing law "between the two points." For instance, a 50-50 mixture would appear halfway between them, and 1/3 of one and 2/3 of the other would appear 1/3 of the way from one point to the other, and so on. However, this is a *pure* color, not a mix, because the light doesn't have the same make-up; positive coefficients are needed, the active triangle, which contains "inside" of the colors that we can see, because all the colors that we can see are contained in the solid triangle bounded by the sides. Where is this zero come from? That turns out to mark a very special mixture of the colors, but we can see against it, or spectral pure. But we do not have to check all colors that we can see, we only have to check the pure spectrals. Because of the spectrum, any light can be considered as a sum of various positive amounts of various pure spectral colors, just like the physical wave point. A given light will have a certain amount of red, yellow, blue and even—greenish colors. So if we know how much of each of our three chosen primaries is needed to make each of these pure components, we can calculate the total amount of each to make our given color. So, if we add up what the color coefficients are, the spectral colors sum for any given color are my colors, then we can work out the whole color mixing rule.

An example of such experimental results for mixing three lights together is given in Fig. 33-4. This figure shows the amount of each of three different primary primaries red, green and blue, which is required to make each of the spectral colors. Notice at the left end of the spectrum, yellow is next to violet, and the two are close. Notice that at some places minus signs are necessary. It is from such data that it is possible to write the proportion of all of the colors in a chart where the x - and y -coordinates are relative to the sum of the different proportions that are used. That is the way that the curved boundary $\Delta\alpha\beta$ has been found. It is because of the pure spectral relations. Every spectral color can be made by adding spectral pure, of course, and so on, but then anything that can be produced by connecting one part of the curve to another is a color that is visible in nature. The straight line connects the end of the spectrum with the extreme mixture. This is because of the purple. Inside the triangle you can see that can be made with lights, and outside it are colors that cannot be made with lights, and nobody has ever seen them except possibly in after-images.

33-5 The mechanism of color vision

Now the next aspect of color vision, in the question, why do colors behave in this way? The simplest theory, proposed by Young and Helmholtz, suggests that in the eye there are three different regions which receive the light and that these three different regions absorb light at different places strongly, say, in the middle, either they have mostly in the blue, another absorbs in the green. And when we shine a light on them we will get different amounts of energy lost in the three regions, and these three pieces of information are somehow transmitted to the brain and in the brain it's converted to decide what the color is. It is easy to convince that all of the rules of color mixing would be a consequence of this proposition. There has been a considerable debate about the thing because the next problem, of course, is to find the share of absorption by each of the three pigments. It turns out, unfortunately, that because we can transform the color coordinates in a *very* manner we want to, we can only find *certain* kinds of linear combinations of absorption curves by the orthogonality experiments, but not the curves for the individual pigments. People have tried all various ways to obtain specific ones which also describe some particular physical properties of the eye (though not in a very satisfactory way), something like in Fig. 33-5. In this figure are two curves, one for eyes in the dark, the other for eyes in the light; and here is the cone being measured. This is measured by finding what is the smallest amount of colored light we need to add to the white to just see it. This measures how sensitive the eye is to different spectral regions. Eyes are very interesting way to measure this. If we take two colors and make them appear in unison by flickering back and forth from one to the other, we see a flicker if the frequency is too low. However, as the frequency increases,

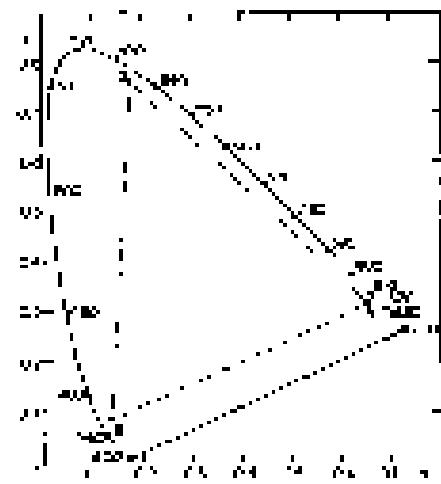


Fig. 33-4. The standard chromaticity diagram.

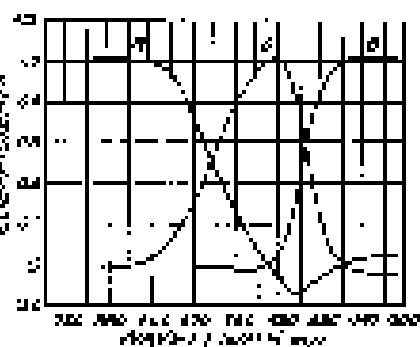


Fig. 33-5. The color coefficients of pure spectra values in terms of a cone in one of standard primary colors.

The color will change slightly depending on a certain frequency that depends on the brightness of the light. If we try to see repetitions per second, now if we adjust the brightness of the stimulus so one color cannot be seen, there comes an intensity where the flicker of the colors disappears. To see flicker with the brightness so adjusted, we have to go to a much lower frequency in order to see a flicker of the colors. So, we get what we call a flicker of the brightness at a higher frequency and, it is known that there is a flicker of the colors. It is possible to match two colors for "equal brightness" by this frequency technique. The results are almost, but not exactly, the same as those obtained by measuring the threshold sensitivity of the eye for seeing weak light by the cones. Most workers use the Flicker system as a definition of the brightness curve.

Now, if there are three other sensitive pigments in the eye, the problem is to determine the shape of the absorption spectrum of each one. Now,¹ we know there are people who are color blind—eight percent of the male population and one-half of one percent of the female population. Most of the people who are color blind are abnormal in either one has a different degree of sensitivity. There others may see some of colors but they still need three colors to match. However, there are some who are called dichromats, for whom the color can be matched by only two primary colors. These are called anomalous trichromats. They are missing one of the three pigments. If we can find three kinds of color-blind dichromats who use different sensitivity rules, one kind should be missing the red sensitive pigment, and another the blue sensitive pigment. By measuring all these types we can determine the three curves. It turns out that there are three types of dichromats and three types. There are two anomalous types and a third very rare type and from these three it is then possible to deduce the spectrum absorption spectra.

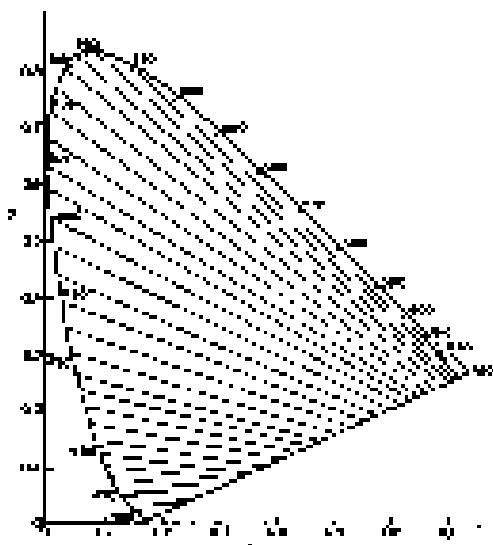


Fig. 25-6. Loss of color caused by dichromats.

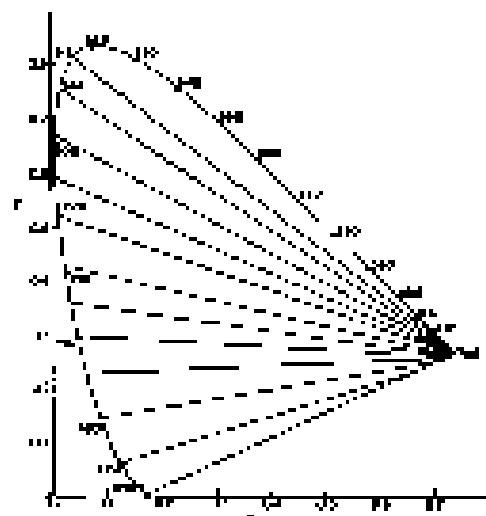


Fig. 25-7. Loss of color caused by protanopes.

Figure 25 follows the color missing of a particular type of color-blind person called a dichromat. For him, the set of curves above are not quite the certain lines, along each of which the color appears to him to be the same. If the theory that he is missing one of the three colors is confirmed, it might well be that all the lines should intersect at a point. If we carefully measure on the graph, they do intersect perfectly. Obviously, therefore, this has been made by a mathematical calculation and does not represent real data. And in fact, if we look at the lined paper with real data, it turns out this in the graph of Fig. 25-6, the point of intersection of all the curves is not exactly at the right place. Using the lines as in the above figure, we can find a reasonable spectrum; we need negative and positive absorptions in different regions. In using the new sets of curves, it turns out that each of the absorption curves is everywhere positive.

Figure 35-7 shows a different kind of color blindness, that of the protopigment, which has a peak near the red end of the incoming curve. You can just approach the same position in this case. Using the same color-matching curves of color blindness, the three-pigment response curves have fully been determined, and are shown in Fig. 35-8. *Really?* Perhaps there is a question as to whether the three-pigment idea is right. We know color blindness results from loss of one pigment, and even whether the following data are color blindesses are right. Different workers get different results. This field is still very much under development.

35-6 Physiochemistry of color vision

Now, what about checking these curves against real pigments in the eye? The pigment then can be obtained from a retina consisting mainly of a pigment cell (retinoblastoma). The most remarkable feature of this are, first, that it is the eye of almost every vertebrate animal, and second, that its response curve fits beautifully with the sensitivity of the eye, as seen in Fig. 35-9, in which we plotted on the same graph the absorption of visual purple and the sensitivity of the dark-adapted eye. This pigment is evidently the pigment that we see with it; the dark-adapted eye. This pigment is evidently the pigment that we see with it; the dark-adapted eye. This fact was discovered in 1877. From today it can be said that the other pigments of the cones have never been obtained in a live eye. In 1938 it could be said that the other pigments had never been seen at all. But since that time, two of them have been discovered by techniques by a very simple and beautiful technique.

The trouble is, presumably, that since the eye is so weakly sensitive to bright light compared with light of low intensity, it needs a lot of visual purple *as pigment*, but not much of the other pigments for seeing colors. Reddish's idea is to have the pigment in the eye, and measure it anyway. What he does is this. The eye is instrumented with an ophthalmoscope for sending light into the eye through the lens and even focusing the light that comes back out. With a photogate measure how much is reflected. So one measures the reflectance coefficient of light that has passed once through the pigment (reflected by a back layer in the eyeball, and coming out through the pigment) reflected by a back layer in the eyeball, and coming out through the pigment of the cone again. Nature is not always so beautifully designed. The cones are interestingly designed so that the light that comes into the cone bounces around and works its way down into the light-sensitive points of the eyes. The light goes right down to the sensitive point, bounces to the bottom, and comes back out again, having lost only a considerable amount of the intervening pigment; also, by lossless reflection, where the cone membrane is not exercised by visual purple. But the color of the retina has been seen a long time ago: it is sort of orangy pink; then there are all the blood vessels and the color of the muscle of the back, and so on. How do we know when we are looking at the pigment? Answer: First we take a color-blind person, who can never distinguish between it is therefore easily to make the analysis. Second, the various pigments, like visual purple, have an intensity excess when they are bleached by light; that is, when light on them they change their composition. So, while looking at the absorption spectrum of the eye, R. takes pulsed beam in the whole eye, which changes the contrast ratio of the pigment, and he measured the change in the spectrum, and the difference of course has nothing to do with the amount of blood or the color of the intervening layers, and so on. Just only the pigment, and at this measurement obtained a curve for the pigment of the photoreceptor, which is given in Fig. 35-10.

The second curve in Fig. 35-10 is actually obtained with a normal eye. It is obtained by taking a normal eye and, having already determined what one pigment was, bleaching the other one in the red where the last one is insensitive. Red light has no effect on the photopigment eye. In this is the normal eye, and thus one can obtain the curve for the missing pigment. The shape of one can fit perfectly with Yoshida's green curve, but the red curve is only fit displaced. So you know we are getting at the right track. Or perhaps not—the latest work with spectrometers does not show any definite pigment missing.

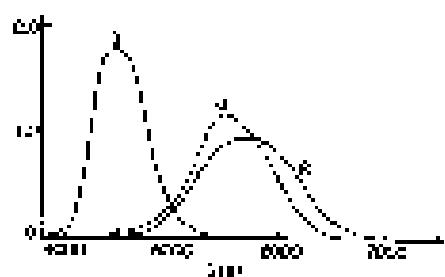


Fig. 35-8. The spectral intensity curves of three normal trichromatic receptors.



Fig. 35-9. The sensitivity curve of the dark-adapted eye compared with the absorption curve of visual purple.

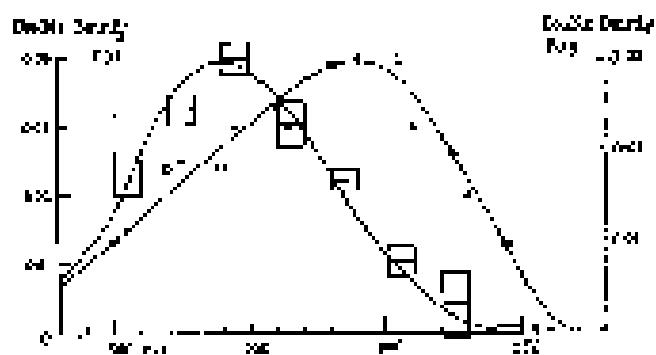


Fig. 15-10. Absorption spectrum of the color pigment of a person with anomalous trichromacy and a normal eye pigment.

Color vision, as one of the products of the life of the individual, is a common, and not unusual, function in different living organisms. For instance, if we have a pink light, made by superimposing crossing beams of white light and red light full spectrum, made with white and red incandescent lamps, we may observe that white light may appear pink. If we place an object in the beam, it casts its shadow, one illuminated by the white light alone and the other by the red. But most people see "white" shadow of an object, like blue, but, if we keep superimposing this shadow until it covers the entire screen, we see that it suddenly appears white, not blue! We can get other effects of these combinations by using red, yellow, and white light. Red, yellow, and white light can produce only orange, yellow, and so on. See if we mix red lights roughly equally, we get only orange light. Now, since, by mixing 3 different kinds of colors in the light, we have some coverage of colors, one gets quite a series of beautiful colors, which are not in the light themselves (that is only a concept, but in our sensations). We clearly see many colors and colors that are quite unlike the "physic" ones in the beam. It is very important to appreciate that a retina is already "thinking" about the light, it is comparing a color it sees in one region with what it sees in another, although not consciously. What we know of how it does this is the subject of the next chapter.

REFERENCES

- Communication Relativity, United States of America, U.S. Bureau of Census, Thomas J. Cleveland Company, New York, 1953.
- Horn, R. S., and John W. H. Pritchard, "Colour Vision and Vision," Journal of General Physiology, 1942, 25, 819-842.
- Margot, Gurney and Burt Stiles, *Physiological Psychology*, 2nd ed., McGraw-Hill Book Company, Inc., 1941.
- Nelson, N. O. and E. M. Yerkes, "Researches on Dicromatism Vision and the Special Sensitivity of the Receptors of Trichromats," presented at Symposium No. 6, First Congress of Colour, Vol. II, National Physical Laboratory, Teddington, England, September 1937. Published by Her Majesty's Stationery Office London, 1938.
- Rushton, W. A., "The Color Pigments of the Human Eye in Colour Blind and Normal," presented at Symposium No. 6, First Congress of Colour, Vol. I, National Physical Laboratory, Teddington, England, September 1937. Published by Her Majesty's Stationery Office, London, 1938.
- Watterson, R., Warren, R., *Myths and Marvels of Psychology*, Henry Holt and Company, New York, 1933. Revised edition, 1934, by Robert S. Woodward and H. Salvadore.

Mechanisms of Seeing

26-1 The sensation of color

In discussing the sense of sight, we have to realize that function of a gallery of modern art! succeeds not so much in space of colors or space of light. When we look at art, what we see is not always the thing; in other words, the artist interferes when we want. How it does this, no one knows, and it does it, of course, at a very high level. Actually we evidently learn to recognize what the man knows after much experience, this and a number of features to vision which are more clever are not which are in the combining information from different parts of what we see. To help us a diagram shows we can be an integrator of all sorts of images. It is not so difficult to study the earlier stages of the cutting together of information from the different visual cells. In the present chapter we do "experiments" with the aspect of vision, although we shall also mention a number of theories as we go along.

An example of the fact that we have an accommodation, at a very elementary level, of information from several parts of the eye at the same time, beyond the voluntary control of auditory to learn, was that the shadow which was produced by white light when both white light and were shining on the same screen. The other is, how interestingly, that the background of the screen is white enough, when we are looking at the blue shadow, i. e. only "white" light coming into a particular cell in the eye; nevertheless, years of information have been put into it. The more complete and familiar the context is, the more the eye will make corrections for peculiarities. In fact, Land has shown that if we mix that equal blue and the red in various proportions, by using two photographic transparencies with a stopper in front of the red and a slide in different proportions, it can be made to receive a real scene, with no attempt to keep it fully. This is because, a lot of information to apparent colors too, analogous to what we do, if you're mixing red and blue green; it seems to be a blue + magenta sort of color, but if we look very hard at them, they are not very good. Then so, it is surprising how much can be obtained from just red and blue. To make the white looks like a real silverman, the more one is able to see patient. In the first, total the light is actually decreasing, but think

Another example is the appearance of "colorless" like black and white rotating disc, whose black and white rings are as shown in Fig. 26-1. When the disc is rotated, the sensations of light and dark for any one cell are exactly the same; it is only the background that is different for the two kinds of "colorless". You see all the "rings" appear colored with one color, and the other with another.¹ You may yet understand the reason for these colors, but it is clear that attention is being put together at a very elementary level, in the eye itself, most likely.

Although presenting these kind of color vision, that is not mixing discs in fact. But the way you're segments in the cases of the eye and the, it is the spectral absorption in these three pigments that functionally produces the color sense. But the total sensation that is associated with the absorbed chromatic is less than the three segments acting together is not necessarily the sum of the individual sensations. We all agree that yellow and green seem to be reddishness; but that though this is somewhat surprising to most people to discover that light is, in fact, a mixture of colors, because presumably the spectrum of light

26-1 The sensation of color

- 26-2 The physiology of the eye
- 26-3 The cell walls
- 26-4 The compound (double) eye
- 26-5 Other eyes
- 26-6 Psychology of vision

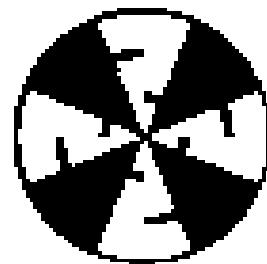


Fig. 26-1. When a disc like above is spun, colors appear in only one of the two outer "rings". If the spin direction is reversed, the colors appear in the other ring.

¹ The colors depend on speed of rotation, on the magnitude of illumination, and the size of the white bands at the central line in order to occur in them.

is due to some other process than a single neuron like a chord in muscle, where two other notes are there at the same time and if we listen hard we can hear them individually. We cannot look hard and see the red and the green.

The earlier responses of Valen said that there are three pigments and three kinds of cones, each kind containing one pigment; each a different form with cones in the brain, so that the three pieces of information are carried to the brain; and this is the brain, anything can happen. This, of course, is an incomplete idea; it does not seem to discover that the information is carried along the eye to nerve to the brain, however we have not even started to solve this problem. We must ask more basic questions: Does it make any difference where the information is put together? Is it simpler, but it is carried right up into the brain in the optic nerve, or think the retina do some analysis first? We have given a picture of the retina as an extremely complicated thing with lots of interconnections (Fig. 33-2) and it might take some analysis.

As a matter of fact, people who study anatomy and the development of the eye have known that the retina is in fact the brain; in the development of the embryo a piece of the brain comes out at birth, and long fibers grow from the retina to the brain. The retina is organized in just the way the brain is organized and, as someone has beautifully put it, "The brain has developed a way to look out upon the world." The eye is a piece of brain that is catching light, so in effect, it is already. So it is, at all unlikely that some analysis of the color has already been carried by the retina.

This gives us a very interesting opportunity. None of the other senses involves such a large amount of calculation as happens before the visual gets to the brain. One can make measurements on the calculations for all the rest of the senses mostly happen in the brain itself, which is very difficult to get at specifically, or to be measurements, because there are so many interconnections. But, with the visual sense, we have the optic nerve with lots of cells making calculations, and the results of the calculations being transmitted through the optic nerve. So we have the first chance to observe physiologically how, perhaps, the last layers of the brain work in their test steps. It is this sort of thinking which, too, simply furnishes the vision, but interesting in the whole problem of physiology.

The fact that there are three pigments does not mean that there are the three kinds of sensations. One of the other theories of color vision has it that there are really opposite color schemes (Fig. 33-3). That is, one of the nerve fibers carries a lot of impulses if there is yellow being seen, and less if it is not for that. Another nerve fiber carries green and red impulses, but in the same way, and, moreover, red to red black. In other words, in this theory someone has already started to make a connection in the system to sorting the sensations of calculation.

The problem we are trying to solve by posing in these first calculations, are questions about the sensations we report are seen on a pink background, when happens when things are adapted to different colors, and about the sorts of psychological phenomena. The psychological phenomena sort of the nature, for instance, that white does not "feel" blue and yellow and blue, and this has been well known. Because the psychologists say that there are four apparent pure colors: "There are four colors which have a nervous tendency to evoke sensations, namely, simple blue, yellow, green and red blue respectively. Unlike the magenta, purple, or mixtures of the circumlocutable colors, these simple blue are reported in the same time, these particles of the nature of the other; specifically, blue is not yellowish, and blue, in greenish, and so on; these are psychologically-primary hues." That is a psychological term, so-called. To find out from what evidence this psychological fact was deduced, we must search very hard indeed through all the literature. In the modern literature, all we find on the subject is a remark of the author of the book, or of one by a famous psychologist who was one of his professors Leonardo da Vinci, who, of course, we all know was a great artist. He says, "Even in the design there were two colors." Then, looking still further we find in a still older book, the evidence for the subject. The book systematically like this, "Purple is reddishblue, orange is reddish-yellow, but can not be seen as, purpleish-orange? And not red yellowish-green, inquiry that purple is strong?" The average person, asked to state which

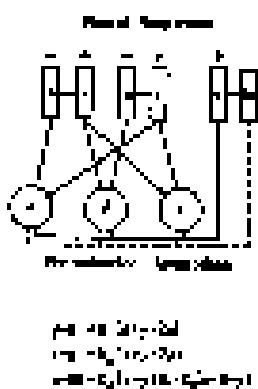


Fig. 33-2. Neural connections according to the "opponent" theory of color vision.

values are ordinary, or red, yellow, and blue, there's blue, and some others are also a faint green. Psychologists are accustomed to accept this form of "value" sense." So that is the situation in the psychological analysis of color vision; if everybody says, because three, and somebody says there are four, and they want it to be four, it will be four. That shows the difficulty with psychologists, resembles. It is clear that we're research findings, but it is very difficult to obtain more information about them.

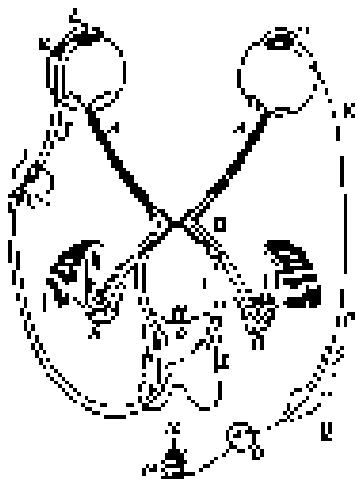
So the other direction we go is the physiological direction, to find out experimentally what actually happens in the brain, the eye, the retina, & whatever and perhaps to discover that same combinations of impulses from various cells travel along certain nerve fibers. Additionally, primary pigments do not have to be the separable cells, one can have cells in which are substances of the secondary pigments, cells with the red and the green pigments, & it's well known that in formation of colors in the white infant, long before birth. There are many ways of looking at the system so you have to find one other way before this ends. It would be no use, necessarily, that after we make out the physiological connections we will have a little bit of understanding of some of those aspects of the psychology, so we look in that direction.

36.2 The physiology of the eye

We begin by talking not only about color vision, but about vision in general, just to remind ourselves. First, the information comes in the retina, shown in Fig. 36.2. The retina is really like the surface of the brain. Although the actual picture through a microscope is a little more complicated looking than the anatomical schematic drawing, by careful analysis one can see all these interconnections. This is in practice the main part of the surface of the retina is connected to other parts, and that the information doesn't go out on the long axons, which generate the axonettes, or ramifications of it between them many cells. There are three layers of cells in the succession of crossover there are retinal cells, which are thinner than the bipolar cells, an intermediate cell which takes information from a single or a few retinal cells and gives it out again to several cells in a third layer of cells and connects it to the brain. There are all kinds of mechanisms in these cells in the layers.

We now can discuss aspects of the structure and performance of the eye (Fig. 36.3). The focusing of the light is done principally by the cornea by the fact that it has a curved surface which "bends" the light. This is why we cannot see clearly under water, because we then get a very large difference between the index of the cornea, which is 1.33, and the of water, which is 1.33. Because the cornea is water, principally with an index of 1.33, and beyond that is air, which has a very interesting structure; it is a series of layers. We can ignore except that it is all transparent and its refractive index is 1.00 in the middle and 1.00 at the outside. It would be nice if we could make optical glass in which we could adjust the "bend" throughout, but then we would not have a curve it so much as we do when we have a uniform index. Furthermore, the shape of the cornea is not that of a sphere. A spherical lens has a certain amount of spherical aberration. This aberration is "haloed" at the outside, it is a halo in in just such a manner that the spherical aberration is less far below than it would be if we put a spherical lens instead. The eye is focused by the curved lens system, onto the retina. As we focus on things that are closer and further away, the lens tightens and loosens and changes the focus to adjust for the different distances. To adjust for the total amount of light the iris may vary, which is what we call the pupil of the eye, it grows or bars, depending on what it is, as the amount of light increases and decreases, the iris moves around.

Let us now look at the neural mechanism. In controlling the accommodation of the lens, the motion of the eye, the muscles which move the eye in the socket, and the visual conduction, ideally in Fig. 36.3. Of all the information that comes out of the optic nerve at the great majority is divided into one of two bundles (which we will talk about later) and comes to the brain. But there are a few fibers, of



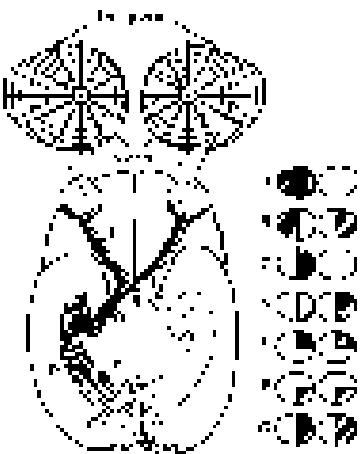


Fig. 26-4. The visual connections from the eyes to the visual cortex.

intense to us now; what do not immediately go to the visual cortex of the brain? These are the fibers which move to the storage spot and make adjustment for the images. If the image loses focus, they try to correct the lenses, if there's a double image, they try to adjust the eye for binocular vision. At any rate, they go through the pretectal area and feed back into the eye. At first the muscles which can be accommodated by the lens are not at all effective, then more is occurring. The orb has two muscle systems. One is extensor muscle, which, when it is active, pulls in and causes diverging motions very easily, and the result is it may connect from the brain through short axons into them. The opposite muscles are called "constrictors," etc., etc. The things you care about the skeletal muscle fibers, here would include you, too. There are however, in many places in the body, a pair of muscles which work in opposite directions, are in almost every spot like the nervous system which consists of two, so it is very delicately regulated that when one goes one way it is tightened, others are automatically and it follows the other. The iris is a peculiar exception, the muscle which moves the iris connects to the cones we have already described, but the nerves which make the iris expand and contract come from nearby where, go down to the spinal cord back of the chest into the thoracic sections out of the spinal nerve, up through the neck to place, and all the way up and back up into the brain in order to run the afferent of the iris. In Fig. 26-4, the optic nerve through a remarkably sufficient nervous system, not the central nervous system at all, but the sympathetic nervous system, so it is a very simple way of making things go.

We have already emphasized another strange thing about the eye, that the optic nerve comes from the wrong side, so that the right eye -- goes through several layers of other cells before it gets to the receptor -- it is quite sure and. Some of the features are very useful and some are apparently stupid.

Figure 26-5 shows the connections of the eye to the part of the brain which is most directly concerned with the visual process. The optic nerve fibers from both eyes pass just beyond D, called the lateral geniculate, whatever you like, and then to a portion of the brain called the optic cortex. Notice the fibers from each eye are sent over to the side of the brain, so the picture formed is incomplete. The optic nerves from either side of the left eye are sent to the right optic chiasm, & while the ones on the left side of the left eye come around and go off some way to the right side of the brain across all the interpeduncular commissure from the left side of the eyeball of each eye, i.e., on the right side of the right optic, while the right side of the brain sees the left side of the visual field. This is the connection in which the information from each of the two eyes is put together. In order to tell how far away things are, this is the system of binocular vision.

The connections between the retina and the visual cortex are interesting. If a spot in the retina is excited or destroyed in any way, even the whole lens will do, and we can thereby find out where it is connected. It turns out that necessarily, the connections are one to one, for each spot in the retina there is one spot in the visual cortex, and each spot and every other together in the retina are very close together in the visual cortex. So the spot which represents the spatial arrangement of the rods and cones, is, of course, much dissected. Those which are in the center of the field, which occupy a very small part of the retina, are represented over many, many cells in the visual cortex. It is clear that it is useful to know things which are originally close together, still close together. The most remarkable aspect of the matter, however, is the following. The point where one would think it would be most important to have things close together would be right in the center of the visual field. Because if or not, the spot above line is your visual field as we have it concerning a certain animal. But the information from all the points on the right side of that line is going into the left side of the brain, and vice versa. From the point on the left side is going into the right side of the brain, and the way this is made, there is a cut right down through the middle, so that the things that are very close together right in the middle are very far apart in the brain. Some of the information has to go to the outside of the brain to do this through some other standards, which is quite surprising.

The question of how the nervous system gets "wired" together is very interesting. The problem of how man is able to walk and move such a learned from childhood. It is not like I thought long ago, that perhaps it does not have to learn, since carefully it is not, it is only just properly interconnected, and then, by experience, the young child learns that what it finds is "out there" it contains some stimulus, in the left eye, it can also tell us who, the young child "knows," but how does one know what a child finds at the age of one? The child, at the age of one, supposedly sees that an object is "out there," gets a certain association, and learns to touch "there," however when he touches "there," it does not work. The approach probably is not correct, because we already see that in many cases there are these special direct interconnections. My illustration is the same that now I think requires them, with a salamander. (Incidentally, with the salamander there is a direct crossover connection, without the optic chiasm, because the eyes are on each side of the head and have no common area. Reptiles do not have binocular vision.) The experiment is this: When one optic nerve to a salamander has been cut, will give out binocular eyes again. Thousands and thousands of cell lines will thus be established. However, now, in the optic nerve, the first feature is a salamander's eye, otherwise it looks a good, slippery multi telephone cable, all the fibers twisting and turning, but when it goes to the brain they are flattened out again. When we cut the eye nerves of the salamander, the interesting question is, will it ever get right again? The answer is remarkable; yes. If we cut the optic nerve of the salamander, and it grows back, the salamander has good binocular sight again. However, if we cut the optic nerve and you see eye suddenly and let it grow back again, it is good visual acuity, of sight, but it has a terrible error, when the salamander sees a fly "up here," it appears it "down there"; and vice versa. Therefore, there is some mysterious way by which the thousands and thousands of fibers find their right place in the brain.

The problem of how much is wired in, and how much is not, is an important problem in the theory of the development of creatures. The answer is not known, but is being studied intensively.

The same eyes in man, in the case of a painful operation, there is a terrible loss like a great star or complication, in the optic nerve where we cut it, but in spite of it, it is the fine axons back to vision again, placed in the brain.

Let's consider this, we may grow into the odd channels of the optic nerve they must make several decisions about the direction in which they should grow. How do they do this? There are in the brain channels that different fibers respond to differently. Pairs of the common nerves of growing fibers, each of which is an individual, I believe, in some way, has its neighbors, in responding to whatever the chemical stimulus, it depends in a unique, unambiguous way to find its proper place for ultimate connection in the brain. This is an interesting—a framework—of some of the quite recently discovered phenomena of biology, and is undoubtedly connected to many other biological problems of growth, organization, and development of organisms, and particularly of embryos.

One of the interesting phenomena is to do with the motion of the eye. The system of the eyes in order to be the two images coincide in different circumstances. These motions are of different kinds: one is to follow something, which requires the left eye to move in one particular direction, right or left, and the other is to swing them toward the same place at varying distances away, which requires that they move, and oppositely. The nerves going to the muscles of the eye are already wound up for just such purposes. There is a set of nerves which will pull the muscles on the inside of one eye and the outside of the other—and others, the opposite muscles, so that the two eyes move together. There is another center where one will know to set the eyes around in a coordinate-like form parallel. Both eyes can be turned out to the exterior, the other eye never leaves the nose, but it is a peculiarly remarkable to move slightly to turn both eyes out at the same time, not however from one to another, but has to be binocularly, so as to align to turn both eyes out, unless we have had an accident or there is something the matter, for instance, if a nerve has been cut. Although the muscles of one eye can certainly steer that eye about, not over, a Yogi is able to move both eyes out freely

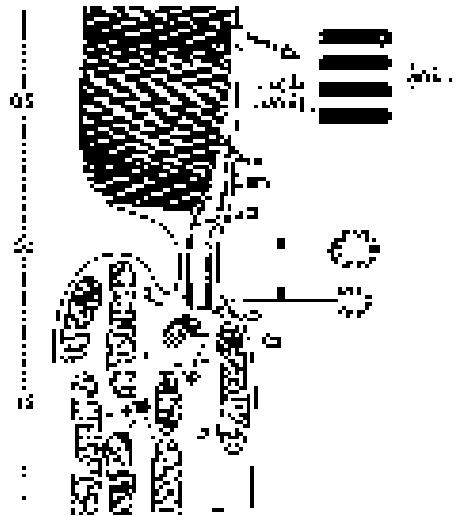


Fig. 26-5. Electron micrograph of a rod cell.

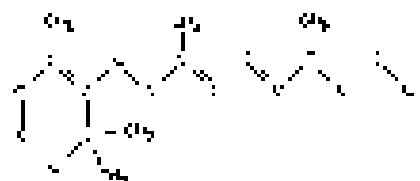


Fig. 26-6. The structure of retinene.

makes it difficult to see, so there does not seem to be any way to do it. We are almost given to a certain point, because most of the earlier books on vision, *y* and conductivity, and so on, the eye apparatus or do not emphasize the fact that we are so amazingly well prepared—they say that everything is just perfect.

26-3. The rod cells

Let us now examine in more detail what happens in the rod cells. Figure 26-5 shows an electron micrograph of the nucleus of a rod cell (the rod cell sleeps going up out of the field). There are layers after layers of plate-like structures, shown here, tilted at the right, which contain the substance rhodopsin (rhodo purple), the eye pigment which produces the effects of vision in the dark. The rhodopsin, which is the pigment, is a big protein which contains a special greenish-yellow retinene, which can be taken out the protein, and which is undoubtedly, the main reason of the absorption of light. We can not measure all the energy that is absorbed, but it is very likely that the best reason for getting all the rhodopsin molecules excited. The chemistry of the retina has been worked out to a large extent, but there might be some physics to it. It may be that it will be measured very accurately in some day of a new century when others measure the extent which is generated, yes, may not all the way down to one-tenth of a volt if it is designed, or something like that. The voltage is very important, and having homogeneous electric fields throughout both bioceramics and solid state physics or something like it, will ultimately be used.

This kind of a structure with layers appears in other circumstances where light is important, for example in higher-order problems, where the light causes photosynthesis. If we magnify those, we find the same thing with almost the same kind of layers, but there we have of "roughly" of course instead of smooth. The chemical form of rhodopsin is shown in Fig. 26-6. It has a series of alternating double bonds along the side chain, which is characteristic of such all strongly absorbing organic substances. I recall, methyl, Ethyl, Propyl, Butyl, and so on. These kinds of acids are impossible for human hair to absorb in their complete—*we have to eat it*. So we eat all the fat of a special substance, which is exactly the same as retinene except that there is a hydrogen tie on the right end; *so called dimethyl*, and if we do not eat enough of it, we do not get a supply of retinene, and the eye becomes blind, we call *night blindness*, because there is *not* enough pigment in the rhodopsin to eat with the rods at night.

The reason why such a series of double bonds absorbs light very strongly is also simple. We have just given a bit: The alternating series of double bonds is called conjugated, which means that there is an extra electron there, and this extra electron is easily shifted to the right or left. When light strikes the electron of each double bond it is shifted over by one step. All the electrons in the whole chain shift, just a single electron moving over, and though each one moves only a little distance we would expect that, in a single atom, we would move the electron only a little distance. The effect is the same as though the right end was moved over to the other end! It is the same as though one electron were to go from the left end to the right end, in this manner, we get much stronger absorption under the influence of the electric field, than if we move only *one* electron a distance which is associated with one atom. So, since it is easy to move the electron back and forth, it must absorb light very strongly; that is the machinery of the physical chemistry of our eye.

26-4. The compound (insect) eye

Let us now return to biology. The human eye is not the only kind of eye. In the invertebrates, almost all eyes are essentially like the human eye. However, in the lower animals there are many other kinds of eyes, especially various eye spots, and other less sensitive things which we have not mentioned yet. But there is one often highly developed eye among the invertebrates, the compound eye of the insect. Most insects living on the ground eyes also have various additional similar eyes usually. A fly's eye is one which has been studied very carefully. It is

only very slightly the properties of vision of bees because they are unadapted to beauty, and we can make experiments in which we identify the honey by putting it on blue paper or red paper, and see which one they come to. By this method some very interesting things have been discovered about the vision of the bee.

In my first glass, in trying to measure how acutely bees can see the violet light area between two pieces of "white" paper, and it was found that they were not very good, and others found they were considerably good. Even if the two pieces of white paper were almost exactly the same, the bees could still tell the difference. The experimenters used true white for one piece of paper and had others for the other, and although these look exactly the same to us, the bee could easily distinguish them, because they reflect a different amount of the ultraviolet. In this way it was discovered that the bee's eye is sensitive over a wide range of the spectrum. Far is it from 3000 Angstroms to 4300 Angstroms, from red to violet, but the bee can see down to 2800 Angstroms into the ultraviolet. This is also the number of different interesting effects. In the first place, bees can distinguish between many flowers which we look alike. Of course, we must realize that the colors of flowers are not designed for our eyes but for the bees' eyes and they train the bees to a specific flower. We all know that there are many white flowers. And only white is not very interesting to the bees, because it turns out that all of the white flowers have different proportions of reflection in the ultraviolet. They do not reflect one hundred percent at the ultraviolet as would a true white. If the light is not coming back, the ultraviolet is missing and that is a color, just as, for me, if the sun is missing, it comes out yellow. So, if the flowers are colored for the bees. However, we also know that red cannot be seen by bees. This might surprise you, but it does not look how to the bee. Instead, a careful study of red flowers shows, first, that even with our own eyes we can see the red, because red flowers have a little edge because they are actually reflecting ultraviolet spectrum in the blue, which is the part that the bees see. Furthermore, experiments also show that flowers vary in their reflection of the ultraviolet and different parts of the public variation. So if we put two flowers together they would be quite more attractive than the full

It has been shown, however, that there are a few red flowers which do not reflect in the blue or in the ultraviolet, and would therefore appear black to the bee. This was of course not known to the people who worry about this matter, because those eyes are not like ours, so they do not know what is hard to tell from a early childhood. It actually turns out that these flowers were chosen by bees, those are the flowers that we should be growing today, and summing up, can say now?

Another interesting aspect of the vision of the bee is that bees can apparently tell the direction of the sun by looking at a patch of bright sky without using the sun itself. We cannot easily do this. If we look outside windows at the sky and see that it is blue, in which the sun is the sun.[†] The bee can tell, because the bee is quite sensitive to the polarization of light, and the general light of the sky is polarized.[‡] There is a million calculations how this sensitivity operates. Whether it is because the receptors of the eye are different in different circumstances, or the bee's eye is actually sensitive, is not yet known.[†]

If you look at the television screen flicker up to 300 oscillations per second, while we see it only up to 20. The motion of bees in the house is very quick, the bee moves with the wind, etc., but it is very hard for us to see these motions with our eye; however, if we could see more rapidly, we would be able to see the motion. It is probably very important to the bee that its eye has such a rapid response.

[†] The wings of the bee have a slight sensitivity to the polarization of light, and this will help to tell the position of the sun. This phenomenon can be described best as called Ruffing's effect, which is a form of break among certain patterns due to the state of the visual field when one looks at a fixed, featureless surface using polarizing glasses. It can also occur in the blue sky when wearing glasses that reduce the field of view and affect the visual system.

[‡] Professor Sherratt's work, which has been given indicates that the eye is directly sensitive.



Fig. 36-7. The structure of an ommatidium (a single "eye" of a compound eye).

Now let us discuss the visual acuity we could expect from the bee. The eye of a bee is a compound eye, and it is made of a large number of special cells called *ommatidia*, which are arranged radially on the surface of a sphere (except for) on the outside of the bee's head. Figure 36-7 shows a picture of one such ommatidium at the top. This is a longitudinal section kind of thing. But actually it is more like a filter or light pipe or something going down along the "anterior fiber," which is where the absorption presumably occurs. Out of the center cell it comes the *optic fiber*. The central fiber is surrounded on its sides by six cells which, in fact, have absorbed the light. That is enough explanation for our purposes; the point is that it is a vertical ring of cells, each next to each other all over the surface of the eye of the bee.

Now let us discuss the resolution of the eye of the bee. If you draw lines (Fig. 36-8) from every ommatidium on the surface, which we suppose is a sphere of radius r , the bee actually receives here with each ommatidium a very small bit of air, and assuming that resolution is as clever as we are! If we have a very large assumption, we can see how much resolution. That is, one cell gets a piece of information from one direction, and the next cell gets a piece of information from another direction, and so on, and the bee cannot see things from too many well, but the angle subtended by each unit in the eye will surely correspond to an angle, the angle of the end of the ommatidium — relative to the center of curvature of the eye. (The eye cells, of course, exist only at the surface of the sphere; incident light is the bead of rays etc.) This angle, from our assumption to the best, is, of course, the diameter of the cone-like device by the ratio of $\lambda/\pi r^2$.

$$\theta_0 = \frac{\lambda}{\pi r^2}.$$

(36.1)

So, we may say, "The lines we make the θ_0 we meet the visual reality. So why doesn't the bee just use very very fine ommatidia?" Answer. We know enough physics to realize that if we are trying to get θ_0 down into a cone like that, we cannot do it really in a given direction because of the diffraction effect. The light that comes from every cone has an interference due to diffraction, **so will** get light coming in at angle $\Delta\theta$ such that

$$\Delta\theta = \lambda/r.$$

(36.2)

Now we see the "If we make the θ_0 too small, then each ommatidium does not look in only one direction, because of diffraction!" If we make them too big, each one looks in a different direction, but there are not enough of them to get a good view of the scene. So we find the best resolution to make minimum the total angle of these two. If we add these two together, and find the place where the sum has a minimum (Fig. 36-9), we find that

$$\frac{\Delta\theta_0 + \Delta\theta_D}{r} = 0 = \frac{1}{r} - \frac{\lambda}{r^2},$$

which gives us a distance

$$r = \sqrt{\lambda r}.$$

(36.3)

If we know that r is about 7 millimeters, since the light that the bee sees is 1000 wavelengths, we put the two together and take the square root, we find

$$r = (7 \times 10^{-3} \times 1 \times 10^7)^{1/2} \text{ m} \\ 7.0 \times 10^{-2} \text{ m} = 7.0 \mu.$$

(36.4)

The book says the diameter of the eye of the bee is about 1 mm, so, apparently it really works, and we can understand what determines the size of the bee's eye. It is also easy to you can observe that hawk which had 20 times greater the visual acuity is 100 angstrom resolution; it is very poor relative to man, who. We can see things the same thirty times smaller in apparent size, i.e., how the bee has a rather sharp out-of-focus image; this is what we call *see*. Nevertheless it is still right, and it is the best they can do. The important why the bees do not develop

be like eye like our own, with a lens and so on. There are several interesting reasons. In the first place, the bee is very small; if it had our eye like ours, but in 10⁻³ scale, the eye-ball would be about 20 cm in size and ditto the eye would be as important there as would not be able to see very well anyway. The eye is not good if it is so small. Secondly, if it were as big as he was itself, then the eye would occupy the whole body of the bee. The beauty of the compound eye then is that up to species, it is just a very thin layer on the surface of the bee. & when we argue that they didn't have done it our way, we need remember the "They had their own systems."

34-5. Other eyes

Besides the bees, many other animals can see color. Fish, butterflies, birds, and insects can see color, but it is believed that the mammals cannot. The primates can see color. The birds certainly see color, and that means a lot for the colors in birds. There would be no point in having such ordinary colored males if the females could not notice it. That is, the evolution of the sexual "show-off" is that the body color is a result of the female being able to see males. So, for example we have all these peacock feathers, think of what a brilliant display of colors it is and how beautiful the colors are, and what a wonderful aesthetic sense it takes to appreciate all that. But, we should not compliment the peacock, but should compliment the vision, sensory, and aesthetic sense of the peacock, because that is what has generated the beautiful scene.

All vertebrates have poorly developed eyes or compound eyes, but all the vertebrates have eyes very similar to man's eyes, with one exception. If we consider the higher form of animal, we usually say "the eye is not," but if we take a less prejudicial point of view and reflect causatives to the invertebrates, we find we cannot ignore insects, and ask what is the highest invertebrate animal? Most zoologists agree that the octopus is the highest in animal. It is very interesting then, besides the development of the brain and its reactions and so on, which is certainly good for an invertebrate, it has a well-developed, independent, a different eye. It is not a compound eye or an eye spot—it has a retina, it has lens, it has iris, it has pupil, it has a region of walls, it has a lens capsule. It is extremely like ours as the eye of the vertebrates. It is a remarkable example of a coincidence in evolution where we have twice discovered the same solution to a problem, with one slight improvement. In the octopus it is larger, and, consequently, that the retina is a piece of the brain that has come out in the same way in its embryo, it develops right from the start very early on, but the interesting thing which is different is the cells which are sensitive to light are on the inside and the cells which do the processing are on the back of them, rather than "face on," as in our eye. So we see, perhaps, that there is no general reason for us being upside out, the other invertebrates much like us are straightforward (Fig. 34-10). The largest eyes in the world are those of the giant squid. They have been found to be 8 inches in dia. (80 cm).

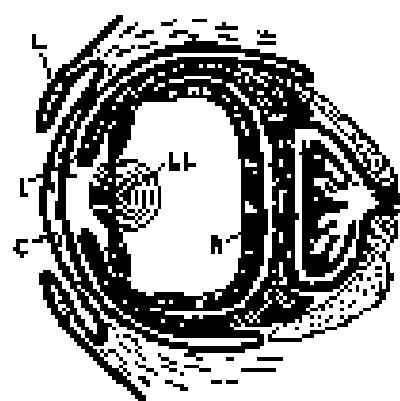


Fig. 34-10. The eye of an octopus.

34-6. Neurology of vision

One of the main points of our subject is the transmission of information from one part of the eye to the other. Let us consider the compound eye of the housefly, on which considerable experimental work has been done. First of all, we must appreciate what kind of information it conveys to consciousness. As we know, it is a kind of disturbance which has an electrical effect, that is easy to detect, a kind of wave-like vibration that which can flow, the nerve and processes can affect at the other end: a long piece of the nerve cell, called the axon, carries the message. You always, as I have said, kind of surprise, called a "spike," you know if it is excited or not and when one spike goes along the nerve, another comes immediately follow. All the spikes are of the same size, so it is not that we get bigger spikes when the thing is more strongly excited, but that we get more spikes per second. The sort of energy is determined by the "PSP". It is important to understand this in order to see what happens next.



Fig. 36-11. The compound eye of the horseshoe crab. (a) Normal eye. (b) Crushed eye.

(from Dr. G. H. D. T. Gray, Duke University Marine Laboratory, Beaufort, North Carolina, with permission.)

Figure 36-11(a) shows the compound eye of the horseshoe crab; it is just as much of an eye as has only skin as transverse commissure. Figure 36-11(b) is a crushed compound eye system; one can see the commissure, which however does not run across them and go into the brain. But note that even in a horseshoe crab there are little transverse lines. They are much less elaborate than in the human eye, and it is worth while to study a simpler example.

Let us now look at the experiments which have been done by cutting the electrodes in the simple nerves of the horseshoe crab and stimulating one or more of the commissures, which is easy to do with Lucas. If we turn a light on, it stimulates α_2 , and because the other ion channels will come out, we find that there is a spike, followed by three or four series of discharges which gradually slow down to a uniform rate, as shown in Fig. 36-12(a). When the light goes off, the discharge stops. Now, it is very interesting that if, while our stimulus is still on in the same nerve fiber we shine light on a different commissure nothing happens; no signal.

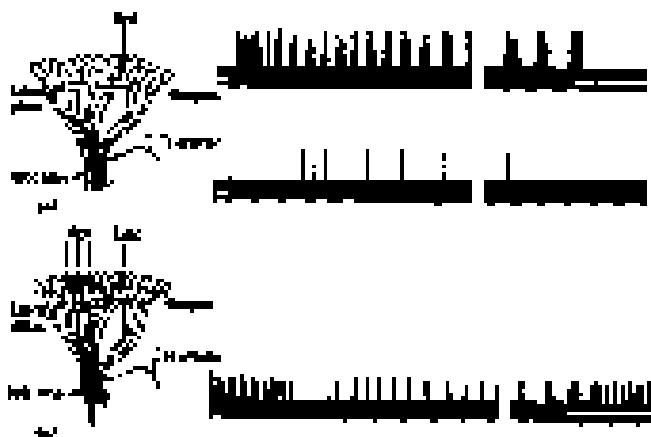


Fig. 36-12. The response in light of the nerve fibers of the eye of the horseshoe crab.

Now we do another experiment, as shown in the figure on the original commissure and get the same response, but if we now turn light on and the time as they as well, the pulses are in different spikes and then come a much slower rate (Fig. 36-12b). The case of rest is inhibited by the impulses which are coming out of the other. In some ways each nerve fiber carries the information from one commissure, but is stronger than it is inhibited by the signals from the other. So, for example, if the whole eye is uniformly or nonuniformly illuminated, the inhibition coming from any one commissure will be relatively weak because it is balanced by so many. In fact the inhibition is enhanced if we shine light on several nearby commissures the inhibition is very great. The inhibition is weaker when you stimulate one fiber and if the stimulated fiber is enough away from one another, inhibition is practically zero. So it is not dependent on the distance, there is a first example of inhibition from different parts of the existing commissure in the eye itself. We can see, perhaps, that link also in a vertebrate, that has to do due to inhibition occurring at the edges of retinula, because if part of the retina is light and a part is black, then the inhibition in the liquid and you impulsive nature

inhibited by all the other lights in the neighborhood, so it is relatively weak. On the other hand, an *occluding* edge, the boundary which is given a "dark" impulse is also inhibited by others in the neighborhood, but there are not as many of them, since some are exactly in the right place, so it therefore strengthens. The result would be a curve something like that of Fig. 26-11. The crab will see an enhancement of the concave.

The fact that there is an enhanced *out-of-concave* boundary has been known; it fact it is a remarkable thing that has been commented on by psychologists many times. In order to view an object, we have only to close our eyes. How does one get an enhanced picture than is merely the outline? What is the answer? The solution is only the edge difference between light and dark in one corner and another. It is not even *strongly* defined. In other words, it is not that every object has one arcuate in. There is no such line. It is only in our own psychological make-up that we do this; we are beginning to understand the reason why the "line" is enough of a cue to go the whole thing. Presumably our own eye works in a similar manner, much more complicated, but similar.

Finally, we can briefly describe one more elaborate work, the beautiful, elaborate work that has been done on the frog. Doing corresponding experiments on a frog, by pulling very fine, beautifully-woven woodlike strips over the eye, one can obtain the signals, but > a living thing, one will pull over and over again, just as in the case of the horseshoe crab, we find that the information does not depend on just one spot in the eye, but is a sum of information over several spots.

The most recent paper¹ of the operation of the frog's eye is the following. One can find four different kinds of optic nerve fibers, in the sense that there are four different kinds of responses. These experiments were undertaken by Stimberg and *et al.* in 1958 of the frog, because that is not *that* a frog does. A frog just sits there and his eyes never move, unless the fly just flies around, and he turns, and in that case his eyes will like just right so that the image stays just. He uses *not* both his eyes, if anything moves in his field of vision, like a little bug flies by, while in the something small moving in the field backs, needs to turn so that there are four of the kinds of fibers which carry the visual properties are summarized at Table 26-1. Sustained edge detection... interesting, means that if we bring an object with a *edge* into the field of view of the frog, then there are a lot of impulses in this particular fiber while the object is moving in. If you suddenly let a *sharp* edge impulse just experience suddenly as the edge is there, even if it is standing still, then you can see the light, the impulses stop. If we turn it on again, while the edge is still in view, they start again. They are *reversible*. Another kind of fiber is very similar, except that if the edge is straight, it does not work. It may be convex only with its back side. How conspicuous must be the edge and how small is the number of the edges the frog is forced to pay attention to in order to notice what has moved in! Furthermore, if through this fiber does sustain temporal, if there are sustained impulses along the center, and if we turn the fiber around, it immediately goes back up again. It depends on the *rocking* of the nerve.

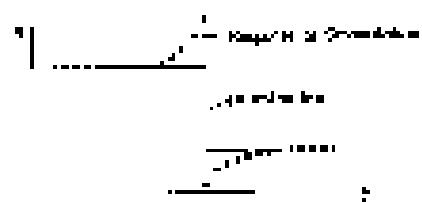


Fig. 26-10. The net response of an omniotic crab omniotic eye to three concaves in Horseshoe.

Table 26-1
Types of responses to optic nerve fibers of a frog

Type	Size	Angular field
1. Steepest edge detector (concave)	0.1 mm radius	1°
2. Convex edge detector (convex)	0.5 mm	2-3°
3. Convex detector	1.2 mm	10-15°
4. Dynamic detection	Up to 2 m, sec.	Up to 10°
5. Vertical detector	1	Very large

surface. The eye sees it come in and announces that it is there, but if we merely turn out the light for a moment, it simply forgets it and no longer sees it.

Another example is change-in-intensity detection. If there is an ugly moving it, we feel there are pulses, but if the thing stands still there are no pulses at all.

Then there is a third detection. If the light intensity is uniformly decreased it creates pulses, but if it goes down or goes up, the response depends on the source where the pulse is coming.

Then, finally, there comes the third which is dark detection. A most amazing thing they are all the same! If we increase the light, they fire less rapidly than all the time. If we decrease the light, they fire more rapidly than ever. In the dark they are like mad, perpetually saying "It is dark! I is fired! I is it!"

Now these responses seem to be rather complicated to classify, and we might wonder whether perhaps the sensations are better distinguished. But it is very interesting that these same classes are very clearly separated in the anatomy of the frog. By this measurement, after these responses had been classified by observation, the next step is to about 1000. It was discovered that the visual system of the different fibers there was not too simple, so there was another, independent way to check what kind of a fiber we have found.

Another interesting question is from how big an area is one particular fiber making its contributions? The answer is different for the different classes.

Figure 24-14 shows the surface of the so-called optic tectum of a frog, where the nerves come into the brain from the eye is shown. All the nerve fibers coming in from the optic nerve make numerous anastomosing layers of the tectum. This typical structure is analogous to the retina. To large, by which we know that the brain and retina are very similar. Now, by taking an electron microscope it does not succeed in finding the fibers, we can find out which kinds of optic nerves end where. And the beautiful and remarkable result is that the different kinds of fibers end at different layers. The first ones end in number 1 type, the second in number 2. The third and fourth end in the same place, and deeper of all is number four (which is remarkable, they put the number eleven in the right place! No, that is not true, they numbered them that way, and the paper had the numbers in a different order!)

We may easily summarize what we have just learned this way. There are three types, presumably. There may be many different kinds of receptor cells containing the three pigments in different proportions, but these are ready cross connections which may permit utilization of some colors in much detail and discrimination in the nervous system. So before we really understand color vision, we will have to understand the final connection. This subject is still to be done, but these researches with electron microscopy and so on will perhaps ultimately give us more information which we see each.



Fig. 24-14. The tectum of a frog.

REFERENCES

- Deutsche, C. C., Color Theory, Optical Society of America, The Society of Color, Thomas Y. Crowell Company, New York, 1931.
"Inhibition of Vision," 2nd Supplement to *Journal of General Physiology*, Vol. 41, No. 6, Part 2, July 1958, Rockefeller Institute Press.
Eccles, J. C., *Neurophysiology*, Methuen & Co., London, 1957.
Goldschmid, E., "Some Observations on the Unconscious and Morphogenesis of Photophoresis," p. 1-16.
Lewin, L. M. and D. J. Lewin, "Perceived Color, Individual Effect, and Optokinetic Response Mechanism," p. 1143.
Rogers, W. A. et al., *Science Communication*, Massachusetts Institute of Technology Press, Cambridge, Mass., 1932.
"Sight, Sense of," *Encyclopedia Britannica*, Vol. 18, 1951, pp. 623-624.

Quantum Mechanics

37-1 Atomic mechanics

In the last few chapters we have treated the essential ideas necessary for an understanding of most of the important physical and optical effects in electromagnetism in general. We have left some specific topics for later years. Specifically, the role of the idea of wave materials had been somewhat neglected. What we now do with it is called the "kinetic theory" of electric waves, which turns out to be a extremely elegant description of motion for a very number of effects. We have not had to worry so about the fact that light energy comes in lumps or "photon."

We would like to take up as our first subject the problem of the behavior of matter in very small pieces of matter, both mechanical and thermal properties, for instance. In studying these, we will find that the "kinetic" (or "classical") theory of matter is inadequate, because matter is really made up of wave-like particles. So, you will deal only with "kinetic" particle theory, i.e., in the style of R.F. Landau, we can understand using the classical mechanics we have been learning. But, we can't make very successful. We shall find out in the case of matter, unlike the case of light, we shall have difficulties, definitely even. We could, of course, even entirely skirt away from the atomic effects, but we shall instead face here a short discussion of what we will describe the basic ideas of the quantum properties of matter, i.e., the quantum theory of atomic physics, so that you will have some feeling for what it is we are leaving out. Let us will now review and sum up important subjects that we are not using, coming down to.

So, we will give now the representative to be object of quantum mechanics, but we will be able actually to get into the subject much much later.

"Quantum mechanics" is the description of the behavior of matter in all its detail and, in particular, of the happenings at the atomic scale. This is not *easy*, we believe. No, nothing that you can see directly exceeds above. They do not behave like waves, they do not behave like particles. They do not be like clouds, or billiard balls, or weights on springs, or like anything that you have ever seen.

Everyone thought that light was a corpuscular effect, but that was discovered, as we have seen here, that it behaves like a wave. Later, however, (at the beginning of the twentieth century), we found that light did indeed sometimes behave like a particle. Historically, the electron, for example, was thought to behave like a particle and then it was found that it more closely behaved like a wave. So, clearly behaved like neither. Now we have given up. We say, "It is the *activity*."

This is not only better, however, electrons behave just like light. The system exhibits a "wave effect" when one passes between screens, and when it is too coarse-grained, they are a "particle effect," or vice versa you will not call them. So, while we can, abstractly speaking, electrons (which we shall do for our example) will apply to both "particles," including photons of light.

The general conclusion of information that almost all one could be having about the fine quarks of this creature, which gave some indications about how *tiny* things we cannot produced in the using equipment which was finally developed in 1926 and 1927 by Schrödinger, Dirac, Heisenberg and Pauli. They finally obtained a description description of the behavior of matter on a small scale. We take up the main features of their derivation in this chapter.

Because atomic behavior is so unlike ordinary experience, it is very difficult to get used to and it appears paradoxical and mysterious to every new book to the

37-1 Atomic mechanics

37-2 An experiment with bullets

37-3 An experiment with waves

37-4 An experiment with electrons

37-5 The importance of electrons

37-6 Wavelike the electrons

37-7 First principles of quantum mechanics

37-8 The uncertainty principle

music and to the experimental physicist. But the experts do not understand it in any way they would like to, and it is probably conceivable that they wouldn't because all of our human experience and of human intuition applies to large objects. We know how large objects will act but things on a small scale just do not act that way. So we have to learn about them in a sort of abstract or mathematical fashion & don't have connection with our direct experience.

In the chapter we shall locate immediately the basic element of the mysterious behavior in its most strange form. We shall see a situation of polarization which is impossible physically impossible to explain in any classical way, and which lies in the heart of quantum mechanics. In reality, it contains the only mystery. We cannot explain the mystery in the name of "probability" how it works. We will tell you how it works. In telling you how it works we will not tell you about the probabilities of all quantum mechanics.

37-1 An experiment with bullets

To try to understand the quantum behavior of electrons, we do something similar and contrast their behavior at a particular experiments setup, with the more familiar behavior of particles like bullets. We want the behavior of bullets like the other waves. We consider first the behavior of bullets in a experiments setup shown diagrammatically in Fig. 37-1. We have a machine gun that shoots a stream of bullets. It is not a very good gun, in that it sprays the bullets randomly over a fairly large angular spread, as indicated in the figure. In front of the gun we have a wall (made of dense, plated lead bricks) in our house just about big enough to let a bullet through. Beyond the wall is a backstop made of thick wall which will "absorb" the bullet when they hit it. In front of these we have an object which we shall call a "detector" of bullets. It might be a box containing sand. And, finally, that's the detector will be stopped and accumulated. When we wish we can empty the box and count the number of bullets that have been caught. The detector can be moved back and forth so when we will collect a direction. With this apparatus we can find out experimentally the answer to the question "What is the probability that a bullet which passes through the hole in the wall will arrive at the backstop at the distance x from the center?" First, you should realize that we should talk about probability, because we cannot say definitely where any particular bullet will go. A bullet which happens to hit one of the holes may become at the edges of the hole, red over and over again in a wall. By "probability" we mean the chance that the bullet will arrive at the distance x , which we can measure by counting the number which arrive at the detector at a certain time and then taking the ratio of the number to the maximum that will hit the backstop during that time. Or, if we assume that the gun always shoots at the same rate during the measurement so the probability we want is just proportional to the number of bullets which the detector is some standard time interval.

For our present purposes we would like to imagine a somewhat idealized experiment in which the bullet are not real bullets, but are mathematical bullets they cannot break in half. In our experiment we find that bullets always come in jumps, and when we find something in the detector, it is always one whole bullet. At the instant which the machine gun fires a bullet, say how, we find that any given moment either nothing arrives at our sand box, or exactly one—exactly one—bullet arrives at the backstop. Then, the size of the jump certainly does not depend on the rate of firing of the gun. We shall say: "Bullets also wear their identical jumps." What we mean is that our detector is the probability of arrival of a jump. And we measure the probability as a function of x . The result of such measurements with the apparatus for have not yet done the experiment, we are really imagining, is seen in Fig. 37-2 and plotted in the graph shown a part of Fig. 37-3. In the graph we plot the probability in the right and x vertically, so that the x scale fits the diagram of Fig. 37-1. We call the probability P_x because the bullets may pass completely through hole 1 or through hole 2. You will not be surprised that P_1 is large near the middle of the graph and gets small if x is very large. You may wonder, however, why P_2 has its minimum value at $x = 0$. We can understand

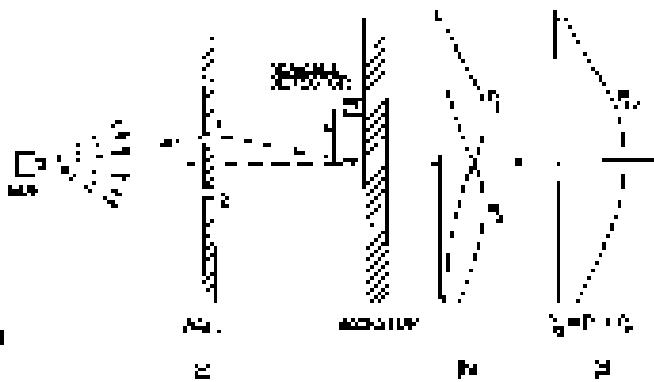


Fig. 27-1. Interference experiment with bullets

to see if we do not expect to get after covering up hole 1, and once more while covering up hole 2. When hole 2 is covered, bullets can pass only through hole 1, and we get the curve marked P_1 in part (a) of the figure. As you would expect, the maximum of P_1 occurs at the value of x which is on a straight line with the gun and hole 1. When hole 1 is closed, we get the symmetric curve P_2 drawn in the figure. P_2 is the probability distribution for bullet positions through hole 2. Comparing part (a) and (c) of Fig. 27-1, we find no surprise there:

$$P_{12} = P_1 + P_2 \quad (27-1)$$

The probabilities just add together. One bullet will be found either in the sum of the paths with each hole open alone. We shall call this result an instance of "no interference" for a reason that you will see later. So much for bullets. They come in lumps, and their probability of arrival does not interfere.

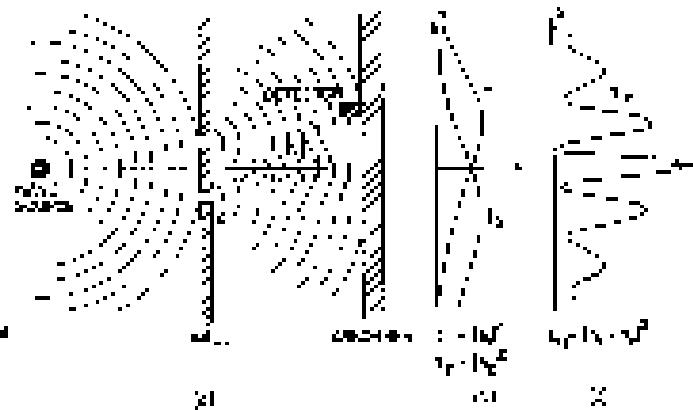


Fig. 27-2. Interference experiment with water waves

Now we wish to consider an experiment with water waves. The apparatus is shown diagrammatically in Fig. 27-2. We have a shallow trough of water. A small object labeled the "wave source" is jiggled up and down by a motor and makes circular waves. To the right of the source we have again a wall, with two holes, and beyond that is a second wall, which, to keep things simple, is an "absorber," so that there is no reflection of the waves that arrive there. This can be done by putting a padded board "back" in front of the second wall. The detector is now a device which measures the "Intensity" of the wave motion. You can imagine a gauge which measures the height of the wave motion, but whose scale is calibrated in proportion to the square of the wave height, so that the reading is proportional to the intensity of the wave. Our detector reads, then, a proportionality to the energy being carried by the waves—*i.e.* either the rate at which energy is carried to the detector.

With our wave apparatus the first thing to notice is that we *excessively* do not have any size. If the gun just makes a very small sound, then there is just a little bit of wave motion at the detector. When there is more twanging, however,

there is some intensity at the detector. The intensity of the wave can have any value at all. We would not say that there was one "component" in the wave, unless it is.

Now let us measure the wave intensity for various values of θ (keeping the wave source operating always in the same way). We get the interesting looking curve recorded in the portion of the figure.

We have already worked out how such patterns can come about; when we studied the interference of electric waves. In this case we would observe that the original wave is diffracted at the holes, and two similar waves expand out from each hole. If we cover one hole at a time and measure the intensity distribution at the detector we find the rather simple intensity curves shown in part (c) of the figure. I_1 is the intensity of the wave from hole 1 (which we find by measuring when hole 2 is "blocked" or open). I_2 is the intensity of the wave from hole 2 (seed when hole 1 is blocked).

The intensity I_3 observed when both holes are open is obviously not the sum of I_1 and I_2 . We say that there is "interference" of the two waves. At some places where the curve I_3 has no maximum the waves are "in phase" and the wave peaks add together to give a large amplitude, therefore a large intensity. We say that the two waves are "constructively interfere" in such places. There will be such constructive interference whenever the distance from the detector to one hole is a whole number of wavelengths larger than the distance from the other hole to the detector.

At those places where the two waves interfere destructively, the distance of θ (where they are "out of phase") just regulating wave motion at the detector will be the difference of the two amplitudes. The waves "interfere destructively," and we get a low value for the wave intensity. We repeat our law values whenever the distance between hole 1 and the detector is different from the distance between hole 2 and the detector by an even number of half-wavelengths. The low values of I_3 in Fig. 32.2 are, according to this principle, due to destructive interference.

You will notice that the square law relationship between I_1 , I_2 , and I_3 can be expressed in the following way: The instantaneous height of the water wave at the detector for the wave from hole 1 can be written as $(\delta_1 + \delta_2)$, where the "amplitude" δ_1 is, in general, a complex number. The intensity is proportional to the mean squared height, when we use the complex numbers as $|\delta_1|^2$. Similarly, for hole 2 the height is δ_2 and the intensity is proportional to $|\delta_2|^2$. When both holes are open, the wave heights add to give the height $(\delta_1 + \delta_2)^2$ and the intensity $|I_3| = |\delta_3|^2$. On this the equations of proportionality for all present purposes can serve as follows for anything waves are:

$$I_1 = |\delta_1|^2, \quad I_2 = |\delta_2|^2, \quad I_3 = |\delta_3|^2. \quad (32.2)$$

You will notice that the intensities of waves from the different holes follow (Eq. 32.1). If we expand $|\delta_3|^2$ we see that

$$|\delta_3|^2 = |\delta_1|^2 + |\delta_2|^2 + 2[\delta_1 \delta_2 \cos \theta], \quad (32.3)$$

where θ is the phase difference between δ_1 and δ_2 . In terms of the intensities, we could write

$$I_{12} = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \theta. \quad (32.4)$$

The last term in Eq. 32.4 is the "interference term." We may for most waves. The intensity can have any value, and it always is, when

32.4 An experiment with electrons

Now we propose a similar experiment with electrons. It is shown diagrammatically in Fig. 32.3. We make an electron gun which consists of a tungsten wire heated by an electric current, and a cathode ray tube, a metal box with a hole in it. If the wire is at a negative voltage with respect to the box, electrons emitted by the wire will be repelled and cannot hit the wire and some will pass through the hole. All the electrons which come out of the gun will after (just like) the water waves in front of the sun is seen as a wall (just a thin misty pattern with maxima and minima) if

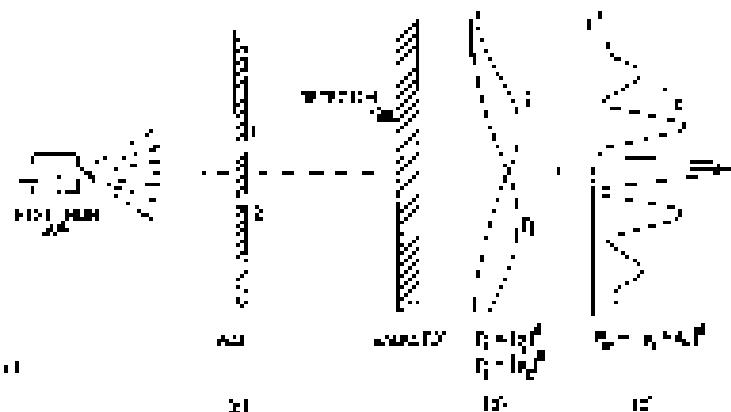


FIG. 37-2. Interference experiment with electrons.

Before we do another glove which will serve to cover up the front of the backstop we place a more elaborate detector. This detector consists of a paper tape counter, perhaps followed by an electrical integrator, which is connected to a loudspeaker.

We should say right away that we should not try to set up this experiment (as you could have done with the one we have already described). Our experiment has never been done in just this way. It remains to be seen if the apparatus would ever in fact be made or if it possibly would fail to show the effects which we expect. We are doing a "thought experiment," which we have chosen because it is easy to think about. We know the results that would obtain because there are many similar ones that have been done, in which the scale and the proportion have been chosen to give the effects we shall describe.

The first thing we notice with our electron experiment is that we hear sharp "clicks" from the detector (that is, from the loudspeaker). And all "clicks" are the same. There are no "half-clicks."

We would also notice that the "clicks" come very erratically. Something like click . . . click click . . . click . . . click click . . . click . . . etc., just as you have no doubt heard a radio timer operating. If we count the clicks which arrive in a sufficiently long time—say for many minutes—and then count again for another equal period, we find that the two numbers are very nearly the same. So we can speak of the average rate at which the clicks are heard (say, about many clicks per minute at the end of a day).

As we move the detector around, the instant at which the clicks appear is fast or slow; but the overall character of each click is always the same. If we lower the temperature of the gun or the gun tube, or of the detector, it goes down, but still each click sounds the same. We would notice also that if we put separate detectors at the backstop, say 8 inches apart, we still click, but never both at once. (Even if this does not occur in a while, if there were two clicks very far apart in time, you might not notice the separation.) We conclude, therefore, that whatever arrives at the backstop comes in "umps." All the "umps" are the same size, only while "umps" arrive, and they arrive one at a time at the backstop. We shall say "electrons always arrive in individual lumps."

Just as in our experience with bullets, we can now proceed to "feel experimentally" the answer to the question: "What is roughly the probability that an electron 'ump' will arrive from the backstop various distances from the center?" As before, we calculate a relative probability by measuring the rate of clicks within a given range of position. The probability that humps will arrive in a particular place is proportional to the average rate of clicks at that place.

The result of our experiment is the interesting curve marked $I_0 e^{-x^2/l^2}$ in Fig. 37-2. See this is the way electrons go.

37-5. The interference of electron waves

Now let us try to analyze the curve of Fig. 37-2 to see whether we can understand the behavior of the electrons. Let us first bring in what may be called "electron wave columns," each lump which we may as well call an electron, has an "electron wave column" through hole 2. Let us write this in the form of a "Proposition":

Proposition A: Both electrons enter goes through hole 1 or it goes through hole 2.

Assuming Proposition A, > electrons road arrive at the backplane can be divided in two classes: 1) those that come through hole 1, and 2) those that come through hole 2. So our observed curve would be the sum of two effects of the electrons which come 1. In hole 1 see the electrons whilst come through hole 2. This includes interference by experiment. First, we will make a measurement for those electrons that come through hole 1. We knock off hole 2 and count our counts of the electrons from the detector. From the counting rate, we get P_1 . The result of the measurement is shown by the curve marked P_1 in part (b) of Fig. 3-3. The result seems quite reasonable. In a similar way, we measure P_2 , the probability distribution for the electrons that come through hole 2. The result of this measurement is also shown in the figure.

The result $P_{1,2}$ obtained with both holes open is clearly not the sum of P_1 and P_2 , the probabilities for each hole alone. In analogy with our water-wave experiment, we say: "There is Interference."

$$\text{For interference: } P_{1,2} \neq P_1 + P_2. \quad (37.5)$$

How can such an interference come about? Perhaps we should say: "Well, but course, probability, too, it is not true that the things go either through hole 1 or hole 2, because if they did, the probability would be 1. Perhaps they go in a more complicated way. They split in half and . . ." But no! They cannot they always arrive in lumps . . . "Well, perhaps some of them go through . . . and then may go around through 2, and then around a few more turns, or by some other complicated path . . . then by closing hole 1, we changed the chance that an electron that was going out through hole 1 would finally get to the backplane . . ." But notice. There are some situations which we know very well when both holes are open, but which leaves many electrons if we close one hole, so closing one hole removed the number from the other. But this however, just in the center to the pattern, $P_{1,2}$ is more than which gives $P_1 + P_2$. It is as though closing one hole decreased the number of electrons which come through the other hole. It seems here to explain this effect by regarding the the electrons travel in complicated paths.

"I'm afraid my ideas. And the more you look at it, the more mysterious it seems. Many ideas have been suggested to try to explain the curves for $P_{1,2}$ to terms of individual electrons going around in complicated ways through the holes. None of them has succeeded. None of them can get the right curve for $P_{1,2}$ in terms of P_1 and P_2 .

Yet we mightly suspect, the requirement for relating P_1 and P_2 to $P_{1,2}$ is extremely simple. For $P_{1,2}$ is just like the curve $S_{1,2}$ of Fig. 37-2, and that was simple. What is going on at the backplane, he described by two complex numbers that we will call δ_1 and δ_2 (they are functions of x , of course). The contribution of δ_1 gives the other you only have to add. That is, $P_1 = |\delta_1|^2$. The other were only hole 2 been to given by δ_2 in the same way. That is, $P_2 = |\delta_2|^2$. And the combined effect of the two holes is just $P_{1,2} = |\delta_1 + \delta_2|^2$. The conclusion is the same as one we had for the water waves. (It is hard to see how one could get such a simple result from complicated paths of electrons going back and forth through the hole on some strange trajectory.)

We conclude the following: The electrons arrive in lumps, not particles, and the probability of finding n these lumps is distributed like the distribution of intensity of a wave. It is in this sense that an electron behaves "sometimes like a particle and sometimes like a wave."

Incidentally, when we were working your classical waves we defined the intensity as the trace over time of the square of the wave amplitude, and we used complex numbers as a mathematical trick to simplify the analysis. When quantum mechanics it comes in, that the amplitudes must be represented by complex numbers. The real parts alone will not do. This is a technical point for the moment, because we'll consider back just the same

Since the probability to arrive here of both holes is given as simple addition it is not equal to $(P_1 + P_2)$, that is really to say. But there is a large number of况sories involved in the fact that nature does work this way. We would like to illustrate some of these subtleties for you now. First, since the number that arrives at a particular point is not equal to the number that arrives through 1 plus the number that arrives through 2, as we would have concluded from Proposition A, undoubtedly we should conclude that Proposition A is false. It is not true that the electrons go either through hole 1 or hole 2. But this conclusion can be tested by another experiment.

37-6 Watching the electrons

We do this by the following experiment. To our electron apparatus we add a very strong light source: glass! Behind the wall and between the two holes, as shown in Fig. 37-4. The glass and electric arcs give rather light. So when an electron passes through a glass pane, owing to the reverb., it will scatter some light to our eye, and we can see where the electron goes. If, for instance, an electron were to take the path via hole 2 that is indicated in Fig. 37-4, we should see a band of light coming from the vicinity of the place marked A in the figure. If an electron passes through hole 1 we would expect to see a band B in the vicinity of the upper hole. If it should happen that we get light from both places at the same time, because the electron divides in half... (i.e. we do the experiment!!)



Fig. 37-4. A different electron experiment.

Here is what we see: every time that we hit a "click" from our electron detector δ , the backs up we also see a flash of light either near hole 1 or near hole 2, but never both at once! And we observe the same results no matter where we put the detector. From this observation we conclude that when we think of the electrons we find that the electron goes either through one hole or the other. Ergo immediately, Proposition A is necessarily true.

What, then, is wrong with our argument against Proposition A? Why isn't Proposition A right in Fig. 37-4? Back in experiment 1 let us stop most of the electrons and find out what they are doing. For each position (or location) of x in $y = x$ we will count the electrons that arrive and also keep track of which hole they went through, by watching for the flashes. We can keep track of things, for why, whenever we hear a "click" we will turn a counter in Column 1 if we see the flash near hole 1, and if we see the flash near hole 2, we will turn a counter in Column 2. Every electron which arrives is counted in one of two classes: those which come through 1 and those which come through 2. From the numbers recorded in Column 1 we get the probability P_1 that an electron will arrive at the detector via hole 1, and from the numbers recorded in Column 2 we get P_2 , the probability that an electron will arrive at the detector via hole 2. If we now repeat and a measurement for many values of x , we get the curves for P_1 and P_2 shown in part (a) of Fig. 37-4.

We find that this is certainly surprising. We get the something quite similar to what we got before for P_1 by blocking off hole 2 and P_2 is similar to what we got by blocking hole 1. So there is an very unexpected business like going through both holes. When we watch now, the electrons come through just as we could

move there to come through. Whether the holes are closed or open, those which we see come through hole 1 are distributed in the same way whether hole 2 is open or closed.

But wait! What do we know for the total probability, the probability that an electron will survive all the detectors by any route? We already have that information. We just pointed that we can record a single light flashes, and we keep track of the detector clicks which we have recorded into the two columns. We know, in fact, the numbers. For the probability that an electron will survive the two holes by passing through either hole, we've had $P_{1,2} = P_1 + P_2$. That is, although we succeeded in *writing out* how one electron could pass through, we no longer get the old probabilities $P_{1,2}$ or P_1 , but a new one, $P_{1,2}$, *causing no interference*. It's our fault that $P_{1,2}$ is restored.

We now can think that when we look at the electrons the distribution of them in the screen is coherent with what we saw on each. Perhaps it is something in our light source that causes this? It might be that the electrons are very electric, and the light, when it captures all the electrons, gives them a jolt that changes their motion. We know that the electric field of the light source can change will went a photon on it. So perhaps we should expect the motion to be changed. Anyway, the light exerts a big influence on the electrons. By trying to "watch" the electrons we have changed their motion. That is, by just going to the place on where the photon is considered to be, we try to change the electron's motion enough so that it it might have gone somewhere else. If it was a mass point, it will instead travel where it was a remnant; that is why we no longer see the wave interference effects.

You may be thinking: "Don't use such a big 'gun'—turn the brightness down! The light waves will then be weaker and we can measure the electrons much more surely, by seeing the light dimmer and dimmer, eventually the wave will be weak enough that it will have a negligible effect." It is. Let's say it. The last thing we observe is that the *coherent* light measured from the electrons as they pass by does not get weaker. It always has constant Bush. The only thing that happens as the light is made dimmer is that sometimes we hear a "click" from the detector but see no Bush at all. This situation has gone along with increasing $P_{1,2}$. When we are decreasing is the light also acts like electrons, we know that it is as " $P_{1,2}$," but now we find that it is also " $P_{1,2}$," it always arrives in a scattered—in truth, that we call "photon." As we turn down the intensity of the light source we do not change the ϕ of the photons, only the rate at which they are emitted. This explains why, when we turn our dim, some electrons go by without being seen. They did not happen to be a photon and not in the range the electron went through.

This is all a bit disconcerting. It is because whenever we "see" the electron we see the normalized Bush, i.e., how many times we see are above the detector noise. Let us try the electrons from our own light sources. Now whenever we hear a click in the detector we can keep a count. i.e., how many times in Column 1; those that were seen by hole 1. In Column 2, those electrons seen by hole 2, and in Column 3, those electrons not seen at all. When we work up everything (computing the probabilities) we find these results. These "seen by hole 1" have a distribution like $P_{1,2}$; but the "seen by hole 2" have a distribution like $P_{1,2}$; and those "not seen at all" have a "heavy" distribution just like $P_{1,2}$ of Fig. 10-2. If the electrons are not seen, no more surprises!

That is understandable. When we do see an electron, no photon can be in, and when we do see it, a photon has disturbed it. There is always the same amount of disturbance from the light photons; all photons are registered effects and the effect of the previous being scattered is enough to smear out any interference effects.

Is there not some way we can see the electrons without disturbing them? We learned in an earlier chapter that the momentum carried by a "photon" is linearly proportional to its wavelength by $= \lambda/2\pi$. Considering the job given to the electrons when the photon is scattered, would our eye depend on the momentum that photon carried? And if we want to detect the electrons only

Energy we could not have known! The intensity of the light we should have known if its frequency rose with as increasing its wave length. That is, the light of a greater color. We could have the infrared light, or even waves like radar, and "heat" where the electrons went to form, help of some energy given. But can "heat" light of these higher wave lengths. It is the "geometric" light propagation we will be discussing the electrons so much.

Let us try the experiment with larger waves. We shall keep repeating our experiment with the small light of a longer wavelength. At first nothing seems to change. The results are the same. Then a terrible thing happens. You remember that when we discussed the microscope we pointed you to, due to the wave nature of the light, how a limited number above two spots can be had or can be seen as two separated spots. This distance is of course of the wavelength of light. So now, when we make the wavelength longer than the distance between our holes we see a big fuzzy flash when the light is scattered by the electrons. You can no longer tell which hole the electron went through. We just don't want to see it! And it is just as I light of the radar that we find that the jets grow as the jets are strong enough so that they begin to look like P_1 , that we begin to get some interference pattern. And it is only for wavelengths longer than the separation of the two holes ($\lambda > d$) that we get at all scattering where there is no result that the distance d , i.e., the light goes sufficiently small that we again get the curve P_1 shown in Fig. 31-3.

In our experiment we find that it is impossible to change the light in such a way that we can tell which hole the electron went through and yet the wave still not change. Let me tell. It was suggested by Heisenberg that the "new laws of mechanics" only by connection with the wave mechanics can be experimental possibilities and previous theoretical. He proposed, as a general principle, his uncertainty principle, which we can verify in our experiment as follows: "It is impossible to design an apparatus to find out which hole the electron passes through, but still not to decrease time during which hole the electron goes through. It cannot be so ruled out, if it does not disturb the pattern in an essential way. No one has ever found (or even thought of) a way around the uncertainty principle. So we must assume that it describes a basic structure of nature.

The complete theory of quantum mechanics which we now have is that the charge and, in fact, the matter depend on the interpretation of the uncertainty principle. Since quantum mechanics is such a proceeding theory, you believe in the uncertainty principle is reinforced. But the way in "heat" the uncertainty principle were ever discussed, quantum mechanics would give inconsistent results and you would never discussed it as a valid theory of atoms.

"Well," you say, "What about Proposition A? Is it true, or is it not true? If the electron does pass through hole 1 or it goes through hole 2?" The only answer we can be given is that we have found from experiment that there is a certain special way that we have a third possibility that we do not get in a meaningful ratio. What we can say is avoid taking wrong position and to do the following. If one looks at the detector, then one asks, also, if one has a piece of apparatus which is capable of detecting whether the electron goes through hole 1 or hole 2, then one can say that it is possible through hole 1 or hole 2, but when one does not try to tell which wave the electron goes, when there is nothing in the experiment to disturb the electrons then one may not say that an electron goes either through hole 1 or hole 2. If one does say it, and starts to make any deductions from the statement, he will have errors in the analysis. This is the logical tightrope on which we must walk if we wish to deduce nature successfully.

To the notion of "heat" as well as electrons—must be described in terms of waves, what would we have in our first experiment? Why didn't we see 62 different reprints in there? It turns out that the bullet-like wavelengths were too tiny that no interference pattern became visible. So however fast the wave may

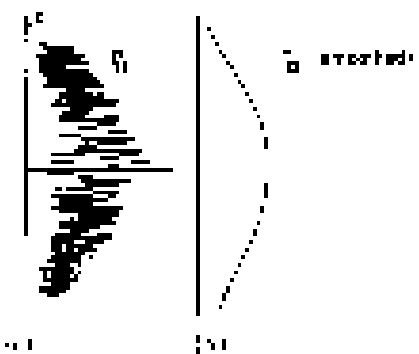


Fig. 37-5. Interference pattern with (a) both slits open; (b) one slit covered.

detector of finite size one could not distinguish between passing and missing. What we see is only a kind of average which is not classical wave. In Fig. 37-4 we have tried to indicate schematically what happens with large-scale objects. Part (a) of the figure shows the probability distribution for one might predict by billiards, using classical mechanics. The two bounces are supposed to represent the interference pattern and get two waves of very short wavelength; b) Any physical detector, however, smudges several wiggles of the probability curve, so that the measurements show DeBroglie's curve shown in part (b) of the figure.

37-7 Basic principles of quantum mechanics

We will now write a summary of the most important of the experiments. We will, however, put the results in a form which makes them more or less a general class of such experiments. We can write our summary more simply if we first define an "ideal experiment" as one in which there is no uncertainty or error influence, i.e., no jiggling or other things going on that we cannot take into account. We should be quite precise if we want an ideal experiment to mean a situation in which all of the initial and final conditions of the experiment are completely specified. When we will call "an event" as, in general, just a specific set of time, position conditions. (For example: "an electron leaves the gun, arrives at the detector, and nothing else happens.") Now for our summary:

Summary

- (1) The probability of an event in an ideal experiment is given by the square of the absolute value of a complex number ψ which is called the probability amplitude.

$$\begin{aligned} \psi &= \text{probability}, \\ &\quad \times \text{probability amplitude} \\ P &= |\psi|^2 \end{aligned} \quad (37.6)$$

- (2) When an event can occur in several alternative ways, the probability amplitude for the event is the sum of the probability amplitudes for each way (classical superposition). There is ADDITION.

$$\begin{aligned} \psi &= \psi_1 + \psi_2, \\ \psi &= \psi_1 + \psi_2 = \psi_1^2 + \psi_2^2. \end{aligned} \quad (37.7)$$

- (3) If an experiment is performed which is capable of determining whether one or another alternative is actually taken, the probability of the event is the sum of the probabilities for each alternative. The difference is lost.

$$P = P_1 + P_2. \quad (37.8)$$

One might still like to ask: "How does it work? What is the machinery behind the law?" We might wonder "machinery" while saying, "No one can explain" any more than we have just "explained". No one will give you any deeper representation of the situation. We have no idea about the mechanism from which these rules were derived.

We would like to emphasize a very important difference between classical and quantum mechanics. We have been talking about the probability that an electron will arrive in a given circumstance. We have recalled that in the experimental arrangement (or even in the best possible one) it would be impossible to predict exactly what would happen. We can only predict the *event*. This would mean, if it were true, that physics has given us on the problem of trying to predict exactly what *will* happen in a definite circumstance. Yet physics has given up. We do not know how to answer what *could* happen in a given circumstance, and we believe now that it is impossible that the only thing that can be predicted is the probability of different events. It can be recognized that this is a mechanism in our castle, ideal of understanding nature. It may be a backward step, but no one has ever a way to avoid it.

The more now is the news. It can suggest that something has been made to be by the description we have given. "Perhaps the electron has some kind of internal structure which is not visible. Or we do not yet know enough. Perhaps that is why we cannot predict where it will appear." If we could look more closely at the electron, we could be able to see where it would end up. So far as we know, this is impossible. We would still be in difficulties. Suppose we were to assume and inside the electron there is some kind of machinery that determines where it is going to end up. That machine must also determine which hole it is going to go through in this way. But we must note again that what is inside the electron should not be dependent on what we see, and in particular you wouldn't expect a choice one of the holes. So the electron, before it comes, has already made up its mind the other hole it is going to pass through when it is going to land. We should then say, for those electrons that have chosen hole 1, P_1 for those that have chosen hole 2, and correspondingly the sum $P_1 + P_2$ for those that come through the two holes. There seems to be no way around this. But we have verified experimentally that that is not the case. And no one has figured a way out of this puzzle. And he probably finds it very difficult to compute probabilities. We say "at the present time," but we expect very strongly that it is something that will be with us forever. But it is impossible to beat that puzzle—that this is the way nature really is.

37-8 The uncertainty principle

This is Heisenberg's and the uncertainty principle originally: If you make the measurement to any greater and greater accuracy the component of the momentum with uncertainty Δp , you cannot at the same time know its position more accurately than $\Delta x = \hbar/k$. The uncertainty in the position and momentum of any particle must have this product \hbar/k . This is Planck's constant. This is a special case of the uncertainty principle that was stated above. Quantitatively, The more general statement says that one cannot design experiments to say anything about a particle's position without, at the same time destroying the pattern of interference.

Let us show for example with this Fig. the kind of relation given by Heisenberg must be true in order to keep from getting contradictory. We imagine a modification of the experiment of Fig. 37-2, in which we go with the holes consists of a polarizer and a plate, both can respectively, stand down (just like 37-2), or stand in Fig. 37-6. By watching the motion of the pulse carefully we can try to tell which 1 or 2 an electron goes. I mean, imagine what happens when the deflector is placed at $x = 0$. We would expect that an electron which passes through hole 1 must be deflected downward by the plate to reach the deflector. Since the vertical component of the electron's initial momentum is unchanged, the plate must receive wave equal momentum in the opposite direction. The pulse will get an upward kick. If the electron goes through the lower hole, the pulse shot off from downward kick. If however the horizontal position of the deflector, the momentum received by the plate will have a different value for a lower hole than for a higher hole. Only if $P_1 = P_2$ will the distribution of the electrons match, but just by watching the plates, we can tell which path the electron took.

Now in order to do this it is necessary to know where the momentum of the electron is before the slit in your slit. So when we measure the momentum after the electron goes by, we ask, figure out how much the electron's momentum has changed. But, of course, according to the uncertainty principle we cannot at the same time know the position of the electron with an arbitrary accuracy. But if we do not know exactly where the electron we cannot say precisely where the two holes are. They will be in a different places for every electron that goes through. This means that the pattern of our interference pattern will have a different location for each electron. The wiggles of the interference pattern will be continually. We shall discuss quantitatively in the next chapter that if we determine the momentum of the plate sufficiently accurately by determining from the result measurement which hole was used, then it is consistent if the position of the plate was according to the con-



Fig. 37-6. An experiment in which the recoil of the wall is measured.

certainty principle, he enough to this case problem observed at the detector up and down in the scattering plane. In other words, there is no need to do better than this. Such a randomness is just enough to smear out the pattern of the interference I observed.

The uncertainty principle "protects" quantum mechanics. Heisenberg imagined that if it were possible to measure the momentum and the position simultaneously with a greater accuracy, the electron would collapse. So he proposed that it must be impossible. Two people sat down and tried to figure out ways of doing it, and suddenly Schrödinger cut a way to measure the position and the momentum of anything—a screen, an electron, a hill and everything—with very greater accuracy. Quantum mechanics remained as perhaps the accurate existence.

The Relation of Wave and Particle Viewpoints

38-1 Probability wave amplitudes

In this chapter we shall discuss the relationship of the wave and particle viewpoints. We already know, from the last chapter, that between the wave viewpoint and the particle viewpoint there is a conflict. Usually we have tried to present things adequately, or at least precisely enough, but they will not have to hand except when we learn more. It may be extended, but it will not be complete. The idea we try to talk about, the wave picture in the particle picture, both are approximate, and both will change. Therefore what we learn in this chapter will not be accurate in a certain sense; it is a kind of half-truthive fragment that will be much more precise later, but certain things will become really clear when we interpret them exactly in quantum mechanics. The reason for doing such a thing, of course, is that we are not going to go directly into quantum mechanics, but we want to have at least some idea of the kinds of effects that we will find. But first we will say again that we think waves and particles, and so it is rather fancy to use the wave and particle ideas except under standing circumstances. In other circumstances either we know the amplitude or the probabilities of the quantum-mechanical amplitudes. We shall try to illustrate the weaker places as we go along, but most of it is very much easier. It is not a matter of interpretation.

Very often we know that the best way of representing the world in quantum mechanics—*the best framework*—is to give an amplitude for every event that we can, and if the event involves the position of one particle then we can give the amplitude to find that particle at different places and at different times. The probability of finding the particle is then proportional to the absolute square of the amplitude. In general, the amplitude to find a particle in \vec{r} is not given by a different expression than position has time:

In a special case the amplitude varies sinusoidally in space and time like $e^{i(k\vec{r}-Et)}$, you must forget that these amplitudes are complex numbers, not real numbers; and involves a definite frequency ω and wave number k . I can ignore all that this can depend on a certain limiting situation where we would have believed that we have a particle whose energy E was known and is related to the frequency ω :

$$\mathcal{E} = \hbar\omega, \quad (38.1)$$

and where ω in electron-volts is also known and is related to the wave number by

$$\omega = \frac{\hbar k}{m}. \quad (38.2)$$

This means that the idea of a particle is limited. The idea of a particle—in location, its momentum, etc.—which we use so much, is in certain ways incorrect. For instance, if we amplitude to find a particle at different places is given by $e^{i(k\vec{r}-Et)}$, where ω and k are constant, that would mean that the probability of finding a particle is the same at all places. That means we do not know where it is—it can be anywhere. There's a great uncertainty in its position.

On the other hand, if the position of a particle is more or less well known and we can predict it fairly accurately, then the probability of finding it at different places will be confined to a certain region, which I will call S . Outside this region, the probability is zero. Now this probability is the absolute square of an amplitude, and since the absolute square is zero, the amplitude is also zero, so that

38-2 Probability wave amplitude

38-2.1 Measurement of position and uncertainty

38-2.2 Crystal diffraction

38-2.3 The size of an atom

38-2.4 Energy levels

38-2.5 Photochemical applications



Fig. 38-1. A wave packet of length Δx .

or have a wave train whose length is Δx (Fig. 38-1), and the wavelength (the distance between nodes of the waves in the train) of the wave train is what corresponds to the particle momentum.

This is an interesting & strange thing about waves: a very simple thing which has nothing to do with quantum mechanics really. It is something that anybody who works with waves, even if he knows no quantum mechanics, knows: namely, we cannot define a unique wavelength for a wave train. Such a wave train does not have a definite wavelength; there is an indeterminacy in the wave number that is related to the size or length of the train, and thus there is an indeterminacy in the momentum.

38-2 Uncertainty of position and momentum

Let us consider two examples of the idea—*to see the reason why there is an uncertainty in the position and/or the momentum if quantum mechanics is right*. We have already seen, before that if there were no such a thing—*i.e.*, if we had a model of the position and the momentum of anything simultaneously—we would have a paradox, it is terrible that we do not have such a paradox, and the fact that such a uncertainty exists naturally from the wave picture shows the consistency is naturally consistent.

There is one example which shows the relationship between the position and the momentum in a continuous way, is very instructive and—Suppose we have a single slit, and particles are coming from very far away with a certain energy. As they are all coming, essentially horizontally (Fig. 38-2). We suppose a constant v_0 is the horizontal component of momentum. All of these particles have a certain horizontal momentum v_0 , say, in a constant sense. So, in the classical sense, the vertical momentum v_y , because the particle goes through the slit, is definitely known. The particle is moving either up or down, because it comes from a source that is far away—and so the vertical momentum is of course zero. But now let us suppose that it goes through a hole whose width is Δ . Then after it has come out through the hole, we know the position vertically—the position—was considerable uncertainty, exactly Δ . That is, the uncertainty in position, Δx , is of order Δ . Now we might also want to say, since we know the momentum is reasonably horizontal, that Δp , is zero; *but that is wrong*. We don't know the momentum was horizontal, but we do not know it any more. Before the particles pass through the hole, we did not know their vertical positions. Now that we have forced the vertical position by having the particle come through the hole, we have lost our information on the vertical momentum! Why? According to the wave theory there is a spreading out, or dispersion, of the waves after they go through the slit, just as the light. Therefore there is a certain angle for that particles coming out of the slit are not coming exactly straight. The pattern is spread out by the diffraction effect, and the angle of spread, which we can define as the angle of the first minimum, is a measure of the uncertainty in the final angle.

How does the pattern become spread? To say it is spread means that there is some chance for the particle to be moving up or down, that is, to have a component of momentum v_y in given. Which starts and particle has, we can detect this diffraction pattern with a screen to the right, say, when the screen receives the particle, say at C (Fig. 38-2), it receives the entire particle, so that it is observed where the particle has a certain momentum, in order to go from the slit up to C.

To get a rough idea of the spread of the momentum, the vertical momentum v_y has a spread which is equal to $\lambda/\Delta x$, where λ is the horizontal wavelength. And how big is Δx is the spreading parameter? We know the first minimum occurs at an angle θ_1 such that the waves from one edge of the slit have to travel one wavelength farther than the waves from the other side. We worked that out before (Chapter 30). Therefore $\theta_1 = \lambda/v_0$, and $\Delta x = \lambda v_0 \theta_1$ by experimentally justified. Note that if we make Δx smaller and make a more accurate measurement of the position of the particle, the CII makes pattern gets wider. Remember, when we choose the slit for the experiment with the monochromes, we had more intensity there, but. So the narrower we make the slit, the wider the pattern gets on the

—you is to like "most" the we could find that the particle has sides be momentum. Thus the uncertainty in the vector momentum is inversely proportional to the uncertainty in λ . In fact, we see that the product of the two is equal to $\hbar/2$. But λ is the wavelength and p_x is the momentum, and in accordance with quantum mechanics, the wavelength from the momentum is De Broglie's formula. So we obtain the rule that the uncertainties in the vector momentum and in the vector position have a product of the order \hbar .

$$\Delta p_x \Delta x \approx \hbar \quad (38.1)$$

We cannot prepare a system in which we know the vertical position of a particle and can predict how it will move vertically with greater certainty than given by (38.1). That is, the uncertainty in the vertical momentum has crossed here, where Δx is the uncertainty in our knowledge of the position.

Sometimes people say that this statistics is all we do. When the particle passes through the slit, its vertical momentum is very small. And now that I have gone through the slit, its position is known. Now position and momentum seem to be known with infinite accuracy. It is quite true that we can receive a particle, and on reception determine what its position is and what its momentum is. We would have had to have been to have gone there. That is true, but that is not what the uncertainty relation (38.1) refers to. Equation (38.1) refers to the probability of a situation, not merely about the past. I then we good to say "I know what the momentum was before it went through the slit, and now I know the position," does not now he moment of knowledge: it does. The fact that it went through the slit no longer permits us to predict the vertical momentum. We are talking about a predictive theory, not just measurements after the fact. So we must talk about what we can predict.

Now let us take the thing the other way around. Let us take another example of the same phenomenon, a little more quantitative. In the previous example we measured the momentum by a classical method. Namely, we calculated the deflection and the velocity and the angles, etc., so we got the momentum by classical analysis. But since momentum is related to wave number, there exists an entirely different way to measure the momentum of a particle: *grating*, otherwise which has no classical analog, because it uses Eq. (38.1). We measure the wave length of the waves. Let us try to measure momentum in this way.

Suppose we have a grating with a large number of lines (Fig. 38-3), and send a beam of particles at the 45° slit. We have often discussed this problem: if the particles have a definite momentum, then we get a very sharp pattern in a certain direction, because of the interference. And we have also allowed that, however, only we can determine our momentum, that is to say, whatever measuring power of such a grating is to be then there is again we refer to Chapter 33, where we found that the relative uncertainty in the wavelength that can be measured with a given grating is $1/N^m$, where N is the number of lines on the grating and m is the order of the diffraction pattern. Then is,

$$\Delta \lambda/\lambda = 1/N^m. \quad (38.4)$$

Now De Broglie (38.1) can be rewritten as

$$\Delta p_x/\lambda = 1/N^m = 1/d, \quad (38.5)$$

where d is the distance shown in Fig. 38-3. This distance is the difference between the total distance that the particle is away in whatever it is sent to travel if it is reflected from the bottom of the grating, and the distance that it has to travel if it is reflected from the top of the grating. That is, the waves which form the diffraction pattern are waves which come from different parts of the grating. The first ones that come come from the bottom edge of the grating, from the beginning of the wave train, and the rest of them come from later parts of the wave train, coming from different parts of the grating until the last one. Finally we have, and that involves a point in the wave train a distance d behind the first point. So in order that we shall have a spot in our spectrum corresponding to a definite momentum

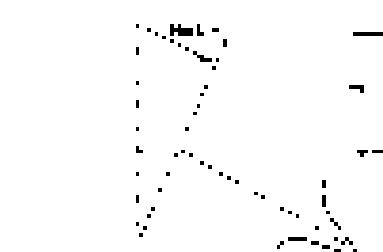


Fig. 38-3. Determination of momentum by using a diffraction grating.

With our birth control plan in (35.4), we have to have wave train of total length L . If the wave train is too short we are not using the entire train. The waves which form the structure are being reflected from only a very short section of the grating. If the wave train is too short, and the grating will not work right—we will find a big angle spread. In order to get a narrower one, we need to use the whole grating, so that at least at some moment the whole wave train is scattering simultaneously from all parts of the grating. Thus the wave train must be of length L in order to have an uncertainty in the wavelength less than that given by (35.5) incidentally,

$$\Delta \lambda \approx \frac{c}{(L/\lambda)} = \frac{c\lambda}{L}, \quad (35.6)$$

Therefore

$$\Delta \lambda = 2\pi/L, \quad (35.7)$$

where L is the length of the wave train.

This means that if we have a wave train whose length is less than L , the uncertainty in the wave number must exceed $2\pi/L$. Or the uncertainty in the wave number times the length of the wave train. We will see that for a moment $\Delta \lambda$ exceeds $2\pi/L$. We call it $\Delta \lambda$ because that is the uncertainty in the position of the particle. If the wave train exists only in a finite length, then that is where we would find the particle, within an uncertainty Δx . Now this property of waves, the "length of the wave train" gives the uncertainty of the wave number associated with it. This is a property that is known to everyone who studies them. It has nothing to do with quantum mechanics. It is simply that if we have a finite train, we cannot count the waves in it very precisely. Let us try another way to see the reason for that.

Suppose that we have a finite train of length L ; then because of continuity it has to have Δx . As shown in Fig. 35-1, the number of waves in the length L is uncertain by something like ± 1 . But the number of waves is $L = cL/\lambda$. Thus λ is uncertain, and we cannot get the result (35.7), a property of many waves. The same thing works whether the waves are in space and L is the number of radiations per centimeter and λ is the length of the train, or the waves are in time and L is the number of oscillations per second and $T \rightarrow L$ is the "length" in time; that the waves exist during L . That is, if we have a wave train lasting only for a certain finite time L , then the uncertainty in the frequency is given by

$$\Delta \omega = 2\pi/L. \quad (35.8)$$

We have tried to emphasize that these are properties of waves alone, and they are well known, for example, in the theory of sound.

The point is that in quantum mechanics we interpret the wave number as being a measure of the momentum of a particle. With the rule that $p = \hbar k$, equation (35.7) tells us $\Delta p = \hbar \Delta k$. This, then, is a formulation of the classical idea of momentum. (No, it's k , it has to be modified in some ways if we are going to represent particles by waves). It is like this: we have found a rule that gives us some idea of when there is a failure of classical ideas.

35-2 Crystal diffraction

Now let us consider the reflection of parallel waves from a crystal. A crystal is a little thing which has a whole lot of smaller crystals— we will include some complications later—in a nice lattice. The question is how to set up energy in this lattice, so that it reflects maximum in a given direction, not a given beam of, say, light (or X-ray electrons, or photons, or anything else). In order to obtain a strong reflection, the scattering from all of the lattice must be in phase. Thus, if each harmonic has a phase in phase and out of phase, the waves will cancel out. The way to arrange things is to find two regions of constant phase, as we have already explained; they are places where waves begin in phase with the initial and final vibrations (Fig. 35-4).

If we consider two parallel planes as in Fig. 35-4, the waves scattered from the two planes will be in phase provided the difference in distance travelled by a wave

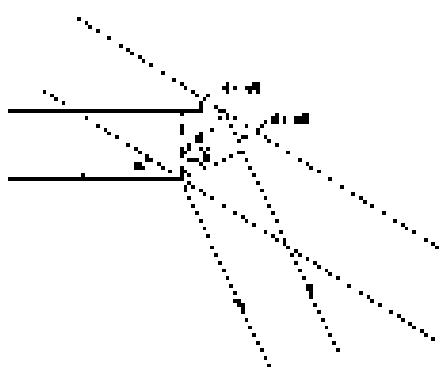


Fig. 35-4. Scattering of waves by crystal planes.

front is an interval on the back of wave-lengths. This difference can be seen to be $2\pi \sin \theta / d$, where d is the perpendicular distance between the slits. Thus the condition for coherent reflections is

$$2\pi \sin \theta / d = (n - 1) \lambda \quad (38.9)$$

In other words, the crystal is said that no wave-changes will occur when the angle of reflection (Eq. 9) with $n = 1$, i.e. there will be a strong reflection. If, on the other hand, there are other n 's instead of the same $n = 1$ (spaced in a certain regularity however), then the inter-plane plates will also scatter equally strongly and will interfere with the others and produce an effect. But it is (Eq. 9) must refer to adjacent planes, not to distant planes far away. Interference will not be this simple.

As a matter of fact, actual crystals are not usually as simple as a single kind of atom arrangement; it is certainly not. Instead, if we make a two-dimensional model, they are much like wallpaper, in which there is some kind of "figure" which repeats all over the wall-paper. By "figure" we mean, in the case of atoms, some arrangement—either more or less random three-arrangements, e.g., for calcium carbonate, or, however, what may happen a practically any number of atoms. But whatever it is, the figure is repeated in a pattern. This basic figure is called a unit cell.

The basic pattern of repeating defines what we call the lattice law: the lattice spacing can be immediately determined by looking at the reflexions and seeing which axes of symmetry it has. In other words, if we see distinct reflections of all directions, the law is true. But in order to determine what is in each of the elements of the lattice one must take into account the intensity of the scattering in the various directions. These directions are independent of the type of lattice, but their intensity each successive is determined by what is inside each unit cell, and in that way the structure of crystals is worked out.

Two photographs of x-ray diffraction patterns are shown in Figs. 38-3 and 38-4. Crystalline scattering from rock salt are monoplatinum respectively.

Incidentally, an interesting thing happens if the spacings of the two unit planes are less than $\lambda/2$. In this case (Fig. 9) the reflection law is violated. Look at Fig. 9 again. Notice the distance between adjacent planes. Now there is no side diffraction pattern, and the light—on which even it's wavelength λ —goes through the material without being scattered or reflected back. So in the case of light, where λ is much bigger than the spacing, when it's scattered through there is no return of reflection from the planes of the crystal!

This behavior is an interesting consequence of waves which make everything there are physically justified. No everybody's money is. If we take two carbons and let them into a tiny block of graphite, the neutron diffuse and work that was done (Fig. 34-7). They diffuse now, so they are bounded by the atoms, but strictly in the wave theory, they are bounded by the atoms because of diffraction from the crystal planes. I think that if we take a very long piece of graphite, the neutrons that come out the far end, in all of long wavelength. In fact, if one plane can, in reality as a function of wave-length, we get nothing except for scattering. It's easier than a certain minimum (Fig. 38-8). In other words, we can get very short neutrons this way. Only the lowest frequency can do. Though, they are not scattered or scattered by the crystal, it is of the property, but been going right through like light through glass, and so, not scattered out the sides. There are many other demonstrations of the validity of neutron wave and waves of other varieties.

46-4 The idea of selection

We have considered another application of the wave-length relation, Eq. (38.5). It must not be taken too seriously, the idea is right but the subject is not fully understood. The idea has to do with the transmission of the wave of atoms, and the fact that, specifically, electrons would reduce light and sound until they settle down right on top of themselves. But that cannot be right quantitatively, because then we would know where each electron was and then tell it was moving



Fig. 38-3



Fig. 38-4

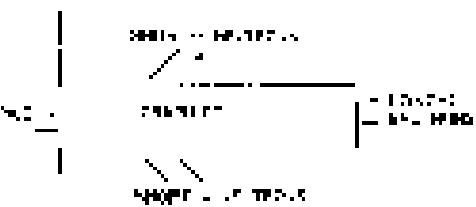


Fig. 38-5. Diffraction of plu neutrons from graphite block.

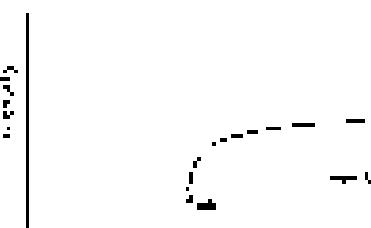


Fig. 38-6. Intensity of neutron scattering from graphite and its behavior at low-wavelength.

Suppose we have a hydrogen atom, and measure the position of the electron; we must not be able to decide exactly where the electron will have the momentum associated with the sum out to be infinite. Every time we look at the electron, it is somewhere, but it has an amplitude to be in different places. These places cannot all be in the nucleus. We shall suppose the radius a is in picometers of order 6. Let us, the distance of the electron from the nucleus is usually about a . We shall determine a by minimizing the total energy of the atom.

The spread of momentum is roughly due to because of the uncertainty relation, so that if we try to measure the momentum of the electron in some manner, such as by scattering x-rays off it and looking for the Doppler effect from a moving scatterer, we would expect not to get zero every time—the electron is not standing still—but the momenta must be of the order $\sim m\omega a$. Then the kinetic energy is roughly $m\omega^2 a^2 = e^2/2m = h^2/2ma^2$. In a word, this is a kind of dimensional analysis. And out in when vary the kinetic energy tends upon Planck's constant, upon a , and upon the size of the atom. We need not treat our atoms as with the classical law $F = kx$. We have now even defined a very precisely just how the potential energy is minus e^2/r for distances from the center, say $-e^2/a$, where the parameter e^2 is the charge of an electron squared divided by $4\pi\epsilon_0$. Now the point is that the potential energy is reduced if a goes smaller, but the smaller a is, the bigger the momentum required, because of the uncertainty principle, and therefore the higher the kinetic energy. The total energy is

$$E = \hbar^2/2ma^2 - e^2/a. \quad (16.1)$$

We do not know what a is, i.e., we know that the atom is going to arrange itself in \rightarrow in some kind of compromise so that the energy is as little as possible. In order to minimize E , we collaborate with respect to a , see the derivative with respect to a . The derivative of E is

$$dE/da = -e^2/m a^2 + e^2/a^2, \quad (16.1b)$$

and setting $dE/da = 0$ gives for a the value

$$a_0 = \hbar^2/m e^2 = 0.055 \text{ angstrom} \\ = 0.055 \times 10^{-10} \text{ meter} \quad (16.1c)$$

This particular distance is called the Bohr radius, and we have thus learned that atomic dimensions are of the order of angstroms, which is right. This is pretty good—in fact, it is amazing, since until now we have had no basis for understanding the size of atoms! Atoms are completely impossible from the classical point of view, since the electrons would spiral in to the nucleus.

Now if we put the value (16.1c) for a , into (16.1b) to find the energy, we come out

$$E_0 = -e^2/2a_0 = -mc^2/2a_0^2 = -13.6 \text{ ev}. \quad (16.1d)$$

What does a negative energy mean? It means that the electron has less energy when it is in the atom than when it is free. It means it is bound. If you give it more energy, it sinks the electron out; it takes energy of the order of 13.6 ev to ionize a hydrogen atom. We have an answer to think that it is not twice or three times this—or half of this, or $1/\pi$ times this, because we have used it as a sloppy argument. However, we have obtained, we have used all the constants in such a way that it happens to come out the right number! This number, 13.6 electron volts, is called a **Radius of energy**; it is the ionization energy of hydrogen.

So we now understand why we do not fall through the floor. As we walk on streets with their masses of atoms push against one floor with its mass of atoms. In order to squash them closer together, the electrons would be confined to a smaller space and, by the uncertainty principle, their momenta would have to be bigger, on the average; and that means high energy; the resistance to atomic compression is a quantum-mechanical effect and not a classical effect. Classically,

we would expect that if we were to draw all the electrons and protons closer together, the energy would be reduced, until further, and the last arrangement of possible and reasonable charges in classical physics is all on top of each other. This was well known in classical physics and was a puzzle because of the existence of the atom. Of course, the early scientists invented some ways out of the trouble—but never mind, we have the right way out now. (Maytag.)

Incidentally, although we have no reason to understand it at the moment in a situation where there are many electrons, it turns out that they try to keep away from each other. If one electron is occupying a certain space, then another does not occupy the same space. More precisely, there are two spin rules, so that two electrons in a single orbital can't both be spinning one way and not the other way. But after that we cannot put any more there. We have to put others in another place, and that is the real reason that matter has volume. If we could put all the electrons in the same place, it would have no volume—more than it does. It is the fact that the electrons cannot all get on top of each other that makes bodies and everything else solid.

Obviously, in order to understand the properties of matter, we will have to use quantum mechanics and not be satisfied with classical mechanics.

38-3 Energy levels

We have talked about the atom in its lowest possible energy condition, but it turns out that the electron can do other things. It can, spiral and wiggle at a more energetic minimum, and so there are many of the low possible minima for the atom. According to quantum mechanics, in a stationary condition, there can only be definite energies for an atom. We make a diagram (Fig. 38-9) in which we plot the energy vertically, and we make horizontal lines for each allowed value of the energy. When the electron is free, i.e., when its energy is positive, it can have any energy; it can be moving at any speed. Our bound energies are not arbitrary. The atom must have one of another out of the set of allowed values, and so there is Fig. 38-9.

Now let us call the allowed values of the energy E_1, E_2, E_3, \dots . If an atom is excited in one of these "normal states," E_1, E_2, \dots , it does not remain in that state forever. Sooner or later it drops to a lower state and radiates energy as the form of light. The frequency of the light that is emitted is determined by conservation of energy plus the quantum-mechanical understanding that the frequency of the light is related to the energy of the light by (38.1). Therefore the frequency of the light that is released in a transition from energy E_3 to energy E_1 (for example) is

$$\omega_{31} = (E_3 - E_1)/\hbar. \quad (38.11)$$

This then, is a characteristic frequency of the atom and defines a spectral emission line. Another possible transition would be from E_3 to E_2 . That would have a different frequency

$$\omega_{32} = (E_3 - E_2)/\hbar. \quad (38.12)$$

Another possibility is that the atom were excited to the state E_2 , it could drop to the ground state E_1 , emitting a similar kind of frequency

$$\omega_{21} = (E_2 - E_1)/\hbar. \quad (38.13)$$

The reason we bring up these transitions is to point out an interesting relationship. It is easy to see from (38.11), (38.12) and (38.13) that

$$\omega_{32} = \omega_{31} - \omega_{21}. \quad (38.14)$$

In general, if we find two spectral lines, we should expect to find another line at the sum of the frequencies (or the difference of the frequencies), and that all the lines can be understood by finding a series of levels such that every line corresponds to the difference in energy of some pair of levels. This remarkable correspondence



Fig. 38-9. Energy diagram for an atom showing several normal conditions.

spectral frequencies are valid before quantum mechanics was discovered, and it is called the *Ritz combination principle*. This is again a mystery from the point of view of classical mechanics. Let us note besides the point that classical mechanics is a failure in its attempt, during two centuries to have determinism, but pretty well.

We have already noted above: quantum mechanics as being represented by amplitudes which behave like waves, with certain frequencies and wave numbers. Let us observe now it comes about from the point of view of amplitude that the atom has definite energy states. This is something we cannot understand at all. etc. This has been said so far, but we recall familiar with the fact that classical waves have definite frequencies. For instance, if sound is confined to the organ pipe or acoustics, like that, then there is a note there, or a note there, or a note there, but it can't have very high or very low frequency. Then an organ in which the waves are confined has certain resonance frequencies. It is therefore a property of waves in a confined space - a cavity, which we will discuss in detail will further tell us that they exist only at certain frequencies. And once the general relation exists between frequencies of the amplitude and energy, we are not surprised to find definite energies associated with the quantum of energy.

28-6 Philosophical implications.

Let us consider briefly some philosophical implications of quantum mechanics. As we know, there are two aspects of the problem: one is the philosophical implications for physics, and the other is the extrapolation of philosophical methods to other fields. The philosophical ideas associated with science are extrapolated into another field, they are usually completely altered. Therefore we shall confine our remarks as much as possible to physics itself.

First of all, the most interesting aspect is the idea of the uncertainty principle; making an observation : *Don't you phenomena?* It has always been known that making observations affects a phenomena, but the point is that the effect cannot be disregarded or minimized in decreased arbitrarily by decreasing the apparatus. When we look for something to phenomena we cannot help but disturb it in a certain minimum way, and the disturbance is necessary for the existence of the observer. The observer, we sometimes, is present in quantum mechanics, but only in a rather trivial sense. Our problem has been resolved if it is really in a sense that there is nobody there to hear it. Does it make a noise? And let's follow our old forest makes a sound, of course, even if nobody is there. Even if nobody is there, to hear it, there are other leaves left. The sound will change some leaves, and if we were careful enough we might find somewhere just some don't had rubbed against leaf and made a tiny scratch that could not be registered unless we examined the leaf with a microscope. So in a certain sense we would have to admit that there is a sound made. We might ask: Was there a sound of sound? No, musicians have to do, presumably, with a mechanism. And what is music and consciousness and whether there was intention before, or whether the tree was conscious, we do not know. Let us leave the problem at that level.

Another thing that people have emphasized since quantum mechanics was developed is the idea that we cannot not speak about those things which we cannot measure. Actually relativistic theory does not say that. If something can be defined by measurement, it is no problem in a theory. And since an exact definition of the mechanics of a localized particle cannot be defined by measurement, it therefore has no place in the theory. The idea that this is what we must do, with classical theory or with the quantum theory, is a complete analysis of the situation. But this means cannot measure position and momentum precisely does not a word mean that we cannot talk about them. To only means that we need not talk about them. The situation is, in the science is this: A concept or an idea which comes by measurement or cannot be reduced directly to experiment not or may not be useful. It need not exist in a theory. In other words, suppose we compare the classical theory of the world with the quantum theory of the world, and suppose that it is found exponentially that we can measure position and momentum only imperfectly. The question is whether the idea of the exact position of a particle and the exact

inclusion of a causal or not. The classical theory admits the idea; the quantum theory does not. This does not in itself mean that classical physics is wrong. When the new quantum mechanics was discovered, most classicists— which included everybody except Heisenberg, Bohr, Dirac, and von Neumann— said: "Look, you know it is not easy to understand quantum mechanics, but nevertheless certain quantities like ψ are the exact position of a particle, which tells about its properties, and state others." Heisenberg's answer was: "I do not need to answer such questions as what is your name, and such a question I can't answer." It is that we do not have to. Consider now theories (a) and (b); they are not necessarily incorrect by classical theory but which is used at the analysis, and the either (b), does not contain the idea of "They always act in their predictions, one could see them, but (b) is false because it cannot explain this idea, that is in (a), because that idea is one of the things that cannot be checked directly. I, & always want to know what ideas cannot be checked directly, but it is not necessary to remove them. From a & b it is clear that we can pursue course completely by using only those concepts whose validity is tested at experiment...

In quantum mechanics the "there is a wave function amplitude, there is a potential, and there are many constants that we cannot measure directly. The basic idea of determinism is probably to predict what will happen in an experimental situation and has been known since Newton: How can we do that? By assuming that we know what is there, independent of the experiment. We must extrapolate the experiments to a region where they have not been done. We must take our knowledge & extend them to places where they have not yet been checked. If we do not do that, we have no prediction. So it was perfectly sensible for the classical physicists to go rapidly along and suppose that "in your room—which obviously has something for a base—... means something else for an electron. It was not stupidity. It was sensible prediction. Today we say that the law of causality is supposed to be true in all energies, but everyday situation may contradict and say how stupid we were. We do not know where we are "stupid" until we "turn over next card" and the whole idea is to put our neck out. And the only way to find out that we are wrong is to find out if we can predict something. That is already necessary to make predictions.

We have already made a few remarks about the indeterminacy of quantum mechanics. That is, that we are unable now to predict what will happen in physics in a given physical situation and what is true as far as exactly as possible. If we drop an atom that is in an excited state and it is going down it is stable. We cannot say when it will emit the photon. It has a certain probability to emit the photon every time, and we can predict only a probability for emission; we cannot predict the precise energy. This has given rise to all kinds of metaphysical questions on the meaning of free will, etc. still, and of the other, that the world is uncertain.

If you ask me, I am notimplausible that classical physics is also indeterminate, in a sense. It is usually thought that this indeterminacy, i.e., we cannot predict the future, is an important quantum-mechanical thing, and this is said to explain the behavior of the mind, feelings of free will, etc. But if the world were classical—if the laws of mechanics were classical—it is not very obvious then the mind would not feel more or less the same. It is true classically that if we knew the position and the velocity of every particle in the world, then, of course, we could predict exactly what would happen. And therefore the classical world is deterministic. Suppose, however, that we have a finite account of the world, but we do not know exactly where just one atom is, say to one part in a billion. Then as it goes along it hits another atom, and because we did not know the position better than to one part in a billion, we hit an even larger error in the position after "one" step. And that is something, of course, i.e., the next collision, so that if we start with only a tiny error, a rapidly magnifies this tiny error exponentially. To give an example: if water falls over a dam, it splashes. If we want exactly, every tiny sand grain a drop will land at our nose. This appears to be completely random, yet such a behavior would be predicted by purely classical laws. The exact position of all the drops depends upon the precise positions of the water before it goes over the dam. Now? The water drops, it's this we recognized in falling, ... that we get complete randomness. Otherwise,

viously, we cannot really predict the position of the drops unless we know the motion of the water molecules exactly.

Speaking more precisely, given an arbitrary accuracy, no matter how precise, one can find a time long enough that the exact predictions will be valid long enough. Now the point is that this length of time is not very large. It is not that the time is millions of years; the accuracy is one part in a billion. The time goes, in fact, only logarithmically with the error, and it turns out that, in only a very, very tiny time we lose all the information. If the accuracy is taken to be one part in billions and billions and billions—no matter how many billions we take provided we do stop somewhere—that we can find a time less than the time it took to state the accuracy—after which we can no longer predict what is going to happen! It is therefore not free to say that from the apparent freedom and indeterminacy of the laws of mind, we should have realized that classical "deterministic" physics could not ever hope to understand it, and to welcome quantum mechanics as a release from a "completely mechanistic" universe. For already, in Classical mechanics there was indeterminacy from a practical point of view.

The Kinetic Theory of Gases

39-1 Properties of matter

With this chapter we begin our study of objects which we occupy in for some time. It is the first part of the analysis of the properties of matter from a physical point of view, in which, recognizing that the gas is made up of a great many atoms, or elementary parts, which interact *exactly* according to the laws of mechanics, we try to understand why various aggregates of atoms behave the way they do.

This is a new and this is a difficult subject, and we emphasize at the beginning that it is in fact a extremely difficult subject, and the we have to deal with it differently than you have dealt with the other subjects so far. In the case of mechanics and in the case of light, we were able to begin with a simple statement of some law, say Newton's law, or the formula for the field produced by an accelerated charge, laws which a whole host of phenomena could be essentially understood, and which would provide a basis for our understanding of mechanics and of light from that time on. That is, we must learn more law, and we do not learn different physics, we only learn better methods of mathematical analysis related to the situation.

We cannot use this approach effectively at studying the properties of matter. We can discuss matter only in a most elementary way; it is much too complicated a subject to analyze directly from its specific basic laws, which are more clutter than the laws of mechanics and electricity. But there are a bit too many for every one the properties we wish to study, and too many ways to put them. Newton's laws to the properties of matter, and these steps are, in themselves, fairly complicated. We will now start to take some of these steps, but the way of our analysis will be quite accurate, but not eventually get back to the results. We will have only a rough understanding of the properties of matter.

One of the reasons can we have to perform the analysis so imperfectly is that the mathematics of it requires a deep understanding of the theory of probability: we are not going to want to know what every atom is actually moving, but rather, how many move here and there on the average, and what the odds are for different effects. So this subject involves a knowledge of the theory of probability, and our inabilities is not you, you may really not we don't want to start from here.

Secondly, and more important from a physical standpoint, the actual behavior of the atoms is not according to classical mechanics, but according to quantum mechanics and a correct understanding of the subject cannot be gained until we understand quantum mechanics. Here, in the millions of billions, bulk and poly-nuclei, the difference between the classical mechanical laws and the quantum mechanical laws is very important and very significant, so that many things that we are deduced by classical physics will be fundamentally incorrect. Therefore there will be certain things to be possibly unlearned, however, we shall indicate in every case when a result is incorrect, so as we will know where the "edge" are. One of the reasons for discussing quantum mechanics in the preceding chapters was to get an idea as to why, in fact, less classical mechanics is incorrect in the real world.

Why do we need with the subject now to start? Why not wait a half a year, or a year, until we cover the mechanics of a building tower, and we learn a little quantum mechanics, and then we can do it in a more fundamental way? The answer is that it is a different subject and the best way to learn it is to do it slowly! The first thing to do is to get some idea, more or less, of what ought to happen. If

39-2 Properties of matter

39-3 The pressure of a gas

39-4 Compressibility of radiation

39-5 Temperature and kinetic energy

39-6 The heat equation

different circumstances, and then, later, when we know the laws better, we will formulate them better.

Anytime we want to analyze the properties of matter in a real problem, we might want to start by writing down the fundamental equations and then try to solve them mathematically. Although there are people who do try to do such an approach, these people are the failures in the test; the two, however, come in those who can, from a physical point of view, people who have enough idea where they are going and then begin by making certain kind of approximations, knowing what is big and what is small in a given complicated situation. This problem—so uncomplicated now even an elementary student could follow, although incomplete and incomplete, is something like that, and so the subject will be one that we shall go over again and again, each time with more and more accuracy, as we go through our course in physics.

A good reason for beginning the subject right now is that we have already used many of these ideas in, for example, chemistry, etc., the other four hours of work at least in high school. This is enabling us to use the physical basis for these things.

As an interesting example, we all know that equal volumes of gases at the same pressure and temperature, contain the same number of molecules. The law of multiple proportions, that when two gases combine in a chemical reaction the volumes needed always exist in simple integral proportions, was understood ultimately by Avogadro to mean that equal volumes have equal numbers of atoms. Now why do they have equal numbers of atoms? Can we deduce from Newton's laws that the number of atoms should be equal? We shall address ourselves to that specific matter in this chapter. In succeeding chapters, we shall discuss various other phenomena involving pressure, volume, temperature, and heat.

We have also felt that the subject can be attacked from a mathematical point of view, and that there are many interesting aspects of the group that we can attack. That treatment, when we compare something, it reveals, if we need it, it rewards. There is a relationship between these two facts which must be deduced independently of the machinery one needs. This subject is called thermodynamics. The deepest understanding of thermodynamics comes, of course, from understanding the actual machinery underlying, and that is what we shall do; we shall take the most simple point from the beginning and use it to understand the various properties of matter and the laws of thermodynamics.

Let us, then, discuss the properties of gases from the standpoint of Newton's laws of motion, etc.

34-2 The pressure of a gas

First, we know that a gas exerts a pressure, and we might clearly understand what this is due to. If our ears were a few times more sensitive, we would hear a perpetual rushing noise. Evolution has not developed the ear to this point because it would be useless if it were so much more sensitive—we would hear a perpetual roar. The reason is that the ear is in contact with the air, and air is a lot of molecules in perpetual motion and these bump against the surfaces in bumping against the ear-membrane. Let us imagine, let us, boom, boom, boom... what we do not hear because the atoms are so small that the sensitivity of the ear is not quite enough to notice it. The result of this perpetual bumping is to push the drum away, but since nothing is stopped perfectly, reboundance follows so the side of the ear-membrane, so the net force on it is zero. If we were to take the ear away from one side, or change the relative amounts of air on the two sides, the ear-membrane would then be pushed one way or the other, because the amount of reboundance on one side would be greater than on the other. We sometimes find the unexplainable effect when we go up too fast in an airplane, especially if we also have the cold (when we have a cold, inflammation above the ear which can lets the air on the inside of the ear-membrane with outside air through the throat, so that the two pressures cannot readily equilibrate).

In working out to analyze the situation quantitatively, we imagine that we have a volume of gas in a box, at one end of which is a piston which can be moved (Fig. 39-1). We want to find out what happens if the piston moves fast enough to knock atoms in this box. The atoms of the gas will tend to hit the piston more often and move it back with certain velocities they bring against the piston. Suppose there is a atom, m , on the x -axis of the piston. What will happen if the piston were to advance, and velocity v_0 onto it, each time it got stopped it would pick up a little momentum and it would gradually get kicked out of the box. So in order to keep it from being pushed out of the box, we have to hold it with a force F . The question is how much force? One way of expressing this force is to talk about the force per unit area, which is the area of the piston. Then the force on the piston will be written as a number times the area. We define the pressure, then, as equal to the force that we have to apply on a piston, divided by the area of the piston.

$$P = F/A. \quad (39.1)$$

To make sure we understand the idea let's have us derive it for accuracy purposes anyway; the differential work done on the gas in compressing it by moving the piston in a differential amount dx would be the force times the distance that we compress it, which, worked up as (39.1), would be the pressure times the area times the distance, which is equal to transfer the pressure into the change in the volume:

$$dW = F dx = -F dx = -P dV. \quad (39.2)$$

(The last d has the dimensionality of the volume changed.) The minus sign is there because, as we are pressing x , we decrease the volume; if we think about it we can see that if a gas is compressed, work is done on it.

How much does do we have to apply to balance the bunching of the molecules? The piston moves from rest collides a certain amount of time Δt . A certain amount of atoms — say n — will move into the piston, and it will start to move. To keep it from moving, we must push back into it the same amount of momentum per second from our force. Of course, the force is the amount of momentum per second that we are applying. There is another way. In addition to the force of the piston it will pick up speed because of the bump it makes; with each collision we get a little more speed, and this adds thus increases. The rate at which the piston picks up speed, or accelerates, is proportional to the force on it. So we see that the force which we already have said is the pressure times the area is equal to the momentum per second delivered to the piston by the colliding molecules.

To calculate the momentum per second we may — we can — do it in two parts: first, we find the momentum delivered to the piston by one particle, and then, if collision with the piston then we have to multiply by the number of collisions per second that the atoms have with the wall. The force will be the product of these two factors. Now let us see what the two factors are: In the first place we shall suppose that the piston is a perfect "mirror" for the atoms. If it is not, the whole theory is wrong, and the piston will start to heat up and things will change, but eventually, when eqilibrium is set in, the net result is that the collisions are effectively perfectly elastic. On the average, every particle that comes in bounces with the same energy. Since we'll imagine that the gas is in a steady condition, and we lose no energy to the piston because the piston is standing still. In other circumstances, if a particle comes in with a certain speed, it bounces off with the same speed and, you see, key, with the same mass.

If v_0 is the velocity of an atom, and v_1 is the resultant v , then v_1 is the component of the momentum "lost"; but we also have an equal component of momentum " v_0 " and so the total momentum, delivered to the piston, by the n atoms in one collision, is mv_{1x} , because it is "lost."

Now we need the number of collisions made by the atoms in a second, or in a certain number of time Δt , but we divide by Δt : How many atoms are hit? This is a question that does not fit easily to the volume A , in $m = N/V$ is each cubic volume. By how many atoms hit the piston, we note that, given a certain



Fig. 39-1. Action of a gas in a box with a Mc力学斯 piston.

amongst x -values, if a particle has a certain velocity towards the piston it will hit it in the time t , provided it is close enough. If it is too far away, it goes only part way toward the piston; if the time t but does not reach the piston. The above is clear that only those molecules which are within a distance $v_0 t$ from the piston are going to hit the piston in the time t . Thus the number of collisions in a time t is equal to the number of atoms which are in the region within a distance $v_0 t$, and since the area of the piston is A , the volume occupied by the atoms which are going within the piston is $A v_0 t$. But the number of atoms per unit volume is n_{atom} . Of course we do not care about whether that is in a time t , we want the number per second, so we divide by the time t to get n_{atom} . (This time t might be made very short, if we feel we want to take into account molecular velocities, but it is however, being.)

So we find that the force is

$$F = n_{\text{atom}} \cdot A v_0^2. \quad (38.3)$$

See, the force is proportional to the area, if we keep the particle density fixed or we change it enough. The pressure is then

$$P = A v_0^2. \quad (38.4)$$

Now we notice a little trouble with this analysis: first, all the molecules do not have the same velocity, and they do not move in the same direction. So, all the v_i 's are different! So what we must do of course is to take the average of the v_i 's, since each one makes its own contribution. What we want is the square of v_i , averaged over all the molecules:

$$\bar{v}^2 = \langle v_i^2 \rangle. \quad (38.5)$$

But we forgot to include the factor 2π . No. of all the atoms, only half are located toward the piston. The other half are located in another way, and there take (\bar{v}) . We are averaging the negative v_i 's squared, as well as the positive v_i 's. So when we just take $\langle v_i^2 \rangle$, without looking, we are putting in twice as much as we want. The average of v_i^2 , the positive v_i , is equal to the average of v_i^2 for all v_i , taking one half.

Now as the terms become small, it is clear that there is nothing special about the "x-direction", the atoms may also be moving up and down, back and forth, left and right. Therefore, it is going to be true that $\langle \bar{v}^2 \rangle$, the average motion of the atoms in one direction and the average in the other two directions, are all going to be equal.

$$\langle \bar{v}_x^2 \rangle = \langle \bar{v}_y^2 \rangle = \langle \bar{v}_z^2 \rangle. \quad (38.6)$$

It is only a matter of either a very mathematical intuition, therefore, that they are each equal to one-third of the sum, which is of course the square of the mean value of the velocity:

$$\langle \bar{v}^2 \rangle = \langle \bar{v}_x^2 \rangle = \langle \bar{v}_y^2 \rangle = \langle \bar{v}_z^2 \rangle = \langle \bar{v}^2 \rangle / 3. \quad (38.7)$$

This has the advantage that we do not have to worry about any particular direction, and so we write our pressure in a more compact form in this form.

$$P = (f/m)v^2/2. \quad (38.8)$$

In whom we wrote the last factor as $(f/m)^2/2$; f is the *kinetic energy* of the center-of-mass motion of the molecule. We find, therefore, that

$$PV = (f/m)v^2/2. \quad (38.9)$$

With this equation we can calculate how much the pressure is, if we know the species.

As a very simple example let us take helium gas, or any other gas like mercury vapor, or potassium vapor, at high enough temperature, or argon, in which all the molecules are single atoms, for which we may suppose that there is no collision.

motion in the atom. It has had a complex motion. There might be some internal motion, mixed with others, or something. We suppose that we may disregard this; this is actually a serious matter that we have to come back to, but it turns out to be all right. We suppose that the internal motion of the atoms can be disregarded, and therefore for this purpose that the kinetic energy of the center-of-mass motion is all the energy there is. So for a monoatomic gas, the kinetic energy is the total energy. In fact, if we are going to add up the total energy, it is immediately realized that this internal energy may differ very, since there is no external energy to a gas, i.e., all the energy of all the molecules is the sum of the others, whatever it is.

For a monatomic gas we will suppose that the total energy E is equal to one-half of the average kinetic energy of mV^2 , because without distinguishing any particular of rotation or vibration motions from the others. Then, in these circumstances, we would have

$$PV = \frac{3}{2}E. \quad (39.10)$$

Incidentally, we can stop here and find the answer to the following question. Suppose that we take a can of gas and do nothing; the gas slowly loses much pressure; do we still dissipate the same amount? This is to ask what does the pressure loss of the energy divided by P . As we see now, when we work out the gas and we thereby increase the energy E , we must be willing to have some kind of a differential equation: "If we start in a given circumstance with a certain energy and a certain volume, we then know the pressure." Now we start to suppose, that the important we do, the energy E increases and the volume V decreases, or the pressure goes down.

So, we have to solve a differential equation, and we will solve it in a moment. We don't just emphasize, however, that as we are compressing this gas, we are dissipating that all the work goes into increasing the energy of the atoms inside. We may say, "Isn't that necessary? Where else could it go?" It turns out that it can go another place. There are what we call "heat leaks" through the walls; the heat (i.e., fast-moving) atoms can penetrate the walls, heat the walls, and energy goes away. We shall suppose for the present that this is not the case.

For somewhat more generality, although we are still not in some very special assumptions about our gas, we shall write, for $\delta E = \delta U + \delta T$

$$PV + \delta U = TdV. \quad (39.11)$$

It is exactly $(P - 1) dU + TdV = 0$, conventional results, because we will deal with a few other cases later where the number in front of V will not be $\frac{1}{2}$ but will be a different number. So, in order to do the thing in general, we call it $\gamma = 1$, because people have been calling it that. We already one hundred years ago. This γ , then, is $\frac{1}{2}$, because $\gamma = \frac{1}{2}$ for a monoatomic gas like helium.

We have already noticed that when we compress a gas the work done is PdV . A compression in which there is no heat energy added or removed is called an adiabatic compression, from the Greek α (not) + $\delta\mu\eta$ (through) + $\nu\mu\nu$ (to go). (The word adiabatic is used in physics in several ways, and it is sometimes hard to see what is meant in each theory. That is, for an adiabatic compression all the work done goes into changing the internal energy. That is the essential; there are no other uses of energy.) So, then we have $PdV = -\delta U$. But since $\gamma = PV/\delta U = 1$, we may write

$$\delta U = (PdV - PVdP)/(\gamma - 1). \quad (39.12)$$

So, we have $PdV = (PVdP - \delta U)/(\gamma - 1)$, or, rearranging the terms, $\gamma PdV = -\delta U$.

$$\gamma PdV/PV + \delta U/V^2 = 0. \quad (39.13)$$

Incidentally, assuming that γ is constant, as it is for a monoatomic gas, we can integrate this. It gives $\gamma \ln P - \ln V = C$, where C is the constant of integration. If we take the exponential of both sides, we get two new

$$PV^\gamma = C \text{ (constant).} \quad (39.14)$$

In other words, under adiabatic conditions, where the temperature rises as we compress because no heat is going out, the pressure is proportional to the β power of a constant for a monoatomic gas. Although we derived it theoretically, this β , in fact, is the way monoatomic gases behave experimentally.

24.3 Compressibility of photons

We may give one other example of the kinetic theory of γ gas, one which is not used in everyday science but is useful for science. We take a large number of photons in a box in which the temperature is very high. (The box, of course, is not in a very hot star. The star is not hot enough; these are still too many atoms, but at all higher temperatures in certain very hot stars, we may not use the kinetic model, because the only objects that we have in the box are photons.) Now then, a photon has a certain momentum p . (We always find that we can interpret variables in the kinetic theory; p is the momentum, m is the mass-momentum; v is the velocity, h is the velocity. T is the temperature; but V is the kinetic energy or the time or the energy; sometimes both can't be used nicely. This p is momentum, it's a vector. Going through the same analysis as before, π is the component of the vector p in the direction in which is given in the kick. Thus $\Delta\pi$ indicates Δv , and in evaluating the number of collisions, π is still $\sim \pi_0$ when we get all the way through, as this last term in Eq. (24.11) is listed.

$$\langle \pi \rangle = 2\pi_0 v_0 n. \quad (24.14)$$

Then, in this averaging, we average out the average of $\pi_0 v_0$ (the same factor of β will, finally, pulling in the other two directions, we find

$$PV = Np \cdot v_0^2 / 3. \quad (24.15)$$

This agrees with the formula (24.9), because the number N is now nV ; it is a little more general, that is all. The pressure times the volume is the total number of photons times $(p \cdot v)^2$ averaged.

Now, for photons, what is $p \cdot v$? The momentum and the velocity are in the same direction, and the velocity is the speed of light, so this is the momentum of each of the photons times the speed of light. The momentum times the speed is E , since physics is its energy. $E = pc$, so these terms are the energies of each of the photons, and we should, if we can, take an average energy, times the number of photons. So we have π of the energy, inside the $\langle \rangle$:

$$PV = Nc^2 \langle p c \rangle n \langle pc \rangle \langle pc \rangle \quad (24.16)$$

For photons this gives us zero, in front of $\langle \rangle$ in (24.16); $\langle pc \rangle$, or $\langle \rangle = 0$, and we have discovered that radiation in a box obeys the law:

$$PV^{1/2} = C. \quad (24.17)$$

So we know the compressibility of radiation! That is what "use" is in analysis of the contributions of radiation pressure to a star, but is how we calculate it until there is changes when we compare it. What wonderful things are already within our power!

24.4 Temperature and thermal energy

So far we have not dealt with temperature; we have purposely been avoiding the temperature. As we compare a gas, we know that the square of the momentum increases, and we are not saying that the gas is \propto hotter; we would like to understand whether this is due to the temperature. Is we try to do the experiments, not adiabatically but at what we call constant temperature, whatever it is? We know that two like boxes of gas and the their will move to each other long enough that π at the start they were at what we call different temperatures, they

all at the end come to the same temperature. Now what does this mean? That means that they get to a condition that they would get to if we left the tube long enough. What we mean by this temperature is just that—the final condition when things have been striking each other long enough.

Let us consider now what happens if we have two gases in a container separated by a movable piston as in Fig. 10-2; and for simplicity we shall take two monoatomic gases, say helium and neon. To summarize (by the atoms have mass m_1 , velocity v_1 , and there are n_1 atoms) and others, n_2 is the other containing the atoms have mass m_2 , velocity v_2 , there are n_2 atoms per unit volume. What are the conditions for equilibrium?

Obviously, the bombardment from the left side must be such that it causes the piston to move right and compresses the other gas until its pressure builds up, and the thing will then stand back and forth, and will probably come to rest at a place where the pressures are equal on both sides. So we can assume that the pressures are equal when just atoms from the initial sample get time to escape the piston, in fact the number n times the average kinetic energy on each side are equal. What we have to try to prove, however, is that the average velocities are equal. So first of all we know is that the number times the kinetic energies is equal.

$$n_1 m_1 v_1^2 / 2 = n_2 m_2 v_2^2 / 2,$$

from (10-8), because the pressure is the equal. We must realize that this is not the only equilibrium over the long run, but something else may happen more slowly as the true complete equilibrium corresponding to equal temperatures sets in.

To see the idea, suppose that the pressure on the left side was developed by having a very light piston but a low velocity. By having a very small and a small n we can get a certain pressure by having a small and a large v . Let atoms may be moving slowly but be packed much more closely, or them may believe that they are hitting harder. Will it stay like that? No, I think we might think so, but the two ends swap and find we have forgotten one important point. That is, that the intermediate piston does not receive a steady pressure; it jiggles just like the piston, but we were just talking about it, because the language is not clear of their uniform. There is not a general, steady pressure but a source of the pressure which makes the thing jiggles. Suppose that the atoms on the right side are not jiggling much, but those on the left are fast and far between and very energetic. The piston will, now and then, get a big impulse from the left, and will be driven against the slow atoms on the right, giving them some speed. (As each atom collides with the piston, it either gains or loses energy, depending upon whether the piston is moving one way or the other when the atom strikes it.) As a result of the collisions, the piston finds itself jiggling, jiggling, jiggling, and this shakes the other gas—it gives impetus to the other atoms and they build up faster movement, until they balance the jiggling and the piston is going to them. The system settles to some equilibrium where the piston is moving at a uniform consequent speed that picks up energy from the atoms at about the same rate as it puts energy back into them. So the piston picks up a certain mean, reasonably in speed, and it is our problem to find it. What we do that is, we consider one problem better, because the two will adjust their velocities—just like the rate at which they are taking in power energy is a even other through the piston will become equal.

It is quite difficult to figure out the details of the piston in this particular circumstance, although it is readily simple to understand, it is not so easy to be a little harder to analyze. Before we analyze this, it is another deeper problem to when we have a gas of gas but now we have two different kinds of molecules in it, having masses m_1 and m_2 , velocities v_1 and v_2 , and so forth; there is no a much more intricate calculation. If all of the N_1 molecules, like the scattering off oil, the molecules is not going to last, because they get located by the N_2 molecules and to pick up speed. If they are not jiggling much faster than the N_2 molecules, then maybe that will not last either—they will pass the energy back to the N_2 molecules. So the both gases under the same law. The problem is to find the next and determine the the speed of the two

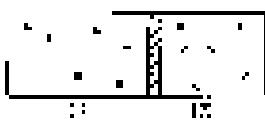


Fig. 39-2. Atoms of two different monoatomic gases are separated by a movable piston.

This is still a very difficult problem, but we will solve it as follows. First we examine the following subproblem: again this is one of those cases where however simple the derivation is, the result is very simple to remember, but the derivation is not. I hope you'll suppose that, at least, two molecules, of different mass, are moving, and that the collision is viewed on the center-of-mass (CM) system. In order to analyze a simplification, we look at the collision in the CM. As we know from the laws of classical mechanics, the conservation of momentum and energy, after the molecules collide, the only way they can move is such that both molecules do not change speed—only they just change their direction. So we have a "circular" collision that looks like that in Fig. 39-1. Suppose, for a moment, that we want to see collisions with the CM at rest. Suppose we're going fast, they are all initially moving horizontally. Of course, at the first collision some of them are moving at an angle. In order words, if they were all going horizontally, then at least some would have to be moving vertically. Now, in such a collision, they would be coming in from another direction, and then they would be forced to still cross the angle. So even if they were randomly organized in the beginning, they would get sprayed around at all angles, and that is why they would be sprayed some more, and sprayed some more, and sprayed some more. Ultimately, what will be the distribution? Answer: it will be equally likely to find any particle moving in any direction in space. After all, since each molecule changes its direction,

They are equally likely to go in $\pi/2$ dimensions, but how do we see this? This is of course an likelihood that they will go at this particular angle; but, because a specific direction is chosen, so we have to talk about *not* "correcting". The fact is that any atom on a sphere centered at a collision point will have many, many molecules going through it as go through any other point on the sphere. So the result of the collisions will be to distribute in directions so that equal areas on a sphere will have equal probability.

Indeed, if we just want to discuss the original dimension and some other direction at angle θ from it, it is clear, intuitively, that the differential area of a sphere of unit radius is $d\Omega = \sin \theta d\theta 2\pi$, and that is the same as the differential of $\cos \theta$. Another reason is that the cosine of the angle between any two directions is equally likely to be anything from -1 to $+1$.

Next, we have to worry about the actual case, where we do not have the collision in the CM system, but we have two atoms which are moving relative with velocities v_1 and v_2 . What happens now? We can analyze this collision with the vector velocities \mathbf{r}_1 and \mathbf{r}_2 in the following way: We can say that there is a certain CM; the velocity of the CM is given by the "average" velocity, which is proportional to the masses, so the velocity of the CM is $v_{CM} = m_1 v_1 + m_2 v_2 / (m_1 + m_2)$. If we have this collision in the CM system, then we have a collision just like that in Fig. 39-1, with a certain relative velocity v equal to $|v|$. The relative velocity is just $v_1 - v_2$. Now consider a real, live CM frame, moving, and in the CM frame v is a relative velocity, v , and the molecule rounds and rounds off in some new direction. All this happens while the CM keeps right on moving without any change.

Now then, what is the distribution resulting from this? From our previous argument we conclude this: for an equilibrating, all directions for v are equally likely, relative to the direction of the motion of the CM.¹ There will be no particular orientation, in the end, between the direction of the motion of the system to the relative velocity, and that is the motion of the CM. Of course, if there were, the collisions would scatter it about, and v is all sprayed around. So the cosine of the angle between v and v_{CM} is zero on the average. That is,

$$\langle \cos \theta_{CM} \rangle = 0. \quad (39-19)$$

¹ This argument, which was not discussed by Maxwell, involves some subtleties. Although he concluded correctly the results does not follow purely from the consideration of symmetry that we used before class, by going to a relative frame moving enough, we may find a different velocity distribution. We can not furnish a simple proof of this result.

But in $\langle \tau_1 \rangle$ we have expressed $\langle v_1^2 \rangle$ in terms of v_1 and v_2 or m_1 :

$$\begin{aligned} \langle v_1^2 \rangle_{\text{ave}} &= \frac{(m_1 v_1) \cdot (v_1 m_1 + m_2 v_2)}{m_1 + m_2} \\ &= \frac{m_1 v_1^2}{m_1 + m_2} + (v_2 - \bar{v}_1) \bar{v}_1 \cdot v_2. \quad (29.20) \end{aligned}$$

Let's go back to $\langle \tau_1 \rangle$ and $\langle \tau_2 \rangle$, what is the average of v_1^2 ? That is, what is the average of the square of velocity of a molecule in the direction of another? Such there is just equal likelihood of finding an atom moving one way or another. The average of the velocities is my definition now. Consider then, is the direction of v_1 , v_2 has zero energy. So, the average of v_1^2 , v_2^2 is zero. Therefore we conclude that the average of v_1^2 must be equal to the average of $m_1 v_1^2$. Then & the average kinetic energy of the molecules must be equal:

$$\langle m_1 v_1^2 \rangle = \langle m_2 v_2^2 \rangle \quad (29.21)$$

If we have two kinds of atoms in a gas, it can be shown, and we probably do have shown it, that the average of the kinetic energy of one is the same as the average of the kinetic energy of the other. When they are both in the same gas so the system has no equilibrium. That means the heavy ones will move slower than the light ones, this is easily shown by experimenting with "atoms" of different masses in air. Crash!

Now we could go on saying "In fact, we say that if we have two different gases separated in a box, they will have equal average kinetic energy when they have finally come to equilibrium & even though they are not in the same box. We can make the argument in a number of ways. One way is to argue that if we have a fixed container w/1. a tiny hole in it (Fig. 29.4) so that one gas could leak out through the holes while the other would not, because the molecules are not big, and I know that in undisturbed equilibrium, the molecules that are free, where they are in fact, they have lower average kinetic energy, but could come through the hole without loss of kinetic energy, so the average kinetic energy in the pure gas and in the mixture must be the same. This is not too satisfactory, because maybe there are no holes, for the kind of selection, the separate one from the other.

So we have to go back to the piston problem. We can make an argument which shows that the kinetic energy of the piston must also be fixed. And also, that would be the kinetic energy due to the pure gas horizontal motion of the piston. So, if you push it up and down motion, it will have to be the sum of $\langle m_1 v_1^2 \rangle$. Likewise, from the equilibrium on the other side, we can prove that the kinetic energy of the piston is $\langle m_2 v_2^2 \rangle$. Although this is not in the middle of the gas, but is on one side of the gas, we can still make the argument, although it is a little more difficult, that the average kinetic energy of the system, and of the gas molecules are equal as a result of all the collisions.

If this still does not satisfy us, you may make an artificial example by which the equilibrium is perturbed by an object which can be hit on all sides. Suppose that you're a short rod with a ball at each end sticking through the piston, or a billiard ball gliding along at your. This ball is much like one of the molecules, and can be hit on all sides. This whole object has a certain total mass, m . Now, we have the piston which has mass m_1 and mass m_2 as before. As a result of the collisions, in the analysis that we made before, is that the kinetic energy of m because of collisions with the molecules on one side, must be kept for the average. Likewise, because of the collisions with molecules on the other side, it has to be $\langle m_2 v_2^2 \rangle$ in the average. So, therefore, both sides have to have the same kinetic energy when they are in thermal equilibrium. So, although we only proved it for a mixture of gases, it is easily extended to the case where there are two different, separate gases at the same temperature, i.e.

Thus when we take the gases at the same temperature, the mean kinetic energy of the CM motion is equal.

The "real" molecular kinetic energy is a property only of the macroscopic gas." Being a property of the "macroscopic," not one of the gas, we can make it as a function of the temperature. The mean kinetic energy of a molecule is thus some



Fig. 29.2 Two gases in a box with a movable piston.

function of the temperature. But we'd like⁷ to what scale to use for the temperature? We may arbitrarily choose K as our temperature +, but the mean energy \bar{E} is linearly proportional to the temperature. The best way to do it would be to call the mean energy itself "the temperature". This would be the simplest possible function. Unfortunately, the scale of temperature has been chosen differently, so instead of calling it temperature directly we use a constant conversion factor between the energy of a molecule and a change of absolute temperature called a *Boltzmann factor*. The constant of proportionality $k = 1.38 \times 10^{-23} \text{ joule}$ (or every degree Kelvin).⁸ So if T is absolute temperature, your definition says that the mean kinetic energy is kT . (The k is put in as a matter of convenience, so as to get rid of it somewhere else.)

We point out that the constant k is only associated with the component of motion in any particular direction, so only $\frac{1}{2}kT$. The other independent directions that are involved divide k by 3.

39.2 The Ideal gas law

Now, of course, we can put this definition of temperature into Eq. (39.9), and so find the law for the pressure of gases as a function of the temperature; it is that the pressure times the volume is equal to the total number of molecules, or the universal constant N_A , times the temperature:

$$PV = N_A kT. \quad (39.10)$$

Put in absolute temperature and pressure and volume, the number of atoms is determined, & this is a universal constant. So equal volumes of different gases, at the same pressure and temperature, have the same number of molecules, because of New ton's laws. That is an amazing conclusion!

In practice, when dealing with molecules, because the numbers are so large, the molecules have uniformly chosen a specific number, a very large number, and called it something else. They have a number which they call a mole. A mole is nearly a Avogadro number. Why they didn't choose 10^4 objects, or it could come out even, is a historical question. They happened to choose, for the convenient number of objects on which they standardize, $N_A = 6.02 \times 10^{23}$ objects, and this is called a mole of objects. So instead of measuring the number of molecules in terms of their mass in grams of molecules, in terms of N , we can write the number of moles, times the number of atoms at a mole, times kT , and if we want to, we can take the number of atoms in a mole times N_A , which is a mole's worth of k , and call it something else, and we do—what we call N . A mole's worth of k is 8.317 joules/K , so $k = N_A k = 8.317 \times 10^{23} \text{ J/K}^{-1}$. That is also just the gas law written as the number of moles (also called N) times N_A , or the number of atoms times kT .

$$PV = NWG \quad (39.11)$$

It is the same thing, just a different scale for measuring numbers. We use N as a unit, and choose the 8×10^{23} as a unit.

We now make one more small adjustment for gases, and that has to do with the law for objects other than monoatomic molecules. We know that only with the 0.02 K motion of the atoms of a monoatomic gas. What happens if there are more atoms? First, consider the case that the atom is like a helical spring, and there are forces on it. The exchange of jiggling motion between atoms will pass on, and momenta does not depend on where the system is at that instant, of course. The equilibrium conditions are the same. No matter where the position is, if a speed of motion comes to a gas fast at some energy, all the molecules will use

⁷ The analogous scale is just $\frac{1}{2}k$ Kelvin scale with a zero chosen at 273.16 K , or $T = 273.16 \text{ J/mole/K}$ temperature.

⁸ What the average ball molecule weighs is the mass in grams of a mole of a molecule. The molecular mass of that is N_A times the mass of one mole of isotopes 12 C , 16 O , 14 N , 17 F , and 18 Ne , so is the nuclear weight of 12 protons.

the right way. So it makes no difference about the timing. The speed at which the piston has to move in the average, is the same. Second theorem, that the center of mass of the entire energy - one direction is $\langle v \rangle$, it does whatever there are forces present or not.

Consider, for example, a diatomic molecule composed of atoms m_1 and m_2 . What we have proved is that the motion of the CM of part A and part B are not the same, $v_{CM} = \langle v \rangle$ and $v_{A,B} = \langle v \rangle T$. How can this be? It may seem odd (although they are held together, what they are spending and forcing in the air when something hits them, exchanging energy with them, the only thing that comes is how fast they are moving). That alone determines how fast they exchange energy in collisions. At the molecular level, the force is not an excentric point. Therefore the same principle is right, even when there are forces.

Let's prove, finally, that the gas law is consistent also with a disregard of the individual masses. We did not really include the internal motions before, we just treated a macromolecule as gas. We want now show that an entire object, consisting of a single body of total mass M , has a velocity of the CM such that

$$\text{Addition: } \langle v \rangle = \frac{\langle v \rangle}{M} M. \quad (4.3.1)$$

In other words, we can consider either the separate pieces or the whole thing. Let us see the reason for that. The mass of the electrons added is $m_1 + m_2 + \dots$, and the velocity of the total mass is equal to $\langle v \rangle = (m_1 + m_2 + \dots) \langle v \rangle / M$. Now we need $\langle v \rangle^2$. If we square both, we get

$$\langle v \rangle^2 = \langle v_1^2 \rangle + 2m_1 \frac{(v_1 - \langle v \rangle)^2}{M^2} + m_2 \frac{(v_2 - \langle v \rangle)^2}{M^2} + \dots$$

Now we multiply M and take the average and thus we get

$$\begin{aligned} \langle M v \rangle^2 &= \frac{\langle (M \langle v \rangle + 2m_1(M \langle v \rangle - \langle v \rangle) + m_2(M \langle v \rangle - \langle v \rangle) + \dots)^2 \rangle}{M} \\ &= \frac{3\langle v \rangle^2 + 2M \langle v \rangle (M \langle v \rangle - \langle v \rangle)}{M} \end{aligned}$$

(We have used the fact that $\langle v_1 \rangle + \langle v_2 \rangle + \dots = \langle v \rangle$). Now what is $\langle v_1 \rangle + \langle v_2 \rangle + \dots$? (It is the kinetic energy). To end out, we can for example, that the relative velocity, $v = v_1 - \langle v \rangle$, is not always likely to go in one direction than in another, but its average component in any direction is zero. This we assume then

$$\langle v_1 - \langle v \rangle \rangle = 0$$

But what is $\langle v_1^2 \rangle$? It is

$$\begin{aligned} \langle v_1^2 \rangle &= \frac{\langle (v_1 - \langle v \rangle)^2 + 2m_1(v_1 - \langle v \rangle)^2 + m_2(v_1 - \langle v \rangle)^2 + \dots \rangle}{M} \\ &= \frac{m_1^2 + (m_1 - M)(M - \langle v \rangle)^2 + m_2^2}{M^2} \end{aligned}$$

Therefore, since $\langle m_i v_i^2 \rangle = \langle m_i v_i \rangle^2$, we do not have to add up all the averages, and we are left with

$$(m_1 - M)(M - \langle v \rangle) = 0.$$

Therefore, if $m_1 = M$, we find that $(M - \langle v \rangle) = 0$ and therefore that the healthy motion of the entire molecule - regarded as a single particle of mass M , has a known average velocity, equal to $\langle v \rangle$.

Incidentally, we have also proved at the same time that the average kinetic energy of the separate motions of the atoms is exactly the same as regarding the healthy motion of the CM is $\langle v \rangle^2$. For, the total kinetic energy of the two molecules is $2m_1 \langle v \rangle^2 + 2m_2 \langle v \rangle^2$, where average is $2\langle v \rangle = \langle v \rangle T$, to be precise because of the center of mass motion is $\langle v \rangle$, so the average kinetic energy of the remaining and everything else from the two atoms inside molecule is $2\langle v \rangle^2 = \langle v \rangle^2 T$.

In classical mechanics, the average energy of the CM motion is *constant*, for any initial condition as a whole, with forces present or not. For every independent direction of motion that there is, the average kinetic energy in that motion is $\frac{1}{2}kT$. These "independent directions of motion" are sometimes called the degrees of freedom of the system. The number of degrees of freedom of a molecule composed of n atoms is $3n$, since each atom needs three coordinates to define its position. The entire kinetic energy of the molecule can be expressed either as the sum of the kinetic energies of the separate atoms, or as the sum of the kinetic energy of the CM motion plus the kinetic energy of the internal motions. The latter case can also be expressed as a sum of two terms, kinetic and $\frac{1}{2}kT$, of the molecular and vibrational energy. In this is an approximation. Our theory, applied to the n -atom molecule, says that the molecule will have, on the average $3nT/2$ units of kinetic energy, of which $3kT/2$ is kinetic energy of the center-of-mass motion of the entire molecule, and the rest, $(3n - 3)kT/2$, is *vibrational and rotational kinetic energy*.

The Principles of Statistical Mechanics

40-1 The exponential atmosphere

We have discussed some of the properties of large numbers of interacting atoms. The subject is now kinetic theory, a description of matter from the point of view of collisions between the atoms. It will turn out, we see, that the macroscopic properties of matter should be explainable in terms of the motion of its parts.

We limit ourselves for the present to conditions of thermal equilibrium, that is, to a subsection of all the phenomena of nature. The laws of mechanics, which apply just in the real equilibrium, called statistical mechanics, and in this section we shall be because acquainted with some of the central features of the subject.

We already have one of the theorems of statistical mechanics, namely, the mean value of the kinetic energy for any system at the absolute temperature T is $\frac{1}{2}kT$ for each independent particle, i.e., for each degree of freedom. That tells us something about the mean square velocities of the atoms. Our objective now is to learn more about the positions of the atoms, to discover how many of them are going to be in different places at thermal equilibrium, and also to go into a little more detail on the distribution of the velocities. Although we know the mean square velocity, we do not know how to answer a question such as how many of them are going three times faster than the mean square, or how many of them are going one-quarter of the mean mean-square speed. Or how they all the same speed exactly!

So, these are the two questions that we shall try to answer; how are the molecules distributed in space when the atoms begin hitting at them, and how are they distributed in velocity?

If there is no force, our questions are completely independent, and that the distribution of velocities is always the same. We already carried a hint of the latter fact when we found that the average kinetic energy is the same $\frac{1}{2}kT$ per degree of freedom, no matter what form it takes, in the molecules. The distribution of the velocities of the molecules is independent of the forces, because the collision rate does not depend upon the forces.

Let's begin with an example—the distribution of the molecules in an atmosphere like our own, but without air winds and other kinds of disturbance. Suppose that we have a column of gas extending to a great height, and at the top is equilibrium, unlike our atmosphere, which as we know gets colder as we go up. We could expect that if the temperature varied at different heights, we could demonstrate lack of equilibrium by connecting a red to some blue at the bottom (Fig. 40-1) where they would pick up ΔT from the molecules there and would shake, via the red, the blue at the top and those would shake the molecules in the way. But, obviously, of course, the temperature becomes the same at all heights in a gravitational field.

If the temperature is the same at all heights, the problem is to determine by what law the atmosphere becomes thinner as we go up. If N is the total number of molecules in a volume V of gas at pressure P , then we know $PV = NkT$, or $P = nk$, where $n = NV$ is the number of molecules per unit volume. In other words, if we ignore the number of molecules per unit volume, we know the pressure, and vice versa, they are proportional to each other, since the temperature is constant in this problem. But the pressure is not constant, it must increase as the altitude is reduced, because it has to hold up its weight, the weight of all the gas above it. That is the clue by which we may determine how the pressure changes with height. If we take a small area dA at height h , then the vertical force from below

40-1 The exponential atmosphere

40-2 The Boltzmann law

40-3 Preparation of a liquid

40-4 The distribution of molecular speeds

40-5 The specific heats of gases

40-6 The failure of classical physics

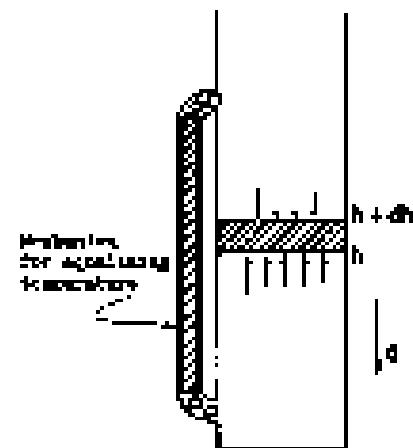


Fig. 40-1. The greater the height h the more chance that $m/h = dh$ by the weight of the intervening gas.

so the unit area, is the pressure P . The vertical component of force pointing down at a height $+ \Delta h$ would be less than in the absence of gravity, but here it is not because the force there is less than zero; the force from above has the weight of gas in proportion between Δh and $\Delta h + \Delta h$. Now $\rho g = m/n$ is of gravity on each molecule, where g is the acceleration due to gravity, and m/n is the total mass of molecules in the unit section. So this gives us the differential equation $P_{\Delta h} - P_{\Delta h + \Delta h} = \rho g \Delta h$ — maybe! Since $P = nkT$, and T is constant, we can eliminate either P or n , say P , and get

$$\frac{dP}{dh} = -\frac{\rho g}{n}$$

or the differential equation, which tells us how the density goes. Given this we go up to energy.

We thus have an equation for the partial density n , which varies with height, h , which has a derivative which is proportional to itself. Now a function which has a derivative proportional to itself is an exponential, and the solution of the differential equation is

$$n = n_0 e^{-\rho g h / n} \quad (40.1)$$

Here the constant of integration, n_0 , is the density at $h = 0$, which can be chosen anywhere, and the density goes down exponentially with height.

Note that if we have different kinds of molecules with different masses, they go down with different exponents. The ones which we breathe would decrease with z faster than the light ones. Therefore we would expect that because oxygen is heavier than nitrogen, it would be higher and higher to an atmosphere with nitrogen, and when the proportion of nitrogen would increase. This does not really happen in our own atmosphere, at least at reasonable heights, because there is no such separator which mixes the gases back together again. It is not an insurmountable barrier, however; there is a limit to the height of air currents like cyclones. A derivative is very great heights in the atmosphere is because the index masses continue to mix, while the other exponentials have all died out (Fig. 40.2).

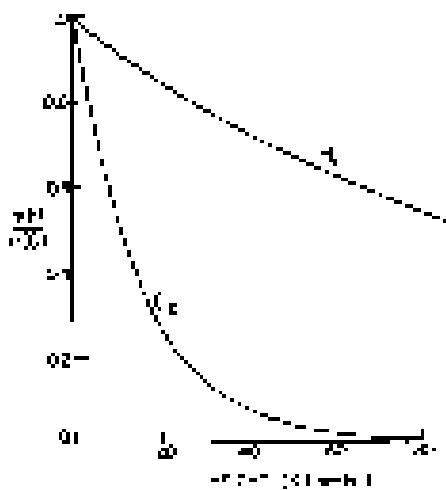


Fig. 40.2. The normalized density n/n_0 versus height z in the earth's gravitational field for oxygen and for nitrogen at constant temperature.

40-2 The Boltzmann law

Here we note the interesting fact that the connection is the exponent of P_0 , exactly in the previous example of expansion. Therefore we can deduce this particular answer: the density n any point is proportional to

$$e^{-\text{constant} \cdot \text{energy of molecule}}$$

This may be an accident, i.e., may be true only for this particular case of a uniform gravitational field. However, we can show that it is a more general proportion. Suppose that there were some kind of force which the molecule feels on the molecules in a gas. For example, the electrons may be charged electrons only, and may be caused to be an electric field or another charge that attracts them. Or, because of the mutual interaction of the atoms for some reason, or for the wall, or for a solid, or something, there is some force of attraction which varies with position x , which acts on all the molecules. Now suppose, for simplicity, that the molecules are all the same, and that the force acts on each molecule i in such a way that the total force on a piece of gas would be simply the difference between the charges on those molecules. To avoid unnecessary complication, let us choose a coordinate system with the x -axis in the direction of the force. In

In the same manner as before, if we take a small Δx distance the i -th molecule moves by a distance dx , then the force on i due from j times the volume per unit Δx is a generalization of the previous single times dx , must be denoted by the previous change, $F_{ij} dx = \partial P = \partial P/dx$. Or, to put the box in a form which will be useful to us later

$$F_i = -\partial \frac{\partial P}{\partial x} \text{ (inert)} \quad (40.2)$$

For the present, however, let us take the work we would do in skin your work from $x = 0$ to $x = \infty$, and if it comes from a potential, i.e., if the work done can be represented by a potential energy at x_0 , then this would be the difference of the potential energy (P.E.). The negative differential of potential energy is the work done, δA_x , and we find that $A_x = -\delta P.E./\delta x$. We also recognize,

$$A_x = \frac{1}{kT} e^{-E_x/kT} \exp(-E_x/kT). \quad (40.2)$$

Therefore what we called to a specific case may well be true in general. When this does not come from a potential, then (40.2) has no solution at $x = \infty$. Energy can be generated, as well as consumed, in small intervals of position for which the work done is not zero, and the equilibrium can be maintained at x^* . Thermal equilibrium cannot exist if the external forces on the atoms are non-conservative! Equation (40.2), known as Boltzmann's law, is another of the principles of statistical mechanics. Let the probability of finding molecules in a given "specific" arrangement varies exponentially with the negative of the potential energy of that arrangement divided by kT .

Now, how much will the distribution of molecules change? Suppose that we had a probability p in a given, interacting system where each atom, how many of them would have off-set distances? If the potential energy is known, we can calculate, then the proportion of them at different distances is given by this law, and so on through many applications.

40-3 Evaporation of a Liquid

In more developed statistical models we are likely to solve the following important problem. Consider an assembly of molecules which scatter each other, and suppose that the force between any two, say i and j , depends only on their separation r_{ij} , and can be represented as the derivative of a potential function $P(r_{ij})$. Figure 40-3 shows a form such a function might have. For $r > r_0$, the energy decreases as the molecules come together, because they attract, and then the energy increases very sharply as they come still closer together, because they repel at short range, which is characteristic of the very molecules behave, roughly speaking.

Now suppose we have a whole collection of such molecules and we would like to know how they arrange themselves on the average. The answer is $e^{-E_x/kT}$. The total potential energy E_x is obviously the sum over all pairs of atoms, plus, that the forces are all in pairs (there may be three-body forces in more complicated things like electricity, for example, the potential energy is \propto $1/r^2$). Then the probability for having molecule in any particular combination of r_{ij} 's will be proportional to:

$$e^{-\frac{1}{kT} \sum_i P(r_{ij})/kT_j}$$

Now, if the temperature is very high, so that $kT \gg E_x$ [Fig. 40-3], the exponent is relatively small, being everywhere and the probability of having a molecule at some fixed position. Let us now, however, take two molecules far apart, $r_{ij} \gg r_0$; then $P(r_{ij})$ would be the probability of finding them at extreme mutual distances r . Clearly, when the potential goes most toward zero, the probability is largest, and when the potential goes toward infinity, the probability is almost zero, which occurs for very small distances. This means that for molecules in a gas, there is no chance that they are on top of each other, since they repel so strongly. And here it goes without saying that, along them far apart, there are no points, the r , at any other point. Now, more pointed, again, on the temperature. If the temperature is not large compared with the difference in energy between $r = r_0$ and $r = \infty$, the exponent is always one by unity. In the case, when the mean kinetic energy is about kT , greater than the potential energy, the boxes do not make much difference. But as the temperature falls, the probability of finding the atoms too far apart (at distance r) gradually increases relative to the probability of finding them close together and, in fact, if kT is much less than $-P(r_0)$, we have a relatively

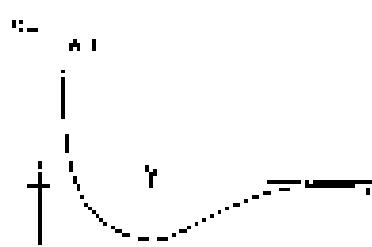


Fig. 40-3. A potential-energy function for two molecules, which depends only on their separation.

large positive exponent at the neighbor level. Furthermore, in a given volume they are much more likely to be at the distance of minimum energy than far apart. As the temperature falls, the atoms fall together, closer in liquids, and nearer to liquids, and solids, and molecules until you heat them up they evaporate.

The experiments for the determination of exactly how things evaporate, exactly how things should happen in a given circumstance involving the following law, to describe the correct nucleation-form: law $\propto e^{-\frac{E}{kT}}$ which does come from something else quantum mechanics, say, or experiment. But given the law of force by most the molecules, to discover what a billion molecules are going to do necessarily consists of studying the function $e^{-\frac{E}{kT}}$ surprisingly enough since it is such a simple function and with an easy idea, given the potential, to have it numerically computed; the difficulty is the huge number of variables.

In short, it's difficult to be subject to quite exciting and interesting. It is often called an example of a "many-body problem," and it really is an extremely interesting thing. In that single formula one has contained all the details, the example, about the solidification of gas, or the form of the crystals that the solid can take, and people have been trying to squeeze it out, but the mathematical difficulties are very great, not in writing the law, but in dealing with so enormous a number of variables.

Then there is the distribution of particles in space. That is the end of classical statistical mechanics, practically speaking because if we know the forces, we can, in principle, find the distribution in space, and the distribution of velocities is something that we can work out, more or less, and is not something that is different for the different cases. The given problems are to get the particular information out of our formal solution, and that is the main subject of classical statistical mechanics.

40-4. The distribution of molecular speeds

Now we go on to discuss the distribution of velocities, because sometimes it is interesting or useful to know how many of them are moving at different speeds in order to do this, we may make use of the fact which we discussed with respect to the gas in the triangle. We took it to be a perfect gas, as we have already assumed it, writing the potential, energy, dissolving the energy, "inside" distribution of the atoms. The only potential energy that we included in our first example was gravity. We would, of course, have something more complicated if there were forces between the atoms. Then we assume that there are no forces between the atoms and, for a moment, disregard collisions also, returning later to the justification of this. Now we saw that there are fewer molecules at the height than there are at the height 0 , according to formula (40.1), they decrease exponentially with height. How can there be fewer at greater heights? After all, consider the molecules which are moving up at height 0 with a v of 0 ft./sec. Because some of these which are moving up at 0 are going too slowly, and cannot climb the potential hill up. Well, therefore, we can calculate how many must be moving at various speeds, because from (40.1) we know how many are moving with low enough speed to climb given distances. These are just the ones that occur for the fact that the height at h is lower than at 0 .

Now let us put that idea a little more precisely. In a room, how many molecules are passing down below to cross the plane $v = 0$ (by calling it height $= 0$, we do not mean that there is a floor there; it is just a convenience), and there is gravitational field. These are molecules not moving around in every direction, but some of them are moving through the room, and at any instant a certain number per second of them are passing through the plane "from below" above with different velocities. Now we note the following: if we call v the velocity which is just needed to get up to the height h (kinetic energy $m v^2/2 = mgh$), then the number of molecules per second which are passing upward through the lower plane in a vertical direction with velocity component greater than v is exactly the same as the number which pass through the upper plane with low v of velocity. These molecules whose vertical velocity does not exceed v cannot get through the upper plane.

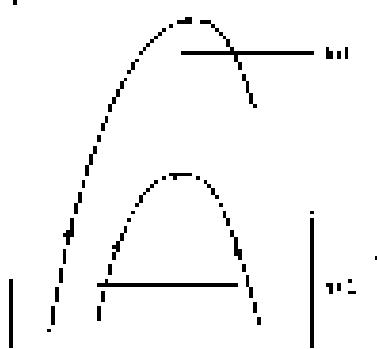


Fig. 40-4. Only those molecules moving up at h with sufficient velocity can cross at height 0 .

where we find that

$$\text{Number passing } h = \text{constant} \cdot v > v - \text{number passing } h = \text{constant} \cdot v > h.$$

But the number which pass through h with any velocity greater than h is less than the number which pass through any lower height with any velocity greater than h , because the number of atoms is greater, *except at the top*. We know already that the distribution of velocities is the same after the argument we made earlier, the temperature being constant all the way through the gas molecule. So since the velocity distributions are the same, *i.e.* v is just that there are more atoms having, clearly the number of atoms passing with positive velocity v height h over the number passing passing with positive velocity v height 0 are in the same ratio as the densities of the two heights which is constant. If $\rho_{\text{gas}}(h) = \rho_{\text{gas}}(0)$ and therefore we find that

$$\frac{\rho_{\text{gas}}(h)}{\rho_{\text{gas}}(0)} = e^{-\frac{mv^2}{2kT}} = e^{-\frac{m(v-h)^2}{2kT}},$$

since $h < v$ right. Thus in words, the number of molecules per unit area per second passing the height h with a component of velocity greater than h is *exactly* $e^{-\frac{m(v-h)^2}{2kT}}$ times the total number passing through the plane with velocities greater than zero.

Now this is not only true at the initial $h=0$ chosen height h , but of course it is true at any other height h , and thus the distributions of velocities are *all* the same. The final statement does not involve the height h , which appeared only in the intermediate argument. The total *one-particle* proportion that goes over the distribution of velocities. It tells us that if we drill a little hole at the side of a gas pipe, a very tiny hole, so that the exit ports are permanent for bacteria, *i.e.* are *farther* apart than the diameter of the tube, then the particles which are coming out will have *different* velocities, but the fraction of $v>h$ molecules which come out is *exactly* greater than $e^{-\frac{m(v-h)^2}{2kT}}$.

Now we return to the question about the reflection coefficient. Why does it not make any difference? We could have pursued the same argument, now with a finite height h , but with an infinite initial height h , where is so small that there would no longer be v values between them? h ? But that was not necessary, the $v>h$ term is evidently *exact* in an analysis of the energies involved, the conservation of energy, and at the collisions that occur there is an exchange of energies between the molecules. However, we do *not* really *know* whether the following statement makes sense: energy is entirely exchanged with another molecule. But it is *extremely* hard if the collision is *inelastic* irreversibly, and it is more difficult than ρ_{gas} , to do a rigorous job, and it makes no difference in the result.

It is interesting that the velocity distribution we have found is per

$$n_{\text{gas}} = \rho_{\text{gas}}^{1/2} m^{1/2} e^{-mv^2/2kT}. \quad (40.4)$$

For very often they the distribution of velocities, by giving the number of molecules that pass a given area with a certain maximum component, is not the most convenient way of giving the velocity distribution. For instance, inside the gas, one often wants to know how many molecules are moving with a component of velocity between two given values, and that of course, is not directly given by Eq. (40.4). We would like to state our result in the more convenient form, even though what we already have written is quite general. State that it is not possible to say with any moderate law exactly how many velocity components of v has a velocity exactly equal to v , 795289/73 meters per second. But if one wants to be a meaningful statement, we have to ask how many are to be found in some range of velocities. We have to say how many v 's exist between v , 795289/73, and $v+dv$. Or, *another* way to ask it is to be the fraction of all the molecules which have velocities between v and $v+dv$, *i.e.* what is the corresponding if this is calculated, all that have a velocity in that range dv . Figure 40-3 shows a graph for the function ρ_{gas} , and the shaded part, of width dv of v , which height ρ_{gas} , represents this fraction $\rho_{\text{gas}}(v) dv$. That is, the ratio of the central

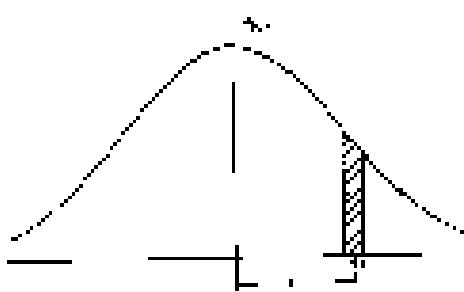


Fig. 40-3. A velocity distribution function. The shaded area is $\rho_{\text{gas}}(v) dv$, the fraction of particles having velocities within a range dv .

area to the total area of the curve is the relative proportion of molecules with velocity v within dv . If we define $f(v)$ as the fraction having a velocity in this range, it is given directly by the shaded area, that is, the total area must be 100 percent of room, that is,

$$\int_{-\infty}^{\infty} f(v) dv = 1. \quad (40.4)$$

Now we have only to get this distribution, by comparing it with the fluxes we derived before. First we set what is not the flux of molecules passing through an area per second with a velocity greater than v , expressed in terms of flux. At first we might think it is merely the integral of $\int_v^{\infty} f(v) dv$, but it is not because we might not have lost the passing flux in just v . The latter case may occur often, we know, than the above case, and in order to compare both fluxes, you have to multiply by the velocity. (We discussed this in the previous chapter when we talked about the number of collisions.) In a given time the total number which pass through the v flux is $\frac{1}{2} v$ of those which have been able to rise in the stream, and this number will then be given from a calculation of the number of molecules which arrive at exactly the number which pass here. But the numbers are there per unit volume, multiplied by the distance that they sweep through in rising for the time through which they are supposed to go, and that distance is proportional to v . It is evident the integral of a curve $f(v) dv$, an infinite integral with a lower limit v , and this must be the same as we found before, namely $e^{-\frac{mv^2}{2kT}}$, with a proportionality constant which we "fix" as follows:

$$\int_v^{\infty} v f(v) dv = \text{constant} e^{-\frac{mv^2}{2kT}}. \quad (40.5)$$

Now if we differentiate the integral with respect to v , we see the thing just is inside the integral, i.e., the integral (with a definite upper limit) is zero, and if we differentiate the other side, we get v times the same exponential (and some constants). This is canceled and we find

$$f(v) dv = C e^{-\frac{mv^2}{2kT}} dv. \quad (40.6)$$

We make the air or heat value as a standard that it is a probability, and this is what the proportion is that velocity between v and $v + dv$.

The constant C must be so determined that the integral is unity, according to Eq. (40.4). Now we can prove that

$$\int_0^{\infty} e^{-\frac{mv^2}{2kT}} dv = \sqrt{\pi}.$$

Using this fact, it is easy to find that $C = \sqrt{\pi/(2m k T)}$.

Since velocity and momentum are proportional, we may say that the distribution of molecules is also proportional to $e^{-\frac{p^2}{2mkT}}$ per unit momentum range p . It is noted that this relation is more or less relative now, if it is in terms of momentum, while it is in velocity it is not, so it is best to keep momentum instead of velocity:

$$dP/dp = C e^{-\frac{p^2}{2mkT}} dp. \quad (40.7)$$

So we find that the probability is of different conditions of energy, kinetic and potential, are both given by $e^{-\frac{E}{2mkT}}$, a very easy thing to remember and a rather beautiful proportion.

* The general value of the integral is:

$$\int_{-\infty}^{\infty} e^{-\frac{p^2}{2mkT}} dp = \text{constant}$$

$$= \sqrt{\pi} \cdot \int_{-\infty}^{\infty} e^{-x^2} dx = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi},$$

which is a double integral over the whole xy plane. But the condition of solution is polar coordinates $r\theta$:

$$P = \int_0^{\infty} r e^{-\frac{r^2}{2mkT}} dr = \pi \int_0^{\infty} r e^{-r^2/2} dr = \pi$$

So far we have, of course, only the distribution of the velocities "vertically." We might want to ask, what is the probability that a molecule is moving in a given direction? Of course if we do this for all directions, and one can obtain the complete distribution from the one we have, because the complete distribution depends only on the square of the magnitude of the velocity, not upon the component. It must be something that is independent of direction, and there is only one function involved, the probability of different orientations. We have the distribution of the components, and therefore we can get the distribution of the other components from it. The result is then the probability is still proportional to $e^{-\frac{1}{2}kT\cos^2\theta}$, but now the kinetic energy involved there is $\frac{1}{2}m(\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2)$, summed in the expression. So we can write it as a product:

$$f(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3)$$

$$= e^{-\frac{1}{2}kT} \cdot e^{-\frac{1}{2}\dot{x}_1^2/kT} \cdot e^{-\frac{1}{2}\dot{x}_2^2/kT} \cdot e^{-\frac{1}{2}\dot{x}_3^2/kT} \quad (40.3)$$

You can see that the formula must be right because, first it is a function only of T , as required, and secondly the probabilities of various values of x_i are given by integrating over all \dot{x}_j and x_j is just (10^{-3}) . But the x_i function (10.9) can do both these things.

40-5 The specific heats of gases

Now we shall look at some ways to test the theory, and to see how useful is the classical theory of gases. We expect that, if E is the internal energy of N molecules, then $PK = NkT = E/N = \frac{1}{2}N$ holds, since as we saw before, maybe, if it is a monoatomic gas, we know this is also true. In a diatomic molecule, if the vector of mean motion of the atoms T is independent of mass, then the kinetic energy is equal to the potential energy, and therefore $T = \frac{1}{2}E$. But suppose it is, say, a monoatomic molecule made of one spin zero atom, and it's a simple system (turns out to be the according to classical mechanics) that the energies of the internal vibrations are also proportional to kT . Then at a given temperature, in addition to kinetic energy ET , it has thermal vibrational or rotational energy. So the total E includes not just the internal kinetic energy, but also the rotational energy, and we get a different value of T . Let's do γ , the best way to measure γ is by measuring the specific heat, which is the change in energy with temperature. We'll return to γ later. For our present purposes, we may suppose γ is found experimentally from the E/T curve for different temperatures.

Let us make a table of γ for some cases. First, for a monoatomic gas N is the total energy, the sum of the kinetic energy, and we know already that right. If it's a diatomic gas, we may take, as an example, oxygen, by adding double, hydrogen, etc., and suppose that the diatomic gas can be approximated as two atoms held together by some kind of force, as in one of Fig. 40-1. We may also suppose, and it seems to be quite true, that all the energies of γ that are of interest for the diatomic gas, are pairs of terms that add linearly to the total energy γ_1 , the distance of potential minimum. If this were not true, if the probability were not strongly varying enough to make the potential energy tilt from the bottom, we would have to remember that oxygen, for instance, is a pure single oxygen along in a 2:1 ratio. We know that these are, in fact, very few single oxygen atoms. This means that the potential energy minimum is very much greater in magnitude than kT , as we have seen. Since they are all held so strongly around r_0 , the only form of vibration that is excited is the one near the minimum, which must be approximated by a parabola. A parabolic potential implies a γ that is usually positive. In fact, to an excellent approximation, the oxygen molecule can be represented as two atoms connected by a spring.

Now γ is the total energy per mole divided at temperature T . We know that for each of the two atoms, each of the kinetic energies should be $\frac{3}{2}kT$, so the kinetic energy of both of them is $\frac{3}{2}NkT = \frac{3}{2}kT$. We can also put γ in a different way; the so-called γ' can also be looked at as $\text{internal energy}/\text{decrease in temp}$

Table 40-1

Values of the quantum level ratio, γ ,
for various gases

Gas	T (K)	γ
He	-140	1.661
Si	19	1.68
Ar	15	1.688
H ₂	100	1.706
O ₂	100	1.709
III	100	1.70
Br ₂	300	1.72
I ₂	385	1.73
NH ₃	71	1.77
C ₂ H ₆	15	1.77

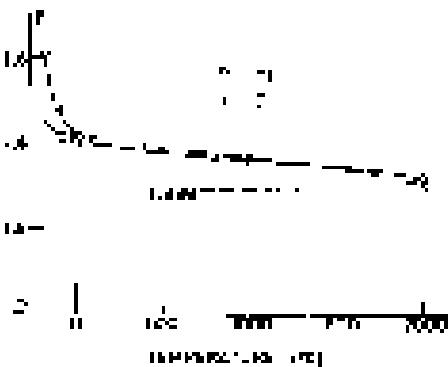


Fig. 40-6 Experimental values of $1/\gamma$ as a function of temperature for hydrogen and oxygen. Classical theory predicts $1/\gamma = 1.726$, independent of temperature.

(6), kinetic energy of rotation (5), and zeroth-order of vibration (3). We know that the kinetic energy of vibration is 0, since there is no displacement involved other than degrees of freedom for θ . Regarding the rotation, it can be a silent player or two seats, so there are two independent choices. We assume that the silent ones are some kind of pairs, and cannot spin around the joining atom, this is something we have in mind, however if we get a disagreement, maybe that is where the trouble is. But we have one more thing, which is the quantum energy of vibration; how much is that? In a diatomic oscillator the average kinetic energy and the zero-point energy are equal, and therefore the total total energy of vibration is $\hbar\omega_0$, also. The ground total of energy is $\hbar\omega_0 + \frac{1}{2}\hbar\omega_0$, or $\hbar\omega_0$ is 20% per atom. That means that $\hbar\omega_0/2$ is below $\hbar\omega_0/6$, i.e., 1.226.

We may compare these numbers with the relevant mean of values shown in Table 40-1. Looking first at helium, which is a monatomic gas, we find very nicely $\hat{\gamma} = 1.661$, and the value reported by experimentalists at such low temperatures may be some forces between the atoms. Hydrogen and neon, both monatomic, agree also within the accuracy of the experiment.

We turn to the diatomic gases and find hydrogen with 1.706, which does not agree with the theory, 1.726. Oxygen, 1.709, is very similar, but again not in agreement. Hydrogen sulfide again is similar at 1.70. Hydrogen to look at enough the right answer is 1.70, but it is not because they took hundred of measurements, we see 1.70, since 2.30 is reasonably close to 1.70, nothing may be said to agree with us, but oxygen is off. So here we have a dilemma. We know it right for one molecule, we do not have it right for another molecule, and we may need to be really ingenious in order to explain how.

Let us look further at a more complex diatomic molecule with three numbers of particles, for example, C₂H₆, which is ethane. It has eight atoms, and they are all rotating and translating in various combinations. The total amount of internal energy must be an even-integer number of \hbar^2 's, at least 12.5% the kinetic energy alone, and $\gamma = 1$ must be very close to zero, or 2 almost every time. If it is lower, but 1.22 is much lower, and is higher than the 1.70 calculated from the kinetic energy alone, and it is just understandable.

Furthermore, this whole system is deep, because the distance and orientation between rigidly coupled atoms make the coupling weaker increasingly, although it might not vibrate much, or would nevertheless keep vibrating. The vibrational energy inside is still $\hbar\omega$, since it does not depend on the strength k of the coupling. But, if we want to give absolute rigidity, erasing all vibrational $\hbar\omega$'s, eliminate a variable, then we would get $\gamma = 2.37$ and $\gamma = 1.66$ for the diatomic case. This looks good for H₂ & O₂. On the other hand, we would still have problems because γ is off the hydrogen or oxygen series with temperature. From the measured values shown in Fig. 40-6, we see that for H₂, γ varies from about 1.8 at -165°C to 1.3 at 200°C. The variation is more substantial in the case of hydrogen than for oxygen, but nevertheless, even in oxygen, γ tends definitely to drop as we go down in temperature.

40-4 The failure of classical physics

So, in all, we realize now that we have some difficulty. We might try some force law other than a spring, but I think that anything else will only make γ higher. If we include more forms of energy, i.e. appreciates only more closely, surrounding the atom. All the classical theoretical things that we can think of will only increase γ . That is to say, there are obstacles in each atom, and we know from their theory that there are internal vibrations; each of the electrons should leave at least 40% of kinetic energy, e.g. in falling for the potential energy, so when these are added in, γ gets off further. It is probably, it is probable.

The big open paper on the dynamics theory of gases was by Maxwell in 1861. One the basis of ideas we have been discussing, he was able accurately to explain a great many known relations, such as Boyle's law, the diffusion theory, the viscosity of gases, and things we shall talk about in the next chapter. He listed all these great successes in a final summary, and at the end he said, "Finally, in 45-1

establishing a necessary relation between the motions of molecules and rotation (he's talking about the $\frac{1}{2}kT$ theorem): if all particles did spherical, we would then have a system of such systems would not precisely satisfy the known relation between the two sp. like heats." He is referring to r (which we've discussed) is due to two ways of measuring specific heat, and he says we know we cannot get the right answer.

Two years later, in a lecture, he said, "I now know well what you asked. I consider it to be the greatest difficulty not surmounted by the molecular theory." These words represent the first discovery that the laws of classical physics were wrong. This was the first indication that there was something fundamentally impossible, because a rigorously general theorem did not agree with experiment. After 1900, Debye was to talk about the previous year. On another level, he said that physicists at the latter part of the nineteenth century thought they knew all the significant physical laws, and that all they had to do was to calculate more decimal places. Some time must have told them that was not so. But a thorough reading of the literature of the time shows they were ... worrying about something that's said above this paper ... but it is a very mysterious concern, and it seems to bring us to the conclusion that certain kinds of motions "frozen out."

If we could assume that the vibrational velocity, did not exist at low temperature and did exist at high temperature, then we might imagine that a gas might exist at a temperature sufficiently low that vibrations continue to be important, say $T = 1.40$, or a higher temperature, at which they begin to cease. In so far, this same might be argued for the rotation. If we can eliminate the rotation, say it "freezes out" at sufficiently low temperature, then we can understand the fact that there is hydrogen appearance. It goes on going down in temperature. Here one sees problems of such a phenomenon. Of course that mass motion "frozen out" cannot be understood by classical mechanics. It was only understood when quantum mechanics was discovered.

Without proof, we may state the results of quantum mechanics of the quantum mechanical theory. We recall that according to quantum mechanics, a system which is bound by a potential, for vibrations, for one atom, will have a discrete set of energy levels, i.e. states of definite energy. Now in quantum mechanics, statistical mechanics is no modified according to quantum-mechanical theory. But it is not, it is amazingly strong, nor although most problems are indeed difficult in quantum mechanics, less in classical mechanics, problems in quantum mechanics are much easier in quantum theory. The simple case we have in classical mechanics, that a $\text{e}^{-E/kT}$, becomes the following very important theorem. If the energies of interest of individual states are listed, say, E_1, E_2, E_3, \dots , the probability in thermal equilibrium, the probability of finding a molecule in the particular state of having energy E_i , is proportional to $e^{-E_i/kT}$. That gives the probability of being in various states. In other words, the relative chance, the probability, of being in state E_i , relative to the chance of being in state E_j , is

$$\frac{P_i}{P_j} = e^{\frac{-E_i/kT}{-E_j/kT}} \quad (40.10)$$

at kT , of course, is the same as

$$P_i = n_p e^{(E_i - E_p)/kT}, \quad (40.11)$$

where $P_i = n_i/n$ and $P_p = n_p/N$. So it is likely to be in a higher energy state than in a lower one. The ratio of the number of atoms in the upper state to the number in the lower state is n times to the power $(E_p - E_i)/kT$, whence, over N , it is very close to preservation.

Now, it turns out that for a harmonic oscillator the energy levels are evenly spaced, including the lowest energy, $E_1 = 0$ (it actually is not zero, but it's different, but it does not matter if we shift the energies by a constant, just that one is then 0). And the second one is $2\omega_0$ and the third one is $4\omega_0$, and so on.

Now let us see why happens. We suppose we are studying the vibrations of a diatomic molecule, which we approximate as a harmonic oscillator. Let us see

that is the relative chance of finding a molecule in state E_1 , instead of in state E_2 . This is given by the chance of finding a molecule E_1 , relative to the probability of finding it in state E_2 , goes down as $e^{-E_1/kT}$. Now suppose that kT is much less than E_2 , and we have a λ to "quantum" corresponds. Then the probability of finding it in state E_1 is extremely small. Obviously all the atoms are in state E_2 . If we change the temperature but still keep kT very small, then the chance of its being in state E_1 is the natural limit, and the energy of the oscillator remains nearly zero in excess of change with temperature so long as the temperature is much less than E_2 . All oscillators are in the fundamental state, and hence motion is effectively "frozen"; there is no contribution of kT to the specific heat. We can judge, then, from Table 40-1 that at 100°C, which is 373 deg. absolute, kT is much less than the vibrational energy in the oxygen or hydrogen molecules, but not so in the iodine molecule. The reason for this difference is that an iodine atom is very heavy, composed with nitrogen, and although the mass may be comparable of iodine and by 2 nitrogen, the iodine molecule is so heavy that the natural frequency of vibration is very low compared with the natural frequency of hydrogen. At the upper limit $kT > E_2$, room temperature the hydrogen has lower frequency, only the latter, incident, exhibits the classical vibrational energy. As we increase the temperature kT , first, starting from a very low value of kT , with the molecule almost all in their lowest state, they gradually begin to have an appreciable probability to be in the next state, and then in the next state, and so on. When the probability is appreciable for many states, the breakdown of the probability method that given by classical physics, because the quantized states become really indistinguishable from one in our mind, and the system can have classical energy. Thus in the temperature range, we should again get the results of classical mechanics, as indeed seems to be the case in Fig. 40-6. It is possible to show in the same way that the molecular vibrations $> kT$ to contribute, but the states are so much closer together that the ordinary correspondence kT is broken: hence the energy "lawn" energy levels are excited, and the rotational thermal energy in the system contributes in the classical way. The same example where this is done up to a certain temperature is for hydrogen.

This is the time that we have already discussed, by comparison, with classical mechanics. I.e., one was comparing wrong with classical physics, and will be looked for a modification of the classical mechanics in mechanics in much the same way as it was done in optics. It was 10 or 20 years before the next difficulty was discovered, and that had to do again with quantum mechanics, but this time the mechanics of a phenomenon. The problem was solved by Planck in the early part of this century.

The Brownian Movement

41-1 Experimental Discovery

The Brownian movement was discovered in 1827 by Robert Brown, a naturalist. While he was studying microorganisms, he noticed little particles of plant pollen jiggling around in the liquid he was looking at in the microscope, and he was wise enough to realize that these were not living, but were just little pieces of dirt moving around in the water. In fact, he tried to determine that this had nothing to do with life by putting from the ground an oak piece of gorse in which there was some water trapped. It must have been trapped for millions and millions of years. But when he would see the same jiggling. When you look at the tiny tiny particles and jiggling particles,

This was later proven to be one of the effects of molecular motion, and we can understand it very easily by thinking of a good push to a jiggling billiard ball from a great distance, with a lot of people understand all pushing the ball in random directions. We can not see the people because we imagine that we are too far away, but we can see the ball, and we notice that it moves straight in a zigzag. We also know, from the treatment that we have developed in a previous chapter, that the mean kinetic energy of a small particle suspended in a liquid or a gas is $\frac{1}{2}mv^2$, even though it is very many compared with a molecule. If it is very heavy, then clearly the speed will relatively slow, but it will still, usually, be. This speed is not really so slow. In fact, we can not see the effect of such a particle very easily because although the mean kinetic energy is $\frac{1}{2}mv^2$, which represents a speed of a million miles per second, it is about a million times larger than this. This is very hard to measure in a microscope, and so the particle immediately comes to a collision and does not get anywhere. How far it does get we will discuss at the end of this present chapter. This problem was first solved by Einstein at the beginning of the present century.

Inside talk, when we say that the total kinetic energy of the particle is $\frac{1}{2}mv^2$, we claim to have derived this result from the kinetic theory, that is, from Newton's laws. We shall find that we can derive all kinds of things—magnetism, things from the kinetic theory, and this is not interesting that we can apparently get much from nothing. Of course we cannot measure the Brownian Brownian "little"

—they are thought to do it easily. What we need is that we did not do very much. How do we prove that? The way is that we have been continually making a certain assumption implicitly, which is this: if a given system is in thermal equilibrium at some temperature, it will also be at thermal equilibrium with anything else at the same temperature. For instance, if we were to let some little granules move in water, and we were counting with water, we could imagine that the water present, composed of another kind of particle, little fine pellets that you supposed do not interact with water, but only by being in contact with them. Suppose the water has a strong sticking surface, all our pellets have to stick to the water. We know that without this kind of a vapor of little sticks at the surface, that is an ideal gas. Water is complicated, it is an ideal gas simple. Now, our particle has to be in equilibrium with the gas of water. Therefore, the total motion of the particle must be zero. We get for average velocity a number if it were not moving at the right speed relative to the water but say, were moving faster, that would mean the pellets would pick up energy from it and get farther from the water. But we had started them at the same height, and if we assume that at time t there is once in equilibrium, it stays in equilibrium— $v = 0$ if it is not pushed and other gas is with, again exactly.

41-1equipartition of energy

41-2 Thermal equilibrium of radiation

41-3 Equipartition and the quantum oscillator

41-4 The random walk

This proposition is true and can be proved from the laws of mechanics, but the proof is very complicated and can be established only by using advanced mechanics. It is much easier to prove in quantum mechanics, but it is in classical mechanics. It was proved by by Heisenberg, but for now we may take it to be true, and then we can argue that one particle has no temperature if it is at rest with zero power, so it can move freely (in space) it is being hit with wave at the same temperature and we take among the particles, as it is good. It is a stronger line of argument, not perfectly valid.

In addition to the rotating cylindrical surfaces for which the temperature was first determined, there are a number of other phenomena, both in the laboratory and in other situations, where one can see Brownian movement. If we are trying to build the most delicate possible thermometer, say a very small mirror on a thin wire fiber for a very sensitive ballistic galvanometer (Fig. 41-1), the mirror does not stay still, but jiggles all the time. On the same wire between we have a spot on it, and when we look at the position of the spot, we do not have a perfect instrument because it is moving all the time. Why? Because we are not kinetic energy. Creation of the vibration has to be in the average. Not

What is the measurement angle over which the mirror will vibrate? Suppose we find the natural vibration period of the mirror by tapping on one side and seeing how long it takes to oscillate back and forth, and we also know the moment of inertia, I . Well now the formula for the kinetic energy of rotation, it is given in Eq. (10.5-3) — $\frac{1}{2}I\omega^2$. That is the kinetic energy and the potential energy that goes with it will be proportional to the square of the angle — it is $E = \frac{1}{2}I\omega^2$. Then if we know the period T , we calculate from this the natural frequency $\omega = 2\pi/T$, then the potential, E_p , is $E_p = \frac{1}{2}I\omega^2T^2$. Now we know that the average kinetic energy is kT , but since it is a cosine it is small, so the average potential energy is also kT . Thus

$$\begin{aligned} \langle E \rangle &= \langle E_p \rangle + \langle E_k \rangle \\ \langle E \rangle &= kT + kT \end{aligned} \quad (41.1)$$

In this way we can calculate the oscillations of a galvanometer mirror, and therefore find what the limitations of our instrument will be. If we want to have small oscillations, we have to cool the mirror. An interesting question is, where to cool it. This depends upon where it is getting its "ticks" from. If it is through the fiber, we cool it at the top. If the mirror is surrounded by a gas and is getting its energy by collisions in the gas, it is better to cool the gas. As a matter of fact, if we know where the damping comes the natural energy from, it turns out that there is always the source of the fluctuations after a point which we will come back to.

The same thing works, amazingly enough, in electric circuits. Suppose that we are building a very sensitive, accurate amplifier for a certain frequency and then a resonant circuit (Fig. 41-2) in having to go to twice it, or, whatever, to the natural frequency, like a radio receiver, but a really good one. Suppose we want to go down to the very lowest limit of things, so we take the voltage very off the inductor, and send it into the rest of the amplifier. Of course, in my circuit like this there is a certain amount of loss. It is not a perfect resonant circuit, but it is very good one (as there is in a resonance). Say (we will be interested in ω) we can see it, but it is supposed to be small. Now we want to find out: How much does the voltage across the inductor fluctuate? Answer: We know that $\langle E \rangle^2$ is the "kinetic energy" — the energy associated with a coil in a resonant circuit (Chapter 22). Therefore, the mean value of $\langle E \rangle^2$ is equal to kT — that means that the inductor is big we can just cut what the rms voltage below the maximum. For this voltage the voltage across the inductor the current is $I = \omega L/V$, and the mean, square voltage on the inductor is $\langle V^2 \rangle = I^2 L^2 / R^2$, and putting in $\langle E \rangle^2 = I^2 L^2 / R^2 T$, we obtain

$$\langle E \rangle^2 = I^2 L^2 / R^2 T \quad (41.2)$$

So now we can design circuits and all when we are going to get what is called Johnson noise, the noise associated with thermal fluctuations.

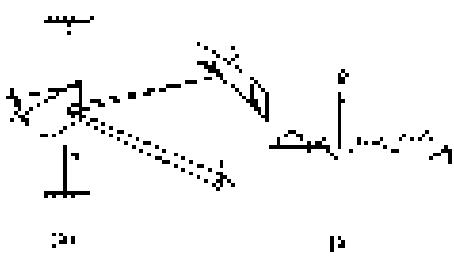


Fig. 41-1. (a) A rotating cylindrical surface. Light from a source S is reflected from a small mirror and is seen. (b) A sinusoidal record of the reading of the scale as a function of the time.

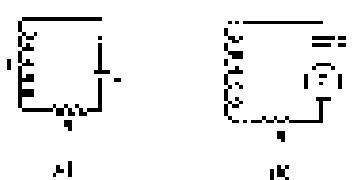


Fig. 41-2. A high-Q resonant circuit. (a) Actual circuit at resonance. (b) Artificial circuit, with or noise (indicated resistors and a "noise generator" G).

Where do the fluctuations come from? They come out from the resistor. They come from the heat that the electrons in the resistor are picking up because they are in thermal equilibrium with the resistor, and they make fluctuations in the density of electrons. They thus make tiny currents that will drive the current circuit.

Electrical engineers represent the current in another way. Physically, the noise is reflecting the source of noise. However, we may replace the real noise by having an instant, brief signal source which is making noise, by an artificial current which contains a little generator that is trying to reproduce the noise, and now the resistor is otherwise dead. So noise comes from it. All the noise is in the artificial generator. And so I think the characteristics of the noise produced by a resistor, if we had the formula for it, then we could calculate what the noise is going to be in response to that noise. So, we need a formula for the noise fluctuations. Now remember that is generated by the resistor is at all frequencies, since the resistor by itself is not resonant. Of course the resistor circuit only "hears" in the part that is near the right frequency, but the resistor has many different frequencies in it. We may describe how strong the generator is as follows. The mean power that the resistor would absorb if it was connected directly across the noise generator would be $\langle P \rangle$, P , $\langle \cdot \rangle$ means the voltage from the generator. But we would like to know τ more than just how much power there is at every frequency. There is very little power in any one frequency; it is a distribution. Let $P(s)$ be the power that the generator would deliver in the frequency range s into the rest of the resistor. Then we can prove (we shall prove it for another time), that the characteristic is exactly the same that the power remains out

$$P(s) ds = \langle P(s) \rangle \delta(s), \quad (41.3)$$

and is independent of the resistance when put this way.

41-1 Thermal equilibrium of radiation

Now we go on to consider a still more advanced and interesting problem in that is called "Suppose we've got a charged oscillator like those we were talking about when we were discussing light, let's say an electron oscillating up and down in an atom. If it oscillates up and down, it radiates light. Now suppose that this oscillator is in a very thin gas of oscillators, and let's say it's time to time the atoms collide with it. Then it's at equilibrium after a long time this oscillator will pick up energy until the total kinetic energy of the oscillator is kT , and since it is a harmonic oscillator, its total energy of motion will become kT . That is, $\langle E_{kinetic} \rangle$, a strong distribution of E , because the oscillator carries electric charge, and if it loses an energy kT it is shaking up and down and radiating light. Oscillating it is impossible to have equilibrium unless matter along with it is changing in it emitting light, and as light it is losing energy. Hence away, the oscillator loses its kT as time goes on, and then the whole gas which is oscillating with the oscillator gradually cools off. And that is, of course, the way a hot star cools off and might be radiating the light into the sky, because the atoms are losing their charge and they are continually recombining, and thereby breaking off from radiation. So it's continually cooling off.

On the other hand, if we enclose the whole thing in a box so that the photons are going out continually, then we've eventually get thermal equilibrium. We may either put the gas in a box where we can say that there are other oscillators in the box walls sending light back to us, or take a other example, we may say that the box has mirror walls. It is easier to think about that case. Then we assume that all the radiation that goes out from the oscillator stops running around in the box. Then, of course, it is true that the oscillator tends to radiate, but pretty soon it can maintain its kT of total energy in spite of the fact that it is radiating, because it is being compensated, we may say, by its own light reflected from the walls of the box. This is like a while there's a present of light running around in the box, and through the oscillator is radiating some, the light comes back and reflects off of the walls of the box and is radiated.

We shall now determine how much light there must be in such a box at frequency ν in order for the shifting of the light on the oscillators will generate just enough energy to account for the light it receives.

Let the two boxes be very few and far between, so that we have an idealized "intermittent" or "resonance" case of the lattice resistance. Then we consider that at thermal equilibrium the oscillator is doing two things at the same time. First, it has a mean energy $\mu_0 \omega T$, and second it is doing some radiation as well. Second, this radiation should be exactly the amount that would result because of the fact that the light shifting of the oscillator is not zero. Since there is nowhere else the energy can go, the effective radiation is really just scattered light from the light that is in there.

This we call spontaneous luminescence, but is generated by the oscillator at second, if the oscillator has a certain energy. (We learn more later, Chapter 12 on radiation resistance a number of equations without going into over their derivation.) The energy radiated per second divided by the energy of the oscillator is called β/α (Eq. 12.3): $\beta/\alpha = 2\pi^2 \mu_0 \omega_0 \sinh(\hbar \omega_0/kT)$. Unlike the quantity α , the damping constant, this can also be written as $\beta/\alpha = 2\pi \omega_0$, where ω_0 is the natural frequency of the oscillator. (Remember why ω_0 ? ω_0 is very large. The energy scattered per second is, then,

$$\frac{dE}{dt} = \frac{\omega_0 E^2}{2} = \frac{2\pi^2 \mu_0 \omega_0}{m} \cdot \omega_0 E. \quad (41.4)$$

The energy radiated per second is thus simply gamma times the energy of the oscillator. Now, the oscillator starts to have an average energy $\mu_0 T$, so we see that gamma β/α is the average amount of energy radiated per second:

$$\langle dE/dt \rangle = \beta/\alpha T. \quad (41.5)$$

Now we only have to know what β/α is. Gamma is easily found from Eq. (32.72), that is,

$$\gamma = \frac{\omega_0}{Q} = \frac{2\pi \omega_0^2}{\lambda \cdot e}, \quad (41.6)$$

where $r_s = \lambda^2/m^2$ is the classical electron radius, and we know $e = -e\hbar c/4\pi r_s$.

Our "int" result for the average rate of radiative of light and the frequency ω_0 is therefore

$$\frac{dE}{dt} = \frac{2\pi \omega_0^2 \mu_0 T}{\lambda^2 e}. \quad (41.7)$$

Next we ask how much light must be shifted on the oscillator. It must be enough that the energy received from the light (and the upshift scattered) is just exactly this much. In other words, the emitted light is accounted for as scattered light from the light that is shifting on the oscillator. In order to do this, we must now calculate how much light is scattered from the oscillator if there is a certain amount, unknown, of radiation incident on it. Let $I(\nu, d\nu)$ be the energy of light energy density $d\nu$ at the frequency ν within a certain range $d\nu$ (because there is no light at exactly one particular frequency, it is spread out over the spectrum). So $I(\nu, d\nu)$ is a certain spectral distribution which we are now going to "integrate" to get a value of a quantity at temperature T . Just as you do when you open the door and look at the hole. How more light is absorbed? We worked out the amount of radiation absorbed from a given incident light frequency, and we substitute it in terms of a cross section. It is just as though we did that since the light can thus on a certain cross section is absorbed. So the total amount that is radiated (additively) is the incident intensity $I(\nu, d\nu)$ multiplied by the same factor σ .

The form to for the cross section which we derived (Eq. 21.16) is $\sigma = \sigma_0$ and has the damping included. Let us look at it through the derivation again and put in the results less terms which we neglected. If we do that, and calculate the cross section the same way, we get

$$\sigma_0 = \frac{8\pi^2}{3} \left(\frac{\omega^2}{(\omega_0^2 + \omega^2)^2 + \gamma^2 \omega^2} \right). \quad (41.8)$$

Now, σ is a function of frequency, ω , so of significant size only for ω very near to the natural frequency ω_0 . (We realize that the σ for an oscillating oscillator is equal to σ_0^2). The oscillator scatters very strongly when ω is equal to ω_0 and very weakly for other values of ω . Therefore we can replace ω by ω_0 and $\omega^2 - \omega_0^2$ by $2\omega_0\omega - \omega_0^2$, and we get

$$\sigma_0 = \frac{2\pi^2 k T}{\pi(\omega_0^2 - \omega_0^2 + \gamma^2/4)} \quad (41.7)$$

Now the whole curve is broad and near $\omega = \omega_0$. (We do not really have to make any approximation, but it is easier to do the integral if we simplify the equation). Now we multiply this intensity in a given frequency range by the cross section σ_0^2 scattering power, the amount of energy scattered in the range $d\omega$. The total energy scattered is then the integral of this for all ω . That

$$\begin{aligned} \frac{dE}{dt} &= \int_{-\infty}^{\infty} \text{Intensity} d\omega \\ &\cdot \left(\frac{2\pi^2 k T \omega^2 d\omega}{\omega^2 - \omega_0^2 + \gamma^2/4} \right) \end{aligned} \quad (41.8)$$

Now we set $dW/dt = 3kT$. Why? Well: we made our analysis of the cross section in Chapter 17, we assumed the the particle was one such that the light scattered the oscillator. If we had used an oscillator which could move only in one direction, and the light only was polarized in the ω way, then it would scatter very little. So we must take average the cross section of an oscillator which can move in one direction, over all directions of incidence and polarization of the light or, more easily, we can imagine an oscillator which will scatter the field in either which way the field is pointing. Such an oscillator which can move equally in three directions, would have $3kT$ average energy because there are 3 degrees of freedom in that particular σ and not kT because of the 3 degrees of freedom.

Now we have to do the integral. Let us suppose that the uniform spectral distribution $N(\omega)$ of the light is a smooth curve and constant very very much above the very narrow frequency region where ω_0 is peaked (Fig. 41.3). Then the only significant contribution comes when ω is very close to ω_0 , within a small interval, which is very small. So therefore, although $\sigma(\omega)$ may be an unknown and unpredictable function, let's say plus where it is important, is near $\omega = \omega_0$, and there we may replace the strength curve by a flat-bottom "barrel" at the same height. In other words, we simply take $\sigma(\omega)$ outside the integral sign and call it $\sigma(\omega_0)$. We may also take the rest of the constants out in front of the integral, and what we have left is

$$(2\pi^2 k T \omega_0) \int_{\omega_0 - \gamma/2}^{\omega_0 + \gamma/2} \frac{d\omega}{\omega^2 - \omega_0^2 + \gamma^2/4} = 3kT. \quad (41.11)$$

Now, the integral starts from $\omega_0 - \gamma/2$, but this is so far from ω_0 that the curve is finished by then, so we go instead to $\omega_0 - \gamma$. It makes no difference and it is much easier to do the integral. The integral is an inverse tangent function of the form $\tan^{-1}(x) + C$. If we look it up in a book we see that it is equal to $\pi/2$. So what it reduces to for our case is $\pi/2$. Therefore we get, with some rearranging,

$$N(\omega_0) = \frac{3k^2 T}{4\pi^2 \omega_0^2}. \quad (41.12)$$

Then we substitute the formula (41.6) for σ_0 into (41.12) and we get σ_0 equal to ω_0 , since it is true or may we say just collinear with the formula for σ_0 that we know about.

$$\sigma_0 = \frac{c^2 k T}{4\pi^2 \omega_0^2}. \quad (41.13)$$

And this gives the radiatiton of light in a box formula. It is called the black

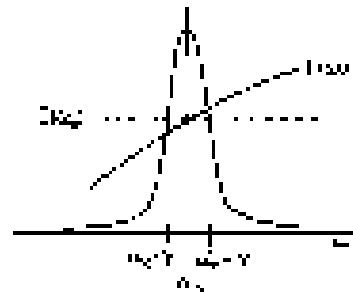


Fig. 41.3. The factors in the integrand (41.11). The peak is the spectrum curve ($N(\omega)$) $\sim \omega^2/4$. To a good approximation the factor that can be replaced by that

black radiation. That became the task in the famous thesis we took up in class when he was a student in 1900.

In a closed box at temperature T , (41.13) is the distribution law of energy of the radiation remaining in classical theory. First let's notice a remarkable feature of this expression. The charge of the oscillator, the mass of the oscillator, all properties specific to the oscillator, are not involved in this distribution function with its equilibrium. We must be at equilibrium with any other oscillator of different mass, or we will be in trouble. So this law imposes a kind of check on the properties that equilibrium does not depend on what we are in equilibrium with, but only on the temperature. Now let us draw a picture of the Rayleigh curve (Fig. 41-1). It tells us how much light we have at different frequencies.

The amount of intensity that there is in our box, per unit frequency range, goes down as the square of the frequency, which means that if we have a box at any temperature of T , then if we look at the x rays that are coming out, there will be a lot of them.

Of course we know this is false. When we open the surface and take a look at it, we do not burn our eyes and catch a fire at all. It is completely false. Furthermore, the total energy in the box, the sum of all the frequency spectrum over all frequencies, would be zero and under this infinite curve. Therefore, something is fundamentally, powerfully, and absolutely wrong.

This was the classic theory absolutely wrong, of course, concerning the distribution of light from a blackbody, just as it was incapable of actually describing the specific heats of gases. My students had better forth over this distribution from many different points of view, and there is no escape. This is the prediction of classical physics. Equation (41.13) is called Rayleigh's law, and it is the prediction of classical physics, and is inherently absurd.

41-3 Equilibrium and the quantum oscillator

The difficulty here was another part of the continual problem of classical physics, which also had with the difficulty of the specific heats of gases, and never has been resolved on the distribution of light in a blackbody. Now, of course, at the same time that theoretical work was being done, there were lots and many measurements of the actual curve. And I found out that the correct curve looked like the dashed curves in Fig. 41-1. That is, the x rays were not there. Low down the temperature the whole curve goes down in proportion to T , according to the classical theory, but the spectral curve also cuts off sooner at a lower temperature. Thus the low-frequency end of the curve is right, but the high-frequency end is wrong. Why? Well Sir James Jeans was worrying about the specific heats of gases, he noted that motions which have high frequency are "chopped out" as the temperature goes too low. That is, if the temperature is too low, if the frequency is too high, the oscillators don't have enough energy to keep going. Now recall our consideration of (41.13) without the aspects on the energy of an oscillator at thermal equilibrium. What did $\langle E \rangle$ ($\langle E \rangle = \langle E \rangle_0$) was, and what the sum at T in (41.13) is, is the mean energy of a 1-m molar oscillator at frequency ω at temperature T . Classically, this is $E_0/2$ experimentally, not necessarily the temperature is too low so the oscillator frequency is too high. At the same time, the curve falls off in the same way as does the specific heat of gases but it is easier to study the blackbody curve than it is the specific heat of gases, which are more complicated. Therefore our job is to proceed on determining the true blackbody curve. However, this curve is a curve which actually tells us, at every frequency, what the average energy of harmonic oscillators actually is as a function of temperature.

Planck studied this curve. He first determined the answer empirically, by fitting the observed curve with a law that can best guess very well. Thus he had to empirical formula for the average energy of a harmonic oscillator as a function of frequency. In other words, he had the right formula instead of E_0 , and then by fiddling around he found a simple definition for it which involved a new postulate. That assumption was that the harmonic oscillator can take up energies only in discrete steps. Before that they can have any energy whatever. Of course, that was not beginning of the end of classical mechanics.

The very first correctly determined quantum-mechanical formula will now be derived. Suppose that the possible energy levels of a harmonic oscillator were equally spaced in frequency, i.e., E_1, E_2, E_3, \dots . In reality, it could take on only these different energies (long, short). Planck made a somewhat more complicated argument than the one that is being given here, because that was the very beginning of quantum mechanics and he had to prove some things. But we are going to make it as short (which is demonstrated in this case) that the probability of occupying a level of energy E is $P(E) = e^{-E/kT}$. If we go along with this, we will obtain the right results.

Suppose now that we have a lot of oscillators, and each is a oscillator of frequency ω . Some of these oscillators will be in the lowest quantum state, some will be in the next one, and so forth. What we would have to know is the average energy of all these oscillators. To find out, let us calculate the total energy over the oscillators and divide by the number of oscillators. That will be the average energy per oscillator in thermal equilibrium, and will also be the energy that is in equilibrium with the "background radiation" and that should be at Eq. 14.13. In place of vE , then let N_ν be the number of oscillators that are in the ground state (the lowest energy state); N_ν the number of oscillators in the excited state; N_ν the number that are in state E_2 ; and so on. According to the hypothesis which we have now proved, that is, quantum mechanics, the law that replaced the principle of $e^{-E/kT}$ or $e^{-Eh\nu/kT}$ in classical mechanics is that the probability goes down as $e^{-Eh\nu/kT}$, where $Eh\nu$ is the excess energy, $h\nu$ the energy of the oscillator, that is, in the first state will be the number N_0 , that are in the ground state, times $e^{-Eh\nu/kT}$. Similarly, N_1 , the number of oscillators in the second state, is $N_1 = N_0 e^{-Eh\nu/kT}$. To simplify the algebra, let us call $e^{-Eh\nu/kT} = x$. Then we simply have $N_0 = N_0 x$, $N_1 = N_0 x^2$, ..., $N_n = N_0 x^n$.

The total energy of all the oscillators must then be worked out. If the oscillator is in the ground state, there is no energy. If it is in the n state, the energy is $E_n h\nu$, and there are N_n of them. So $N_0 E_0$, or $N_0 h\nu$ is how much energy we get from those. Those that are in the second state have $E_2 h\nu$, and there are N_2 of them, so $N_2 E_2 h\nu = N_2 h\nu^2$ is how much energy we get, and so on. Then we add it all together to get $E_{\text{tot}} = N_0 h\nu(0 + x + 2x^2 + 3x^3 + \dots)$.

And now, how many oscillators are there? Of course, N_0 is the number that are in the ground state, N_1 in the first state, and so on, are added together. $N_{\text{tot}} = N_0(1 + x + x^2 + x^3 + \dots)$. Thus the average energy is

$$\langle E \rangle = \frac{E_{\text{tot}}}{N_{\text{tot}}} = \frac{N_0 h\nu(0 + x + 2x^2 + 3x^3 + \dots)}{N_0(1 + x + x^2 + \dots)}. \quad (4.12)$$

Now the two sums which appear here we shall leave for my reader to play with and have some fun with. When we are all finished summing up, after subtracting for N_0 the sum "washout" if we make no mistakes in the sum,

$$\langle E \rangle = \frac{h\nu}{e^{h\nu/kT} - 1}. \quad (4.13)$$

This, then, was the first quantum-mechanical formula ever known, in 1900. Planck, and it was the "quantum" solution of decades of puzzlement. However, knew very then was turning wrong, and the question was, what was right? Here is the quantitative answer of what is right (constant of kT). This expression should, of course, approach kT as $x \rightarrow 0$ (as $T \rightarrow \infty$). See if you can prove that it does—learn how to do the mathematics.

This is the famous "exact factor" that Jeans was looking for, and if we use it instead of kT in (4.12), we obtain for the distribution of light in a black box

$$dN/d\lambda = \frac{h\nu^3 d\nu}{c^2 \lambda^5 (e^{h\nu/kT} - 1)}. \quad (4.14)$$

We see that for a large ν , even though we have a small numerator, that is more used in a denominator power in the denominator, as the curve comes down again and even not "blows up"—we never get a positive light wave—why?—that we did expect them!

E_1	$E_1 = 0$	$E_1 = \text{lowest energy}$
E_2	$E_2 = 2E_1$	$E_2 = \text{second energy}$
E_3	$E_3 = 3E_1$	$E_3 = \text{third energy}$
E_4	$E_4 = 4E_1$	$E_4 = \text{fourth energy}$
E_5	$E_5 = 5E_1$	

Fig. 41-2. The energy levels of a harmonic oscillator are evenly spaced.

You might complain that in our derivation of (41.17) we used the quantum theory for the energy levels of the harmonic oscillator and the classical theory in determining the cross sections. The nonclassical theory of light interacting with a harmonic oscillator gives exactly (41.17), a result that agrees with the classical theory. That is just as it was when just that in studying vacuum fluctuations our analysis of the theory of refraction and the scattering of light, using a model of atoms like that of Section 18, the quantum mechanics substantially disagreed.

Now let us return to the Johnson noise in a resistor. We have already mentioned that the theory of this noise power is exactly the same theory as that of the classical Blackbody Radiation Law. In fact, rather amazingly, we have already done this. When we did it, in a circuit with zero resistance, our "radiation resistance" had no effect because it radiates nothing; a radiation resistance as it would be easy for us to understand, the power would be. It would be just the power loss in multiplying the spectrum from the light that is all around, and we would get the same distribution, unchanged by such noise in two terms. We can also prove that the source is a generator with an unknown power (just from (41.1)). The spectrum is determined by the fact that this noise generator, connected to a resonant circuit at say frequency ω , in Fig. 41.1(b), generates in the inductance a voltage $\sim \omega I$ (approximately as in Fig. 41.2). One has just in the same integral as in (41.10), and the same method works to give Eq. (41.11). For low temperatures (say 57 K) (41.21) must of course be replaced by (41.15). The two theories of resistive noise in solid conductors may be also closely related physically, for we may of course connect a resonant circuit to the resistor, so the resistance is a pure resistive resistance. Since (41.9) does not depend on the physical origin of the resistance, we know the generator to be a real resistance and for resistance resistance is the same. What is the origin of the generated power? First of all the resistance R is only an ideal current in e.g. like number its cross-section at temperature T ? If it is a radiation source, the question is, at what temperature? which impinges on the inductor and is "received again," makes an effective generator. Therefore one can deduce a direct relation of P_{rad} and P_{int} , leading from (41.15) to (41.1).

All the things we have been talking about—the cooled Johnson noise and Planck's distribution, and the exact theory of the Brownian movement—will be considered further in the development of the first theory of noise in the resistivity. Now with these points and that history in mind, we return to the Brownian movement.

41-4 The random walk

Let us consider now the position of a jiggling particle whose change with time, we say, is x , is compared with the time between "steps." Consider a "true" Brownian movement—particle which is jiggling about because it is surrounded on all sides by irregularly jiggling water molecules. (Query: After a given length of time, does the noisy x in (41.1) have to be bunched along? That is, is it often surrounded by lone air and sometimes by water? If we imagine that we divide the time into Δt intervals, let us say a millionth of a second or so, then after the first hundredth of a second it moves here, and in the next hundredth it moves some more, and in the next hundredth it moves to another somewhere else, and so on. In terms of the rate of bounces, n , and with Δt a second, n is very large. The reader may easily verify that the number of collisions a single molecule with water molecules in a second is about 10^{11} , or a hundred million second is about 10^{13} collisions, which is not. There is also a limit on the "bounce rate" per unit time, but what happens before? In other words, how collisions are all measured so that one "step" is not related to the previous "step"? It is like the famous drunken sailor problem: Can you always get home, and when a sequence of steps has each step is chosen at an arbitrary angle, at random (Fig. 41-6)? The situation is: After a long time, where is the sailor? Of course we do not know! It is impossible to say. (Perhaps we mean he is just somewhere near or has gotten "lost." But, then, in the average, where is he?) On the average, how far off from the spot you've gone? We have already answered this question. (Indeed, that is what we were discussing, the

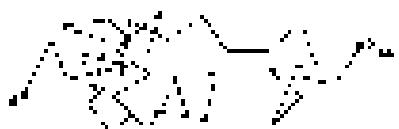


Fig. 41-6. A random walk of 30 steps of length L . Here for $L = 3\lambda$, from 37 days about 6000 λ away.

size position of light from a whole lot of different sources at different angles, and (2) meant adding a list of several different angles (Chapter 12). There, we discovered that the mean square of the angle, from one end to the other of the chain of random steps, which was the intensity of the light in the sum of the intensities of the separate pieces. And so, by the same kind of argument as we did above immediately that if R_N is the mean distance from origin after N steps, the mean square of this distance from the origin, is proportional to N^2 , or, in other words, $\langle R_N^2 \rangle = N^2$, where L is the length of each step. Since the number of steps is proportional to the time in our present problem, the mean square distance is proportional to the time.

$$\langle R^2 \rangle = \text{const} \quad (11.13)$$

This does not mean that the mean distance is proportional to the time. If the mean distance were proportional to the time it would mean that the driving is at a nice uniform velocity. The only way making some steady sensible trajectory, however, such that the mean square distance is proportional to time, that is the characteristic of a random walk.

We may show very easily that in fact $\langle R^2 \rangle$ among the squares of the distances increases, on the average, by L^2 . For, start with $R_0 = R_{N-1} + L$, we find that R_N^2 is

$$R_N \cdot R_0 = R_0^2 + R_{N-1}^2 + 2R_{N-1} \cdot L + L^2,$$

and averaging over many trials, we have $\langle R_N^2 \rangle = \langle R_0^2 \rangle + L^2$, since $\langle R_{N-1} \cdot L \rangle = 0$. Thus, by induction

$$R_N^2 = \text{const}^2 \quad (11.14)$$

Now we would like to calculate the last constant in Eq. (11.13), and here we are most likely to trip up. We are going to suppose that if we were to put a force on the particle, nothing would go with the Brownian movement. We are taking a side track for the moment; then it would just increase following way again. In fact, there would be inertia, m , or, at the coefficient of inertia, the effective mass of the object, not necessarily the sum of the real mass of the solid particle, because the water has to move around the particle if $m \neq M$ (unlike). Thus if we talk about myself, I would realize, there is a term like m/L^2 , when one side. And next, we want also to assume that if we kept a steady p_\perp on the object, there would be a drag on it from the fluid, proportional to its velocity, v . Besides the inertia of the fluid, this is a realistic assumption due to the viscosity and the complexity of the fluid. It is absolutely essential that there be some irreversible losses, something that is irreversable, in order that the law $F = \dot{p}$ takes place. There is no way to prevent the fluid under these conditions to be source of the F in the form's being closely related to these losses. We'll the mechanism of this drag is *viscous dissipation*, or, in words talk about forces that are proportional to the velocity and where they come from. That let us suppose to have that there is such a viscous η . Then the formula for the net force due to external forces, whenever we are putting out from motion, is

$$M \frac{d^2x}{dt^2} + \eta \frac{dx}{dt} = F_{\text{ext}} \quad (11.15)$$

One quantity is now determined directly from experiment. For example, water falls in a drop from the ceiling. Then we know that the force is mg , and g is to be divided by the square of L , the drop falling velocity v_0 . Or we could put the lamp in a room, and see how fast it wobbles. Or if it is charged, we can put another field on it. Since it is measurable thing, it is an artificial thing, and it is known for every type of artificial particles etc.

Now let us use the same formula in the case where the force is a constant, not equal to the singular forces of the source, or, in general. We shall then try to determine the mean square distance for the object goes. Instead of taking the distances in three dimensions, let's take the first dimension, and find by means of $\langle x^2 \rangle$, not to express ourselves. Obviously the mean of x^2 is the same as the mean of x , is the square of the size of x , and therefore the mean square of the distance

In just 1 line what we are going to take, since The averagement of the irregular forces is, of course, just as irregular as any other component. What is the rate of change of x^2 ? It is $d(x^2)/dt = 2x(dx/dt)$, so what we have to find is the average of the positive times the velocity. We shall show that this is a constant, and that therefore the mean square radius will increase proportionally in the time, and at what rate? Now, we multiply Eq. (41.19) by $m \frac{d(x^2)/dt}{dx} = m(2x) \frac{dx}{dt} = 2x^2$. We want the time average of $x^2(dx/dt)$, so let's take the average of the whole equation, and multiply the three terms. Now this is a little like Eq. 3.27. If the particle happens to leave your system at distance x from it, since the irregular force is completely irregular and does not know where the particle started from, the next impulse could be in any direction relative to x . It is positive; then in the reason why the average force should then be in that direction. It is just as likely to be one way as the other. The component forces are not during t in a definite direction. So the average value of x^2 times F is zero. On the other hand, for the term $m(x^2)^2/dt^2$ we will have to be a little fancy, and write this as

$$m \frac{d^2x}{dt^2} = m \frac{d(x^2)/dt}{dt} = m \left(\frac{dx}{dt} \right)^2.$$

Now we put in these two terms and take the average of both. So let us see how many x times the velocity should be. Now x times the velocity has a sense that does not change with time, because when it gets to some position it has no remembrance of where it was before, so things are no longer averaging with time. So this quantity, or the average, is zero. We have left the constant, m^2 , and that is the only thing we know, $m^2/2$ is a mean value $\langle x^2 \rangle$. Therefore we find that

$$\left\langle m^2 \frac{d^2x}{dt^2} \right\rangle + \langle m \frac{d(x^2)/dt}{dt} \rangle = \langle x^2 \rangle,$$

implies

$$- \langle m x^2 \rangle + \frac{m}{2} \frac{d}{dt} \langle x^2 \rangle = 0,$$

or

$$\frac{d\langle x^2 \rangle}{dt} = \frac{m}{2} \langle x^2 \rangle. \quad (41.20)$$

Therefore the object has a mean square distance $\langle x^2 \rangle$, at the end of a certain length of t , occurs to

$$\langle x^2 \rangle = \langle x_0^2 \rangle e^{mt^2/2}. \quad (41.21)$$

And so we can actually determine λ for the particles g . We first must determine how many x equal to a theory force, say, that they drift under a known force F for t , and then we can determine how far they go in their random motion. This equation, was of considerable importance historically, because it was one of the first ways by which the constant λ was determined. After all, we can measure λ , the time, how far the particles go, and we can take an average. The reason that the determination of λ was important is that in the law $\langle x^2 \rangle = R t$ for a move we know that R , which can also be measured, is equal to the number of atoms in a mole times λ . A mole was originally defined as a mole of many grams of oxygen—6 (now carbon is used), so the number of atoms in a mole was not known, originally. This, of course, a very interesting and important problem. How big is λ alone? How many atoms? Some of the earliest determinations of the number of atoms was by the determination of how far a single little particle would move if we pushed it patiently until a microscope for a certain length of time. And thus Boltzmann's constant k and the Avogadro number N_A were determined to values λ had already been measured.

Applications of Kinetic Theory

42.1 Evaporation

In this chapter we shall discuss some of the applications of kinetic theory. In the previous chapter we emphasized one particular aspect of latent heat, namely, that the change in latent heat is one degree of freedom of a molecule or other object. Now, the mutual motion of what we call "molecules" on the other hand, is the fact that the probability of finding a particle in different places, per unit volume, is $e^{-E/kT}$. From $e^{-E/kT}$, we shall calculate a number of applications of rays.

The phenomena which we want to study are relatively complicated: a liquid evaporating, a substance in a metal coming out of the surface, or a chemical reaction in which there are a large number of intermediate steps. In such cases it is no longer possible to make from the kinetic theory only simple and correct statements, because the situation is too complicated. Therefore, this chapter, except where otherwise emphasized, is quite modest. One idea to be emphasized is only that we can understand from the kinetic theory, how or how many things ought to be how. By using thermodynamic arguments, or some empirical measurements of certain related quantities, we can get a more accurate representation of the phenomena.

However, it is very useful to know more, not so much why something behaves as it does, so that when the situation is a new one, we can use what we have not yet learned. In addition we can learn more about what ought to happen. So this discussion is highly inaccurate but essentially right—right in idea, not in the fine detailed, let us say, in the specific details.

For example, first, we shall consider is the evaporation of a liquid. Suppose we have a box with a large volume, partially filled with vapor in equilibrium and at the vapor at a certain temperature. We shall suppose that the molecules of the vapor are relatively far apart, and that inside the liquid, the molecules are packed close together. The problem is to find out how many molecules there are in the vapor at a given temperature, and how this depends on the temperature.

Let us say that n equals the number of molecules per unit volume in the vapor. Then, however, N equals nV at the box temperature. If we add heat, we get more evaporation. Now let another quantity, N_1 , equal the number of atoms per unit volume in the liquid. We suppose that each molecule in the liquid occupies a certain volume, σ^3 . If the N_1 are next to each other, they then all together they occupy a bigger volume. But if N_1 is the volume occupied by one molecule, the number of molecules in a unit volume is N_1 and volume divided by the volume of each molecule. For definiteness, we suppose that densities of molecules between the molecules is to help them together in the liquid. Otherwise we cannot understand why it condenses. This requires that there is a force and that there is an energy of binding of the molecules in the liquid which is less when they go into the vapor. That is, we are going to suppose that, in order to take a single molecule out of the liquid into the vapor, it is fair amount of work W has to be done. There is a certain difference, E , in the energy of a molecule in the vapor from what it would have if it were in the vapor, because we have to pull it away from the other molecules which attract it.

Now we get the general principle that the number of atoms per unit volume in two different regions is $n_1 e^{-E/kT} = n_2 e^{-E/kT}$. Let N_1 be the number of particles per unit volume in the vapor divided by N_1 times N_1 per unit volume in the liquid, is equal to

$$N_1' = e^{-E/kT}. \quad (12.1)$$

42.1 Evaporation

42.2 Thermodynamics

42.3 Thermal insulation

42.4 Thermal insulators

42.5 Einstein's law of radiation

because $\ln \frac{p}{p_0}$ is the general rule. It is fine to assume that in equilibrium under gravity, vapor density is proportional to density just below the top because of the work required to lift the gas system to the height h . In the liquid, the molecules are thrown from the vapor because we have to put the latent energy "into" h , and this is if the densities is $e^{-\frac{U}{kT}}$.

This is what we wanted to prove—that the vapor density varies as $e^{-\frac{U}{kT}}$ the more precisely we do it, over 10⁻³⁰ to 10⁻³² the better it is and is extremely interesting to us, because in most cases the vapor density is very small. Now if the liquid density ρ , in these circumstances, were constant, then for the vapor density where they are almost the same, but where the vapor density is much lower than the liquid density, then the fact that ρ is very much less than $1/V_0$ is compensated by the factor ρ is very much greater than $1/V_0$. So to make such as (42.1) are interesting only after ρ is very much larger than $1/V_0$, because in most circumstances, since we are trying to measure a tremendous amount if we change V a little bit. But temperature does change a lot, and the change is indeed kT . The equilibrium factor is very much more important than any change that might occur in the factors not in front. Why should there be any change in such factors as V_0^2 ? I suppose this was an appropriate analysis. After all, there is no really a definite velocity for one molecule, so we change the temperature, the volume V , this is also constant—the liquid expands. There are other little factors like that, and so on, and allusion is more meaningful. There are slowly varying temperature dependent factors all over the place. In fact we find that V_0^2 does vary slightly with temperature, because at a higher temperature, at a kT constant molecular volume, there would be different average attractions, and so on. So while we might think that if we have a formula in which every thing varies at an unknown way with temperature then we have no formula at all. We realize that the exponential factor is, in general, very large, except that in the case of the vapor density as a function of temperature most of the variation is accommodated by the exponential factor and it is to a first moment and the coefficient $1/V_0$, as usually constant, is a good approximation for most temperatures the case. Most of the variation, in other words, is of the general nature $e^{-\frac{U}{kT}}$.

In fact, well then there are many, many phenomena in nature which are characterized by having to use up energy from somewhere, and in which the central feature of the temperature variation is $e^{-\frac{U}{kT}}$ minus the energy over kT . This is a useful fact only when the energy is large compared with kT , without most of the variation is contained in the variation of the $e^{-\frac{U}{kT}}$ and not in the constant and in other factors.

Now let us consider another way of obtaining a somewhat similar result for the vaporization, but seeking still in more detail. To prove it off, we simply suppose a rule which is valid at equilibrium, but in order to understand things better there is no harm in trying to look at the details of what is going on. We may also assume the what is going on in the following way. The molecules can and do leave continually the surface of the liquid; when they do it, they may fly off or they may get stuck. There is an unknown factor for that—maybe 50%, maybe 10 to 90% we do not know. For many the others get stuck—we can assume this over again later on the assumption that they do not always get stuck. Then we can calculate the N_0 , the total number of atoms which are碰撞ing only the surface of the liquid. The number of molecules moving, the number that leave on a unit area, is the number n per unit volume times the velocity v . This velocity of the molecules is related to the temperature, because we know that v^2 is proportional to kT on the average. So it is some kind of a mean velocity. Of course we should integrate over the v 's to get some kind of an average, but it is roughly proportional to the root-mean-square velocity, within some factor. This

$$N_0 = n v \quad (42.2)$$

is the number which units per unit area and are colliding.

At the same time, however, the atoms in the liquid are jiggling about, and from time to time one of them gets kicked out. Now we have to estimate how fast they get kicked out. The idea will be that at equilibrium the number that are kicked out per second and the number that come per second are equal.

How many get kicked out? In order to go kicked out, a gas molecule has to have acquired by accident an excess energy over its equilibrium—a considerable excess energy, because it is all used up straight for the other molecules in the liquid. Ordinarily it does not have that, so it is only slightly off track, but in the collisions sometimes one of them gets an excess energy by accident. And the chance that it gets the extra energy, if which it needs . . . our case is very small $\approx 10^{-10}$. In fact, $\approx 10^{-10}$ is the chance that a collision has picked up enough to do it much energy. That's the problem in kinetics theory; in order to know an excess energy IP have to suppose, he believes at the time the energy that we have to know, over E_0 . Now suppose that some molecules have been given this energy. We now have to estimate how many leave the surface per second. Of course, just because a molecule has the necessary energy does not mean that it will actually evaporate, since it may be buried too deeply inside the liquid or, even if it is near the surface, it may be travelling in the wrong direction. The number that are going to leave a unit area per second is going to be something like this. The number of course there are near the surface, per unit area, covered by the liquid. Is one atom a centip, and is likely the probability $\approx 10^{-10}$ that they are ready to escape in the sense that they have enough energy.

We shall suppose that each molecule at the surface of the liquid has plus a certain excess energy A . Then the number of molecules per unit area of liquid surface will be N_0 . And now, how long does it take a molecule to escape? If the molecule has a certain average speed v , and goes to move, say, one molecular diameter d , the thickness of my film layer, then the time it takes to get across that thickness is the time needed to escape, if the molecule has enough energy. The time will be $t = d/v$. Thus the number evaporating should be approximately

$$n = (N_0 d v / 2) e^{-A/kT} \quad (42.3)$$

Now the area of each atom, times the thickness of the layer, is approximately πd^2 , so if the volume V_0 occupied by a single atom. And we consider again benzene, we must have $N_0 = V_0 / \pi$.

$$n = (V_0 v / 2) e^{-A/kT} \quad (42.4)$$

We may assume the v 's since they are equal; even though one is the velocity of a molecule in the vapor and the other is the "velocity" of an evaporating molecule. These are to be set, because we know their sum, known energy (in one dimension) is A . But one may object, "Well No! These are the especially fast-moving ones; these are the ones that have probably excess energy." Not really, because the moment they start to pull away from the liquid, they have to lose the excess energy against the potential energy. As we decrease in the surface they are pulled down to the velocity; this is the same as was in our discussion of the distribution of molecular velocities in the atmosphere. At the bottom, the molecules had a uniform distribution of energy. The ones that came up, the ones that have the same distribution of energy, because the new ones are not important, and the old ones were thrown down. The ones that are evaporating have the same distribution of energy as the ones inside—a quite remarkable fact. Anyways, it is useless to try to impose artificially slow, or normal, or excess velocities, never, such as the probability of breaking back rather than leaving the "liquid" & so on. Let us now, a rough idea of the rate of evaporation and condensation, and we see, of course, that the vapor density varies in the same way as before, but now we have understood it in a systematical rather than just a qualitative, *ad hoc* mode.

This deeper understanding permits us to analyze some things. For example, suppose that we start evaporates the water at such a great rate that we removed the vapor as fast as it formed. "we had very good pump, and the liquid was evaporating very slowly"; how fast would evaporation be? . . . if we maintained a liquid temperature $\approx T^*$. Suppose that we have already experimentally measured the equilibrium vapor density, at this we know, at the given temperature. How many molecules per unit volume are in equilibrium with the liquid? Now we would like to know *how fast* it will evaporate. Even though we have used only a rough cut *ad hoc* for as the evaporation part of it is concerned. The number of vapor

unknown quantity was now there as badly, aside from the unknown factor of refection coefficient. We therefore we may let the first law number that are leaving at equilibrium, is the same as the number that are hitting. Thus, if vapor is very evanescent and if the molecules are only coming out, but if the vapor were left alone, it would remain in equilibrium - that is, at which the number that come back would equal the number that are evaporating. Therefore, we can easily see that the number that are coming out the surface we denote is equal to the unknown reflection coefficient θ times the number that would come down on the surface second were the vapor still there, because that is how many go to the balancing evaporation in equilibrium:

$$N_1 = \theta N_0 = (\sigma R T_{\text{atm}})^{-\frac{1}{2}} \cdot \theta. \quad (18.1)$$

Of course, the number of molecules that hit the liquid from the vapor is very, very large, since we do not need to know so much about the forces as we do when we are worrying about how they go in and get through the liquid surface; it is much easier to make the argument the other way.

47.2 Thermionic currents

We now give another example of a very practical situation that is similar to the evaporation of a liquid—namely, that it is not water striking a separate surface. It is basically the same problem. In a metal, there is a source of electrons, namely, a heated tungsten filament, and a positively charged plate to attract the electrons. Any electron that escapes from the surface of the tungsten is immediately attracted to the plate. This is our ideal "pump," which is "pulling" the electrons away from the wire. Now the question is: How many electrons per second can we get out of a piece of tungsten, and how does that number vary with temperature? The answer to this problem is the same as (18.1) because it turns out that one process of most electrons is extracted to the ions, or to atoms, of the metal. They are attracted, or may stay in, to the metal. In order to get an electron out of a piece of metal, it takes a certain amount of energy or work to pull it out. This work varies with both the substance of metal. In fact, it varies even with the character of the surface of a given kind of metal. But the total work may be a few electron volts, which, incidentally, is typical of the energy it takes in chemical reactions. We can remember the battery by noting that the voltage in a dry cell will like a flashlight battery, which is produced by chemical reactions, is also a constant.

How can we find out how many electrons come out per second? It would be quite difficult to analyze the effects on the electrons going out; it is easier to analyze the situation the other way. So, we could start out by imagining that we'd not let the electrons out, and that the electrons were held, just, and could come back to the metal. Then there would be a certain density of electrons at equilibrium which would, of course, be given by exactly the same formula as (18.1), where Z_e is the volume per electron in the metal, roughly, and \mathcal{W} is again eV , where V is the work of work function, or the voltage needed to pull an electron off the surface. This would tell us how many electrons would have to be in the metal holding constant striking the metal in order to balance the ones that are coming out. And thus it is easy to calculate how many are coming out if we sweep away all of them, because the number that are coming out is really equal to the number that can fit by going in with the above density of electron "trap." In other words, the answer is just the current of electricity that can be just one ampere equal to the charge on each times the number that comes per second per volt amp, which is the number per unit volume times the velocity, as we have seen many times.

$$I = q \cdot n = (q \cdot \sigma V_e) e^{-\frac{\mathcal{W}}{kT}}, \quad (18.2)$$

Now one electron volt corresponds to 1.602×10^{-19} coulombs. At 1000 degrees, the lifetime τ of the tungsten is 10 seconds at a temperature of, say, 1000 degrees, so the conversion factor is simply $(10^3)^{-1}$; when we change the temperature to

Since λ , the exponential factor changes a lot, thus, again, the central feature of the formula is $e^{-\beta E}$. As a matter of fact, the factor in front is quite interesting— it turns out that the behavior of electrons in metal is not correctly described by the classical theory, but by quantum mechanics, our λ only changes the factor in front a little. Actually, no one has ever been able to get the things right even not very well, even though many people have used the high-class quantum-mechanical theory for their calculations. The big problem is, does H' change a little with temperature? If it does, one cannot distinguish it from changing slowly with temperature but one can distinguish it from H' if it changes linearly, say, with temperature, so that $H' = H_0 - \alpha T$, then we would have

$$e^{-\beta H'} = e^{\beta H_0} e^{-\alpha T \beta} = e^{-\alpha T \beta} e^{-\beta H_0}.$$

Thus a linearly temperature-dependent H' is equivalent to a "slope" ("constant"). It is really quite difficult and usually impossible to try to obtain this coefficient in the first accuracy.

42-3 Thermal ionization

Now we go on to a clear example of the same idea: always the same idea. This time let's do with ionization. Suppose that in a gas we have a whole lot of atoms which are in the neutral state, say, but the gas is hot and the atoms can become ionized. We would like to know how many ions there are in a given circumstance if we know a certain density of atoms per unit volume at a certain temperature. Again we consider a box in which there are N atoms which can hold electrons. (If an electron has come off a atom, it is called an ion, and if the atom is neutral, we simply call it an atom.) Then suppose that, at any given moment, the number of neutral atoms is n_0 , the number of ions is n_1 , and the number of electrons is n_2 , all per unit volume. The problem is: What is the relationship of these three numbers?

In the first place, we have two conditions or two pairs of the numbers. For instance, as we vary different conditions, like the temperature and so forth, n_0 , n_1 , would remain constant, because this would be simply the number N of atomic nuclei that are in the box. If we keep the number n_0 fixed per unit volume fixed, and change, say, the temperature, then as the ionization proceeded some atoms would turn to ions, but the total number of positive plus negative would be unchanged. That is, $n_0 + n_1 + N$. Another condition is that if the entire gas is to be electrically neutral (and if we neglect double ionization), this means that the number of ions is equal to the number of free negative charges, or $n_1 = n_2$. These are certainly requirements that simply express the conservation of charge and the conservation of atoms.

These equations are true, and we ultimately will use them when we consider a real problem. But we want to obtain another relationship between the quantities. We can do this as follows. We again use the idea that it takes a certain amount of energy to lift the electron out of the atom, which we call the ionization energy, and we let V be the volume. In order to make all of the numbers look the same, we let H' equal the energy needed to pull an electron out of an atom and m be the mass. Now we say it's say that the number of free electrons per unit volume is the "mass" is equal to the number of bound electrons per unit volume in the x -axis, times e to the minus the energy difference between being bound and being free, times λ . That is the basic equation again. How can we write it? The number of free electrons per unit volume would, of course, be n_2 , however that is not definition of n_2 . There will be a number of electrons per unit volume that are bound to atoms. The total number of places that we could put the electrons is approximately $n_0 + n_1$, and we will suppose that when they are bound you ignore being within a certain volume V . So the total number of volume which is available to electrons which would be bound is $(n_0 + n_1)V$, so we might want to write it as follows:

$$n_2 = \left(\frac{n_0}{(n_0 + n_1)V} \right) e^{-\beta E_0}.$$

The formula is wrong, however, in one essential feature, which is the following: σ electron is energy or not cross. Another electron cannot come to the volume occupied. In other words, all the volumes of all the possible sites are not really available for the new electron. When trying to make up its mind whether it's to be in x -direction or in the conjugated position, because in this particular there is an available feature that when one electron is where another electron is, it is not allowed to go there & repeat. For that reason, it comes out that we should count only that part of the volume which is available for another nucleation event. That is, those sites are already occupied - so no point in the total available volumes, but the only volumes which is allowed is rest of the site, where the one present does not let the electron to go. Thus, in these circumstances, we find that a better way to write our formula is

$$\frac{d\sigma}{dt} = \frac{1}{V_0} e^{-E/kT} \quad (16.5)$$

This formula is called the Saha ionization equation. Now let us see two other undesirable complications why a formula like this is right by arguing about the same things we are comparing.

First, every once in a while an electron comes to another and they combine to make a atom. And it may come in a while, an atom goes into a collision and breaks up into an ion and an electron. Now these two rates must be equal. How fast do electrons and ions "find each other"? The rate is certainly increased if the number of electrons per unit volume is increased. It is also increased if the number of ions per unit volume is increased. That is, the rate at which recombination is occurring is certainly proportional to the number of electrons times the number of ions. Now consider this at which ionization is occurring due to collisions must be dependent. Heavily on how many atoms there are in volume. And so the rates will balance when there is some relationship between the product of n_e and the number of ions, n_i . This is what the relationship happens to be given by this particular formula, where kT is the ionization energy, but it's a little bit more in question, but we can easily understand that the formula would necessarily involve the concentrations of the electrons, ions, and atoms in the ionization and/or to conclude a result it is independent of the kT , that depends only on E , kT , n_e , n_i . The other three factors are other constant factors.

We may also note that, since the equation involves the numbers per unit volume, if we were to do an experiment with a given total number N of some substance, that is, a certain fixed number of atoms, but using boxes with different volumes, the n 's would all be smaller in the larger box. But since V_0 is constant, since the same, the total number of electrons and ions must be greater in the larger box. To see this, say, say that there are N nuclei inside a box of volume V , and the electron concentration is denoted. Then $n_e = N/V \approx n_i$, and $n_i = N - N/V$. Then, our equation becomes

$$\frac{d^2N}{dt} \approx - \frac{e^{-E/kT}}{V_0} \quad (17.6)$$

In other words, if we take a small V and smaller density of electrons, we have the volume of the our sites bigger and bigger, so each of the electrons had less room increase. And ionization, and then "expansion" is the density goes down, is the reason why we believe that at very low densities, n_e and n_i in the solid state between them, there may be ions present, even though we didn't understand it from the point of view of the available one σ . At density n , takes many, many kT of energy to make them, there are ions present.

Why can there be ions present when there is so much space around, while σ is infinite the density, the ions tend to disappear? Answer: Consider an ion. Every atom in a solid, light or another atom, or an ion, or whatever it is just maintains thermal equilibrium, strikes it. Very hard, because it takes such a little amount of energy, an electron comes off and on the is left. Now the electron, if the space is enormous, wanders and wanders and does not come near anything for years, perhaps. But once in a very long time, it does come back to V .

in motion, and then continue to make collisions. So the rate at which electrons leave moving out from the source is very low. But if the voltage is very small, an electron which has escaped takes so long to find another ion to recombine with that probability of recombination is very, very small, thus, despite of the large excess energy source, there may be a reasonable number of electrons.

42-4 Chemical kinetics

The same idea for the way we have just called "ionization" is also found to be the case in reaction. For instance, if two atoms *A* and *B* combine to form a compound *AB*, then if we think about it, we will observe the *AB* molecule is formed in form *A* \pm *B*. Well we call an electron, and it is what we mean by ion. Well, these calculations the equations of equilibrium are exactly the same in form:

$$\frac{d[R]}{dt} = k_1[A][B] - k_2[R]^2 \quad (42-3)$$

This formula, of course, is not clear, since the "probability" depends on how much value is allowed for the *A* and *B* atoms, and so on, but by the hydrodynamic argument we can identify what the meaning of the *kF* in the coefficient is, and it turns out that it is very close to the energy needed in the reaction.

Suppose that we want to understand it by formulating a model of collision as much as the way that we understand the evaporation formula, by asking about how many electrons come out & how many of them come back *A* & *B* and *C*. Suppose that *A* and *B* something is occurring every time one *A* will be formed compound *AB*. And suppose that the compound *AB* is a complicated molecule which breaks around and is hit by other molecules and for some time it sees enough energy to break and hasn't enough time into *A* and *B*.

Now it actually turns out, in chemical reactions, most of the atoms, come together with *A* and *B* and even though they may be released to the reactant *A* + *B* \rightarrow *AB*, the fact that *A* and *B* may bump each other does not necessarily make the reaction start. It usually is required that the collision be rather hard, i.e. fast enough to "knock" a "soft" collision between *A* and *B* apart, but doing even enough energy may be necessary in the process. So let's suppose that it is very common in chemical reactions that in order for *A* and *B* to form *AB*, they must just hit each other, but they have to hit each other with sufficient energy. This energy is called the activation energy—the energy needed to "knock" the reaction. Of course activation energy, the energy-energy needed in a collision in order that the reaction may really occur. Then the rate *R* at which *A* and *B* produce *AB* would involve the number of atoms or *N* times the number of atoms *N*/*R*, times the rate at which a single atom would do a certain cross section σ , times a factor $e^{-E/RT}$, which is the probability that they have enough energy:

$$R = N \sigma v N e^{-E/RT} \quad (42-4)$$

Now we have to find the opposite rate, *R*. The rate at which chance that *AB* will split. In order to split, it not only must have the energy *E* which it needs to break, it gets over, but, just as it was harder for *A* and *B* to come together, so there is a kind of hill top. And it has to climb over the activation energy, *E*, they don't have just enough energy just to get ready to pull apart, but a certain amount. So the climbing a hill up to a deep valley, they have to climb the hill coming in and they have to climb out of the valley and then over the hill come up back (Fig. 42-1). Thus the rate at which *AB* goes to *A* and *B* will be proportional to the number *N*/*R*, the one present times $e^{-E/RT}$:

$$R = N \sigma v N e^{-E/RT} \quad (42-5)$$

The *v* will depend the volume of atoms and the rate of collisions, where we can work out as $v = \pi d^2 n / 3$, d is the diameter and n is the density and thickness, but we shall not do this. The main feature of interest to us is that when these two

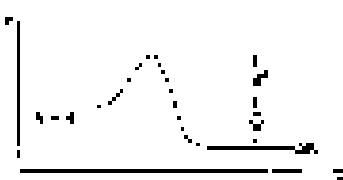


Fig. 42-1. The energy relationship for the reaction *A* + *B* \rightarrow *AB*.

ratio are equal, the ratio of their is equal to unity. This allows that $k_{\text{eff}} = k_1 k_2 \dots k_n = \frac{1}{e^{-E_1/kT} + e^{-E_2/kT} + \dots + e^{-E_n/kT}}$, where E_i denotes the activation energies, v_i denotes the collision velocities, and k is the Boltzmann constant.

The interesting thing is that the rate of the reaction also varies as $e^{-E/kT}$, although the exponent is not the same as that which governs the overall velocity. The activation energy A^* is quite different from the energy E . If you vary the proportion of A , B , and C that are left at equilibrium, then if you want to know how the A^* changes (Fig. 4.9), that is not a question of equilibrium, and there is a different energy, the activation energy A^* , given by the ratio of reaction times² an exponential factor:

Furthermore, A^* is not a fundamental constant like R . Suppose that at the surface of the wall—say at some other place— A and B could temporarily stick there in such a way that they could combine more easily. In other words, we might find a "tunnel" through the hill, or perhaps a lower hill. By the conservation of energy, when we let all particles we have still made ΔE in A of A and B , so the energy difference. If w^* be quite independent of the way the reaction occurs, but the activation energy A^* will depend very much on the way the reaction occurs.

This is why the rates of chemical reactions are very sensitive to outside conditions. We can change the rate by putting A on top of a different kind, we can put it in a "different barrel" and it will go at a different rate, that depends on the nature of the surface. Or, if we put in a third kind of object C , may change the rate very much since things produce enormous changes in rate simply by adding the A^* a little here—they are called catalysts. A reaction might accidentally not occur at all because $A^* > kT$ (Fig. 4). The given temperature, but when we put in this special stuff, the catalyst, then the reaction goes very fast indeed, because A^* is reduced.

Incidentally, there is some trouble with such a reaction, I think. Imagine A , B , C do not interact conserve both energy and momentum when they bump two objects together to make one that is more stable. Therefore, we need at least a rare object C , so the overall reaction is much more complicated. The forward rate would involve the product momenta, and it might seem that one formula is giving wrong, but not. What we look at the rate at which A 's goes, he takes away, we find that it has to collide with C , so there is no change in the reverse rate; that's carried out in the formula for the equilibrium concentration. They are in equilibrium, (4.2), which we first wrote down is obviously guaranteed to be true, no matter what the mechanism of the reaction may be.

42.5 Einstein's law of radiation

We now turn to calculating the energy distribution coming out with the black body radiation law. In the last chapter we worked out the alternative view to the radiation is to carry the way Planck did, considering the radiator, common oscillator. The oscillator had to lose its mean total energy, and since it was oscillating, it would radiate until would keep pumping radiation into the cavity until it piled up enough radiation to balance the absorption and emission. In fact, very few, until just the intensity of emission at frequency ν was given by the formula

$$I(\nu) d\nu = \frac{A \nu^3 h}{e^{h\nu/kT} - 1}. \quad (42.19)$$

This result involved the assumption that the oscillator whose was generated, radiation had definite, equally spaced energy levels. We did not say that light had to be a photon or anything like that. There was no discussion about how when an atom goes from one level to another, the energy must come out at one unit of energy. As, in the theory of light, Planck's original idea was that the nuclear was quantized but not the light, meaning oscillators can't take up just any energy, but have to take it in lumps. Furthermore, he trouble was the derivation is that it was partially circular. We calculated the rate of radiation from an oscillator according to classical physics, then we turned around and said, "No, the oscillator has a lot of energy levels. So gradually, as we do that, the light with the completely quantum mechanical result, from which a slow development.

ment which culminated in the quantum mechanics of 1927. But in the meantime there was one jump by Einstein to convert Planck's viewpoint that only oscillators or atoms were quantized, to the idea that light was really photons and could be considered as a certain way as particles with energy E . Furthermore, Bohr had pointed out that we knew of atoms having energy levels, but they are not necessarily equally spaced like Planck's model says. And so it became necessary to calculate, i.e., first, deduce the accurate law from a more accurately quantum-mechanics viewpoint.

Einstein assumed that Planck's law formula was right, and he used that formula to obtain some new information, previously unknown, about the transactivity of radiation from an atom. This discussion will be taken up later in the chapter on energy levels of an atom, say the nth level, in the Bohr model (Fig. 47-2). Now Einstein proposed that when such an atom has light of the right frequency striking it, it can absorb this photon of light and make a transition from state n to state $n+1$. And the probability that this occurs per second depends upon the two fields, of course, but is proportional to the intensity of light in that it is *linear* with it. Let us call this proportionality constant α , merely because of us that this is indeed a fixed constant of nature, but depends on the particular pair of levels being used. In reality, some levels are bound to exist. Now what is the formula going to be for the rate of emission from $n+1$? Einstein proposed that this must have two parts to it. First, even if there were no light present, there would be some chance that a photon in an excited state would fall to a lower state, emitting a photon; this is called spontaneous emission. An oscillator with a certain amount of energy, even in classical physics, does not keep that energy, but loses it by radiation. Thus the meaning of spontaneous radiation is this. The real situation is that if the atom is in an excited state there is a certain probability β , which depends on the levels again, so it will drop down to a lower level, and this probability is independent of whether light is striking on the atom or not. But then Einstein went further, and by comparison with the classical theory, and by a bit of simple reasoning, concluded that emission was also influenced by the presence of light, and other light of the right frequency is striking on an atom. It has an increase, α , of radiation, a photon loss, proportional to the intensity of the light, with a proportionality constant $\beta\alpha$. Later, if we calculate it, this coefficient turns out, then we will find that Einstein was wrong. Up to now we will find he was right.

This Einstein assumed that there are three kinds of processes: an absorption process from the intensity of light, an emission proportional to the intensity of light, and a pairwise removal or some interconversion involving, and a spontaneous emission independent of light.

Now suppose that we have, at equilibrium at temperature T , a certain number of atoms N_0 in the system and another number N_1 in the state n . Then the total number of atoms that are going up is to be the number that are in the state n . This however assumed that if we take N_0 it gives us to us, we have a formula for the number that are going from n to higher levels:

$$R_{1 \rightarrow n} = N_0 K_B T \beta(\nu). \quad (47-13)$$

The number that we go from n to $n+1$ is proportional to the same number, i.e., the ratio N_1/N_0 , but, in addition, to the fact that per second that each one goes down to n . This gives our expression is

$$N_{n+1} = N_0 (K_B T \beta(\nu) - R_{n+1}(\nu)). \quad (47-14)$$

Now we shall suppose that in thermal equilibrium the number of atoms going up is equal the number coming down. That is the way of heat, in which the number will be sure to stay constant is guaranteed.* So we take these two ratios

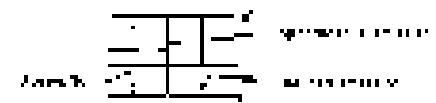


Fig. 47-2. Transitions between two energy levels of an atom.

*. It is not the only way we can change to keep the numbers of atoms in the various levels constant, but it is very naturally works. The steady process, in thermal equilibrium, is balanced by a zero gradient is called the principle of detailed balance.

to be zero, at equilibrium. But we also need the place of information, we know how large δ_{ν} is compared with χ ,—the ratio of those two is $\sim \chi/\delta_{\nu}$. The first Naes Finsen experiment had the only light which was effective in making the transition from n_0 to n_1 is the light which has the frequency corresponding to the energy difference, or $E_{n_1} - E_{n_0}$ in electron-volts. Thus

$$\delta_{\nu} = N_{n_0} \delta_{\nu}^{(1)}, \quad (42.15)$$

Thus if we set the term χ/δ_{ν} equal to $\delta_{\nu}/(\delta_{\nu} - \delta_{\nu}^{(1)})$, or $\delta_{\nu}/(\delta_{\nu}^{(1)})$, and divide by N_{n_0} , we get

$$E_{n_0} (\delta_{\nu})^{1/2} = \delta_{\nu} + \delta_{\nu}^{(1)} \quad (42.16)$$

From this equation, we can calculate δ_{ν} . It is as follows:

$$\delta_{\nu} = \frac{\delta_{\nu}^{(1)}}{N_{n_0} \delta_{\nu}^{(1)/2} + 1}. \quad (42.17)$$

But N_{n_0} has already told us that the formula must be plus 1%. Therefore we can deduce something. First, the N_{n_0} must equal $\delta_{\nu}^{(1)}$, since otherwise we would get the $(\delta_{\nu})^{1/2} = 1$. So Einstein discovered some time ago that he did not know very much about it, namely that the emission probability and the absorption probability must be equal. This is interesting. And afterwards, in order for (42.15) and (42.16) to agree,

$$\delta_{\nu}/\delta_{\nu}^{(1)} \text{ must be } = \delta_{\nu}^2/\delta_{\nu}^{(1)^2}. \quad (42.18)$$

So, if we know, for instance, the absorption rate for a given level, we can deduce the spontaneous emission rate and the stimulated emission rate, or any combination.

This is as far as Einstein or anyone else could go using such arguments. To actually compute the absolute spontaneous emission rate, i.e., the other rates for any specific atomic transition, one needs a knowledge of the mechanics of the atom, called quantum electrodynamics, which was not discovered until eleven years later. This work of Landau was done in 1930.

The possibility of address emission has, today, some interesting applications. If there is big pressure, it is hard to excite the downward transition. The electrons then tend to stick to the available energy, e.g., if there were some atoms stilling in the upper state. Now you can imagine, by some mechanism, caused to have a gas in which the number in the states is very much greater than the number in the states. This is not equilibrium, and so is not given by the formulae $\delta_{\nu}^{(1)}/\delta_{\nu}$, which is for equilibrium. We can even arrange it so that the number in the upper state is very large, while the pressure at the lower state is practically zero. The light which has the frequency corresponding to the energy difference $E_{n_1} - E_{n_0}$ will not be strongly absorbed, because there are not many atoms available. On the other hand, when that light is present, it will induce the emission from this upper state! So, if we took a lot of atoms in the upper state, there would be a sort of chain reaction, in which, the moment the atoms began to emit, more would begin to emit, and the whole lot of them would start again repeating. This is what is called a laser, or, in the case of the忘者, a maser.

Various devices have been obtained in this situation. There may be higher levels to which the atoms can get if we shine in a strong beam of light of high frequency. From these high levels, they may cascade down, emitting various photons, until they get stuck at the state n_0 . If they don't lie in the state of upward emission, the state is called metastable. And then they are all stimulated down together by induced emissions. One more technical point—if we put this system in a ordinary box, it would radiate in so many different directions spontaneously, compared with the induced effect, that we would not be in trouble. But we can enhance the required effect, because its efficiency, by putting nearly perfect mirrors on each side of the box, so that the light which is emitted gets reflected back and forth many, and another chance, to induce the re-emission. Although the mirror is almost one hundred percent reflecting, there is a slight amount of transmission of the mirror, and a little light gets out. In the end, of course, from the conservation of energy, all the light goes out in a single straight direction which makes the strong light beams that are possible today with lasers.



Fig. 42.2. By exciting, say an atom, into a higher state, which may emit a photon having energy $E_{n_1} - E_{n_0}$, the number in this state increases rapidly enough to start laser action.

Diffusion

47-1 Collisions between molecules

We have concentrated so far only the molecular motions in a gas which is in thermal equilibrium. We want now to discuss what happens when things are near, but not exactly in equilibrium. In a situation far from equilibrium things are extremely complicated, but in a situation very close to equilibrium we can easily work out what happens. To see what happens, we must, however, return to the kinetic theory. Since the mechanics and thermodynamics deal with the equilibrium situation, but even from equilibrium we can only analyze what occurs away from equilibrium.

As a simple example of a nonequilibrium circumstance, we can consider the diffusion of ions in a gas. Suppose that in a gas there is a relatively small concentration of free—electrically charged molecules. If we put an ion in it, and let it go, then it will have a force acting which is dependent on the forces on the negative constituents of the gas. If there were no other molecules present, the ion would have a constant acceleration until it reached the wall of the container. But because of the presence of the other molecules, it can see that its velocity increases only until it collides with a molecule and loses its momentum. It starts again to pick up its speed, but then it loses its momentum again. This is not efficient; it does not make a very efficient erratic path, but such is the motion in the direction of the electric force. We shall see that the ion loses on the average half its speed which is proportional to the electric field—the stronger the field, the faster it goes. While the field is on, and while the ion is moving along, it is, of course, not in thermal equilibrium; it is trying to get to eq. equilibrium, which is to be visited at the end of the motion. By means of the kinetic theory we can compute the drift velocity.

I leave out that with our present mathematical abilities we cannot really compute precisely what will happen, but we can obtain approximate results which exhibit all the essential features. We can, without loss of generality, start with one particle, and do it, but it will not be possible to get precisely the exact numerical answers in four or all the terms. We can, therefore, in our derivations, not worry about the precise value of summed factors. They can be obtained only by some much more sophisticated and complex treatment.

Before we consider what happens in nonequilibrium situations, we shall need to look a little closer at what goes on in a gas in thermal equilibrium. We don't need to know, for example, what the average time between successive collisions of a molecule is.

Any molecule experiences a sequence of collisions with other molecules—in a random way, of course. A particular molecule will, in a long period of time T , have a certain number, N , of hits. If we denote the length of time, there will be noncolliding time, so the number of hits is proportional to the time T . We would like to write it this way:

$$N \propto T^{\alpha} \quad (47.1)$$

We have written the constant of proportionality as $1/\tau$, where τ will have the dimensions of a time. The constant is the average time between collisions. Suppose, for example, that in six hours there are one million hits, that is, 6,000,000. We would say that a given molecule is the average time between 1,000,000 hits.

We may often wish to ask the following question: "What is the chance that a molecule will experience a collision during the next small interval of time $d\tau$?" The answer we must intuitively understand is $d\tau$. Let us, if trying to make a more

47-2 Collisions between molecules

47-2 The mean free path

47-3 The drift speed

47-4 Friction coefficient

47-5 Molecular diffusion

47-6 Thermal conductivity

continuing unchanged. Suppose now there were a very large number N of molecules. How many will have collisions in the next interval of time Δt ? If N is so large, nothing is changing on the average over time Δt . At time t we will have the same number of collisions as we have at time $t + \Delta t$. That number we know is $N\delta(t)$. So the number of Δt 's of N molecules is $N\delta(t)/\Delta t$, and the chance, or probability, of a hit for any one molecule is just $1/N$ or $\delta(t)/(N\Delta t)\delta(t) = \delta(t)$, as we guessed above. That is to say, the fraction of the molecules which will suffer a collision in the time Δt is $\delta(t)$. To take an example, if $\delta(t)$ is one billion, then in one second the fraction of particles which will suffer collisions is .0001. What this means, of course, is that 1/1000 of the molecules happen to be close enough to other they are going to hit next. One collision will occur in the next minute.

When we say that τ is the mean time between collisions, in one minute, we do not mean that all the collisions τ occur at times separated by exactly one minute. A particular particle does not have a τ time, wait one minute, and then have another collision. The time between successive collisions is quite variable. We will not need to do our later work over, but we may make a small diversion to answer the question "What are the times between collisions?" We know that for the case above, the average time is one minute, but we might like to know, for example, what is the chance that we get no collision for the population?

We shall find the answer to the general question, "What is the probability that a molecule will go for a time τ without suffering a collision?" At some arbitrary instant—that we call $t = 0$ —we begin to watch a certain molecule. What is the chance that it goes by until t without suffering with another molecule? To compute the probability, we observe what is happening to all the molecules in a container. After we have set out N molecules, how many have suffered up to the time t ? $N\delta(t)$, or fewer, less than $N\delta(t)$. We can think this because we know how it changes with time. If we know that $N\delta(t)$ molecules have got by until t , then $N\delta(t) - dt$ the number which get by until $t + dt$ is less than $N\delta(t)$ by the number that had collisions in dt . The number that collide in dt we have written above in terms of the mean time τ as $\delta(t) - \delta(t+dt)$. We have the equation

$$N(t + dt) = N(t) - N(t) \frac{dt}{\tau}. \quad (41.21)$$

The τ cancels on the left-hand side, $dN = -N/dt$, and we write, according to the definition of derivative, as $dN(t) = -dN(t)/dt dt$. Making this substitution, Eq. (41.21) yields

$$\frac{dN(t)}{N(t)} = -\frac{dt}{\tau}. \quad (41.22)$$

The numbers that are being lost in the interval of time dt is proportional to the number that are present, and inversely proportional to the mean time τ . Equation (41.22) is easily integrated if we rewrite it as

$$\frac{dN(t)}{N(t)} = -\frac{dt}{\tau}. \quad (41.23)$$

Each side is a perfect differential, so the integral is

$$\ln N(t) = -t/\tau + \text{constant}, \quad (41.24)$$

which says the same thing as

$$N(t) = N_0 e^{-t/\tau}. \quad (41.25)$$

We know that the constant must be just N_0 , the total number of molecules present, since all of them start at $t = 0$ to wait for their "first" collision. We can write our result as

$$N(t) = N_0 e^{-t/\tau}. \quad (41.26)$$

If we seek the probability of no collision, (43.6) we can get it by dividing $N(0)$ by $N_{\text{coll}} + 1$:

$$P(0) = e^{-\frac{\tau}{\bar{\tau}}}. \quad (43.8)$$

The result is the probability P that a particular molecule survives at time t without a collision is $e^{-t/\bar{\tau}}$, where $\bar{\tau}$ is the mean time between collisions. The probability starts out at 1 (or certainty) for $t = 0$, and gets less as t gets bigger and bigger. The probability that the molecule avoids a collision for a time equal to τ is $e^{-1} = 0.37 \dots$. This means less than one-half that it will have a greater than average time between collisions. That is all right, because there are enough molecules which go collision-free for times much longer than the mean time between colliding, so that the average time can still be τ .

We empirically define τ as the average time between collisions. The result we have obtained in Eq. (43.7) also says that the mean time from an arbitrary starting point to the next collision is also τ . We can understand this somewhat surprising fact in the following way. The number of molecules which experience their own constant τ at the instant t is the case after an arbitrarily chosen starting time is $N(t)/N_0$. Their "time until the next collision" is, of course, just t . The "average time until the next collision" is obtained the usual way:

$$\text{Avg time until the next collision} = \frac{1}{N(t)/N_0} \int_0^{\infty} t \frac{N(t)}{N_0} dt.$$

Using $N(t)$ obtained in (43.7) and evaluating the integral, we find indeed that τ is the average time from any instant until the next collision.

43-2 The mean free path

Another way of describing the molecular collisions is to talk not about the time between collisions, but about how far the particle moves between collisions. If we say that the average time between collisions is τ , and that the molecules have a mean velocity v , we can expect that the average distance between collisions, which we shall call λ , is just the product of v and τ . This distance between collisions is usually called the mean free path:

$$\text{Mean free path } \lambda = v\tau. \quad (43.9)$$

In this chapter we shall focus on molecular motion, since hardly anyone measures it in any particular case. The various possible averages—the mean, the root-mean-square, etc.—are all usually equal and differ by factors which are near to one. Since a detailed analysis is required to obtain the various numerical factors anyway, we need not worry about which average is required of any particular point. We may also note the reader that the standard symbols we are using for some of the physical quantities (e.g., λ for the mean free path) are not follows generally accepted conventions, mainly because there is no general agreement.

For the time, choose the λ molecule will have a collision in a short time if it is equal to d/v , the chance that it will have a collision in going a distance d is d/v . Following the same line of argument used above, the reader can show that the probability that a molecule will go a distance d before having its next collision is $e^{-d/\lambda}$.

The average distance a molecule goes before colliding with another molecule—the mean free path—will depend on how many molecules there are around, and on the “size” of the molecules, i.e., how big or large they represent. The effective “size” of a target in a collision we usually describe by a “collision cross section,” the same idea that is used in nuclear physics, or in high-energy particle

Consider a moving particle which travels a distance d through a gas which has N molecules (molecules per unit volume; Fig. 43-1). If we look at each unit of area perpendicular to the direction of motion of our selected particle, we find there n molecules. Then, our presents an effective collision area σ , as it is usually called, “collision cross section,” σ , from which a chance induced by the scattering is $n\sigma d/v$.

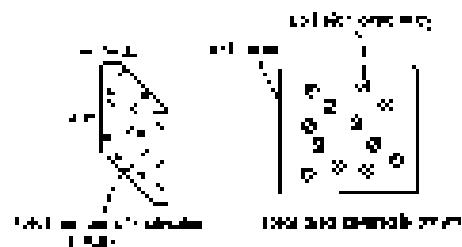


Fig. 43-1. Collision cross section.

By "collision cross section" we mean the area within which the center of our particle must be located if it is to collide with a particular molecule. If molecules were like spheres (a classical picture) we would repeat that as $\pi r_1^2 + \pi r_2^2$, where r_1 and r_2 are the radii of the two colliding objects. The chance that our particle \mathbf{r}_{tot} have a collision is the ratio of the area covered by scattering molecules to the total area, which we let σ_{coll} be to be zero. So, the probability of a collision going a distance d is just σ_{coll}/d .

$$\text{Chance of a collision in } dx = \sigma_{\text{coll}}/dx \quad (43.10)$$

We have seen above that the chance of a collision $= \sigma_{\text{coll}}$ also be written in terms of the mean free path λ as λ^{-1} . Comparing this with (43.10), we can relate the mean free path to the collision cross section:

$$\frac{1}{\lambda} = \sigma_{\text{coll}} \quad (43.11)$$

which is easier to remember than writing σ_{coll}

$$\lambda = \sigma_{\text{coll}}^{-1}. \quad (43.12)$$

The formula can be thought of as saying that the λ should be one collision, or one passage, when the particle goes through a distance λ^2 which leaves letting molecules come just cover the total area. At a cylindrical volume of length λ and radius $\lambda/2$ units, there are $4\pi/3$ surfaces. If each one has a chance, the total chance amounts to $4\pi/3$, which is just one unit of area. The whole area is proportional of course, because some molecules are partly hidden behind others. That is why some molecules go further than λ before they are in equilibrium. It is only for convenience that the mean free path is given by the time they get within area λ^2 . From measurements of the mean free path, we can determine the scattering cross section σ_{coll} , and compare the result with theory based on a detailed theory of atom-molecule. But that is a different subject. So we return to the problem of many-molecule gases.

43-3 The drift speed

We want to describe what happens to a molecule of several molecules, which are different in some way from the large majority of the molecules in a gas. We shall refer to the "numerous" molecules as the "background" molecules, and we shall call the molecule of interest 2, to distinguish the having few molecules ("special" molecule) from 1, the "background". A molecule could be special by any number of ways. It might be heavier than the background molecule. It might be a different element. It might have an electric charge, i.e., be an ion. It may consist of more than molecules. Because of their different masses or charges the forces it may have forces on them which are different from the forces on the background molecules. By considering what happens to these S molecules we can understand the basic effects which come into play in a similar way in many different phenomena. To start, let's use diffusion of gases. Atomic clusters in batteries, colloids, and emulsions, happen to do this.

We begin by concentrating on the basic process of A molecule in a background gas to move on its own specific force F (which might be, e.g., gravitational or due to heat) and to collision by the non-specific force due to collisions with the background molecules. We would like to describe the general behavior of the S molecule. What happens to it, in short, is that it moves around and goes at it will, like a swimmer again with other molecules. But, the point is, we shall see that it does make some net progress in the direction of the force, F. We say that there is a drift speed due to an accumulation. We would like to know what the speed, v_{drift} , is—it's drift velocity due to the force, F.

If we start to observe an S molecule at some instant we may expect that it is somewhere between two collisions. In addition to the velocity it was left with after its last collision, it is picking up some velocity component v_{drift} due to force F. In a

short time for the average, in a time τ , will experience a sufficient number of collisions in a new portion of its trajectory. It will have a new starting velocity, and it gains no acceleration from it.

To keep things simple for the moment, we shall suppose that after each collision our Sand-Dank gets a completely "fresh" start. That is, that it obeys no consequences of its past motion up to t . This might be a reasonable assumption if our Sand-molecule were much bigger than the tiny gas molecule, but it is certainly not valid. (page 2) We shall discuss later the improved assumption.

For the moment, then, our starting fact is that the Sand-molecule leaves each collision with a velocity which may be in any direction or at any speed. The starting velocity will also be equally in all directions and will not contribute to any net motion, so we shall not worry further about its initial velocity after a collision. In addition to its random motion, each Sand-molecule will, however, at any moment, an additional velocity in the direction of the force \mathbf{F} , which it has picked up since its last collision. What is the average *at time t* of the sum of the velocities?

For you, the acceleration F/m , where m is the mass of the molecule, has been the average from many previous collisions. Since the average *at time t* since the last collision must be the same as the average *from time t* to now collision, which we have called τ , equals $F/m\tau$. The average velocity from t to t , of course, is just what is called *the drift velocity*, so we have conclusion

$$\text{d.v.} = \frac{F\tau}{m} \quad (43.17)$$

This last relation is the heart of our subject. There may be some complications in determining what τ is, but the drift velocity is defined by Eq. (43.17).

You will notice that the drift velocity is proportional to the force. There is, unfortunately, no generally accurate or for the constant of proportionality. Different gases have been used for such different kinds of forces. In electrical problems the force is related to the charge times an electric field, $F = qE$. Then the constant of proportionality between the velocity and the electric field is the "mobility." In spite of the possible difficulties of measurement, we shall use the term mobility for the ratio of the drift velocity to the force F in newtons. We write

$$M = \frac{\text{d.v.}}{F} \quad (43.18)$$

In general, and we shall see later why, we have from Eq. 43.17 that

$$\nu = F/M \quad (43.19)$$

The mobility is proportional to the mean time between collisions (there are fewer collisions *as soon as it starts*) and inversely proportional to the mass. From this information we expect M to depend upon the mass.

To get the correct numerical coefficient in Eq. (43.19), which is correct as given, takes some care. Without intending to bore you, we should still point out that the arguments become subtler which can be appreciated only by a careful and detailed study. To illustrate our present difficulties, in spite of agreement, we shall make use again of the argument which led to Eq. 43.17 in a reasonably but unrigorous way (and the way one will find in most textbooks).

We might have said: The mean time between collisions is τ . At a sufficiently particular instant with the random velocity, but it possess a non-additional velocity between collisions, which is equal to the acceleration times the time. Since it takes the time τ to travel at the new velocity, it gets there with a velocity $(F/m)\tau$. At the beginning of the collision, it has zero velocity. Summing the two collisions it has, on the average, a velocity $\frac{1}{2}(F/m)\tau$ of the final velocity on the mean drift velocity $= (F/m)(\tau/2)$ (Eq. 43.17). This result is wrong and the result in Eq. (43.17) is right, although the arguments may sound equally satisfactory. The reason the second result is wrong is somewhat subtle, and we leave with the "fixing." The argument is made as though all collisions were separated by the mean time τ . The fact is that some jumps in velocity and others are longer than the mean. Short jumps occur more often but make less contribution to the drift velocity because they have less

comes "to really get going." If one takes proper account of the distribution of free times between collisions, one can show that there should not be far fewer τ_{coll} than obtained from the second argument. The story was made so trying, in relation to a simple argument, by average free velocity vs. average velocity itself. This relationship is no example, so it is best to concentrate on what is known: the average velocity itself. The first argument we gave determines the average velocity directly. And that's it! That we can perhaps see now why we shall not in general be in all of the correct numerical coefficients in our elementary calculations.

We return now to our simplifying assumption that each collision knocks off all energy of the pair system. Let's check what is made of each collision. Suppose our S -particle is a heavy object at a background of lighter particles. Then our S -particle will not lose its "kinetic" momentum in each collision. It would take many collisions before it became this "unmovable" agent. We should assume, instead, that each collision in some sense loses the energy — in fact a certain fraction of its momentum. We shall not work out the details, but note that the result is equivalent to replacing τ_{coll} by average collision time, by a two-and-fourth τ which corresponds to the average "forgetting time," i.e., the average time to forget its previous momentum. With such an interpretation of τ_{coll} we can use the formula (43.18) for situations which are not quite as simple as we first assumed.

43-4 Tissue conductivity

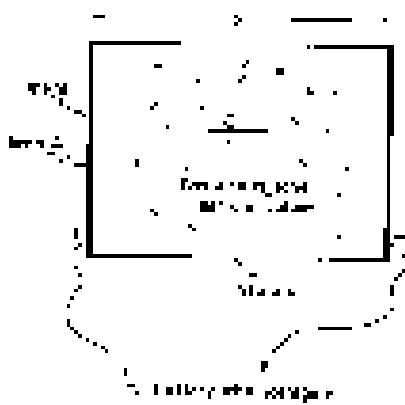


Fig. 43-7 Bedside current in ionized gas.

We now apply our results to a specific case. Suppose we have a gas in a vessel in which there are also some ions — atoms or molecules with a net electric charge. We show the situation schematically in Fig. 43-2. If two opposite walls of the container are metallic plates, we can connect them to the terminals of a battery and thereby produce an electric field in the gas. The electric field will result in motion of the ions, so they will begin to drift toward one or the other of the plates. An electric current will be induced, and the gas with its ions will become like a resistor. By computing the ion flow from the drift velocity we can compute the resistance. We ask, specifically: How does the flow of electric current depend on the voltage difference V that we apply across the two plates?

We consider the case that our container is a rectangular box of length L and cross-sectional area A (Fig. 43-2). If the potential difference in voltage from one plate to the other is V , the electric field E between the plates is V/L . (The electric potential is the work done in carrying a unit charge from one plate to the other. The force on a unit charge is E . If E is the same everywhere between the plates, which is a good enough approximation for now, the work done on a unit charge is just EV , or $V = EB$.) The electric force on an ion of rate q is qE , where q is the charge on the ion. The drift velocity of the ion is then v times this force, or

$$qE = mv^2 \quad \Rightarrow \quad qvE = mv^2 \frac{V}{L} \quad (43.19)$$

An electric current I is the flow of charge in a unit time. The electric current of one of the plates is given by the total charge of the ions which arrive at the plate in a unit of time. If the ions drift toward the plate with the velocity v_{drift} (that they which are within a distance $v_{\text{drift}}T$) are active at the plate in the time T , then, since N ions per unit volume, the number which reach the plate at the time T is $(N_A v_{\text{drift}})^2 v_{\text{drift}}^{-1} T$. Integration over the energy q , we have (43.19)

$$\text{Charge collects in } V \text{ in unit } T. \quad (43.20)$$

The current I is the charge collected in T divided by T , or

$$I = \frac{\text{charge}}{T} = \frac{qV}{L} \quad (43.21)$$

$$I = \mu d^2 n \frac{d}{2} E. \quad (43-19)$$

We find that the current is proportional to the voltage, where d is the length of the tube, and μ is a constant; it is the "value" of the proportionality constant.

$$\frac{I}{E} = \omega^2 n \frac{d}{2}. \quad (43-20)$$

We have a relation between the resistance and the molecular properties n , a , and μ , which depends on m , k , and ν . If we know n and μ from atomic measurements, a measurement of R could be used to determine a , and from a also:

43-5 Molecular diffusion

We turn now to a different kind of problem, one a different size of analysis: the theory of diffusion. Suppose that we have a container, or jar, of thermal expansion α , and that we introduce a small amount of a different kind of gas at some place in the container. We shall call the original gas the "background" gas, and the new one the "special" gas. The special gas will start to spread out through the whole container, but it will spread slowly because of the presence of the background gas. This slow spreading process is called diffusion. The diffusion is controlled mainly by the molecules of the special gas, being knocked about by the molecules of the background gas. After a large number of collisions, the special molecule will spread out more or less evenly throughout the volume of the whole volume. We must be careful not to confuse diffusion of a gas with convection transport, that may occur due to convection currents. Most commonly, the mixing of two gases occurs by a combination of convection and diffusion. We are interested here only in the case that there are no "drift" currents. The gas is spreading only by molecular motion, by diffusion. We wish to see just how fast diffusion takes place.

We now compute the average of molecules of the "special" gas due to the molecular motion. There will be a net flow only when there is some non-uniform distribution of the molecules; otherwise all of the molecular motion would average to give no net flow. Let us consider first the flow in the x -direction. To find the flow, we consider an imaginary plane surface perpendicular to the x -axis and count the number of special molecules that cross this plane. To obtain the net flow, we must count as positive those molecules that cross in the direction of positive x and subtract from this number the number which come in the negative x -direction. As we have seen many times, the number which crosses a surface area in a time Δt is given by the number σ such that the interval Δt is a volume which includes the distance ΔT from the plane. (Note that σ is the actual molecular velocity, not the drift velocity.)

We shall simplify our algebra by giving our surface one unit of area. The mean number of special molecules which pass "forward" right through the x -direction to the right is $\sigma_1 \Delta t$, where σ_1 is the number of special molecules per unit volume to the left within a factor of $\Delta x \times \Delta n$, here we are ignoring such factors). The number which cross "from right to left" is similarly, $\sigma_2 \Delta t$, where σ_2 is the number density of special molecules on the right-hand side of the plane. If we call the molecular current J by which we mean the net flow of molecules per unit area per unit time, we then

$$J = \frac{\sigma_1 + \sigma_2}{2T} \frac{\Delta x \Delta n \Delta t}{\Delta T}. \quad (43-21)$$

 ΔT

$$J = (\sigma_1 - \sigma_2) \tau. \quad (43-22)$$

What shall we use for σ_1 and σ_2 ? When we say "the density of the 'old'" molecules in the left side of the jar? We shall define the density at the place from which the molecules started their "flight," because the number which does such

mass is determined by the number present in the plume. So by the definition the density difference is the "equivalent concentration" with δn being the density at the distance r to the right of our cylinder center.

It is convenient to consider first the distribution of our special molecules in space, described by a continuous function $\rho(x, y, z)$ which we shall call ρ_0 . By $\rho_0(x, y, z)$ we mean the number density of special molecules in a small volume element centered at (x, y, z) . In terms of ρ_0 we can express the diffusion equation (41.18)

$$\partial \rho_0 / \partial r = D \cdot \frac{\partial^2 \rho_0}{\partial r^2} - \frac{\partial \rho_0}{\partial r} / \delta n. \quad (41.23)$$

Substituting this result in Eq. (41.20) and multiplying by a factor of 2, we get

$$J_x = -D \frac{\partial \rho_0}{\partial r}. \quad (41.24)$$

We have found that the flow of special molecules is proportional to the derivative of the density, or to what is sometimes called the "gradient" of the density.

This does not yet tell us how much current through approximations. Besides various factors of two we have left out, we have used a value we should have used 2, and we have assumed that ρ_0 and n refer to a plume in the perpendicular distance r from our surface; whereas for those molecules which do not travel perpendicular to the surface $\rho_0(r)$ is not equal to the density at distance r from the surface. All of these influences can be made; the results of a more careful analysis show that the right-hand side of Eq. (41.24) should be multiplied by 1/2. So a better answer is

$$J_x = -\frac{D}{2} \frac{\partial \rho_0}{\partial r}. \quad (41.25)$$

Similar equations can be written for the currents in the y and z directions.

The current J_x and resistance problem can be measured by macroscopic observations. The experimentally determined ratio is called the "diffusion coefficient" D . That is,

$$J_x = -D \frac{\partial \rho_0}{\partial r}. \quad (41.26)$$

We have been able to show that for a gas we expect

$$D = \eta k. \quad (41.27)$$

So far in this chapter we have considered two distinct processes: scattering of molecules due to "outside" forces; and diffusion by spreading determined only by the internal forces, the random collisions. There is, however, a relation between them, since they both depend basically on the incident molecules, and are zero free both of space in 1-D calculations.

With Eq. (41.27) we substitute $\rho_0 = \rho_0(r)$ and $r = \rho_0$ in the case

$$J_x = -D \rho_0^2 \frac{\partial \rho_0}{\partial r}. \quad (41.28)$$

But ρ_0 depends only on the temperature. We recall Eq.

$$k_B T^2 = 3kT. \quad (41.29)$$

so:

$$J_x = \omega D \frac{\partial \rho_0}{\partial r}. \quad (41.30)$$

We find that D , the diffusion coefficient is just ω times ρ_0 , the mobility coefficient:

$$\omega = \eta k. \quad (41.31)$$

And it turns out that the numerical coefficient in (41.31) is exactly right—no extra factors have to be chosen in order to fit our rough assumptions. We can show,

in fact, that (43.11) must always be correct, even in non-dilute situations (for example, the case of a suspension in a liquid, where the details of our simple calculation would not apply at all).

To prove that (43.11) must be correct, let's do it in a different way, using only one basic principle of statistical mechanics. Imagine a situation in which there is a gradient of "special" molecules, and we have a diffusion current proportional to the density gradient, according to Eq. (43.20). We now apply a force F in the x direction, so that the chemical potential field is $\phi = F$. According to the definition of the mobility μ , this will be a drift velocity given by

$$v_{\text{drift}} = \mu F. \quad (43.12)$$

By our usual arguments, the average number density of molecules which pass a unit of area in a unit of time is

$$J_{\text{drift}} = \mu N_A v_{\text{drift}}, \quad (43.13)$$

or

$$J_{\text{drift}} = \mu N_A F. \quad (43.14)$$

We now consider the force F so that the drift current due to F is zero. In other words, the drift current is zero for all non-special molecules. We have $J_{\text{drift}} = J_{\text{gas}} = 0$, or

$$\mu \frac{\partial \phi}{\partial x} = \nu_{\text{gas}} F. \quad (43.15)$$

This is the "modified" form of Fick's law (with the addition of density ρ in Eq. 1.29)

$$\frac{\partial \phi}{\partial x} = \frac{\nu_{\text{gas}} F}{\mu}. \quad (43.16)$$

But notice: We are describing an equilibrium condition, so our equilibrium laws of dilute ideal molecules apply. According to these laws, the probability of having n molecules at the coordinate x is proportional to e^{-nE} , where E is the potential energy. In terms of the number density n , this means that

$$n_x = n_0 e^{-\beta E(x)}. \quad (43.17)$$

If we differentiate (43.17) with respect to x , we find

$$\frac{\partial n_x}{\partial x} = -n_0 e^{-\beta E(x)} \cdot \frac{1}{kT} \frac{\partial E}{\partial x}, \quad (43.18)$$

or

$$\frac{\partial \phi}{\partial x} = -\frac{n_0}{kT} \frac{\partial E}{\partial x}. \quad (43.19)$$

In our situation, since the force F is in the x direction, the potential energy E is just $-Fx$, and $\partial E/\partial x = F$. Equation (43.19) then gives

$$\frac{\partial \phi}{\partial x} = \frac{n_0 F}{kT}. \quad (43.20)$$

(This is just exactly Eq. (10.2), from which we deduced ν_{gas} in the first place, since here n_0 is constant). Combining (43.16) with (43.19) we get exactly Eq. (43.11). We have shown that Eq. (43.11) which gives the diffusion current in terms of the mobility μ , has the correct coefficient and is very generally true. (Usually one diffusion experimentally measures μ . This relation was first deduced by Flory.)

43.4 Thermal conductivity

The methods of the kinetic theory that we have been using above can be used also to compute the thermal conductivity of a gas. If the gas at the top of a container is colder than the gas in the bottom, heat will flow from the top to the bottom. (We think of the top being "newer" because otherwise conduction would not be

set up and the problem would not keep the end of heat conduction). The transfer of heat from the hotter gas to the colder gas by the diffusion of the "heat" molecules with lower energy decreased and the addition of the "heat" molecules upward. To compute the flow of thermal energy we can ask about the energy at risk downward across an element of space by the mechanism being analyzed, and convert the energy at risk upward across the surface by the downward moving molecules. The difference will give us the net downward flow of energy.

The thermal conductivity κ is defined as the ratio of the rate of which thermal energy is carried versus the instantaneous temperature gradient.

$$\frac{1}{A} \frac{dQ}{dt} = -\kappa \frac{\Delta T}{L}. \quad (43.41)$$

Since the details of the calculations are quite similar to those we have done above at considering the flow of mass in a laminar flow, we shall leave it to the exercises for the reader to show that

$$v = \frac{k_e \sigma}{\tau}, \quad (43.42)$$

where $(\tau = 1.97)$ is the average energy of a molecule at the temperature T .

If we use the relation $(\kappa = 1/v)$, the heat conductivity can be written as

$$\kappa = \frac{1}{v} = \frac{k_e}{1.97 \sigma}. \quad (43.43)$$

We have another surprising result. We know that the average velocity of gas molecules depends on the temperature but not on the density. We know, however, that depend only on the size of the molecules. So our simple result says that the mean free path and therefore the rate of flow of heat in any particular circumstance is independent of the density of the gas! The change in the number of "carries" of energy with a change in density is just compensated by the larger distance the "carries" can go between collisions.

One may ask, "Is the heat flow independent of the gas density in the limit of low density perhaps?" When this is not true at all! One might note the formula (43.43) was derived, as were all the others in this chapter, under the assumption that the mean free path between collisions is much greater than any of the dimensions of the container. Whenever the probability exists for that a molecule has a short distance of crossing from one well of its container to the other without having a collision, none of the results given in this chapter apply. We must at such cases go back to kinetic theory and calculate again the details of what will occur.

The Laws of Thermodynamics

44-1 Heat Engines; the first law

So far we have been discussing the properties of matter from the *LJ* point of view, trying to understand why what happens if we squeeze hot things or stretch them, changing certain properties. However, there was no hope of calculating steps among the properties of substances which can be worked out without consideration of the detailed structure of the materials. The determination of the relationships among the various properties of materials, without knowing their internal structure, is the subject of thermodynamics. Historically, thermodynamics was developed before the understanding of the internal structure of matter was achieved.

As part of this, let us trace from the kinetic theory that the pressure of a gas is caused by molecular bombardment, and we know that if we heat a gas, its bombardment increases, the pressure must increase. Conversely, if the piston in a container of the gas is moved inward against the force of bombardment, the energy of the molecules bombarding the piston will increase, and consequently the temperature will increase. So, on the one hand, if we increase the temperature at a given volume we increase the pressure. On the other hand, if we compress the gas, we will find that the temperature will rise. From the kinetic theory, one can derive a quantitative relationship between these two effects, but one is likely to be right-handed that they are related in some necessary fashion which is independent of the details of the calculation.

Let us consider another example. Many people are familiar with this interesting property of rubber. If you take a rubber band and pull it a *little* more, it goes from its original length, for example, and pulls it out, he can feel a definite warning, and this warning is negligible in the sense that, if he releases the rubber band quickly while it is between his lips, it is definitely visible that occurs (i.e., when we stretch a rubber band it becomes longer when we release the tension of the band it coils). Now our instincts might suggest that if we拉伸 a band, or might pull, then the fact that pulling a band back in might imply that stretching a band should cause it to contract. And, in fact, if we apply a force to a rubber band holding its weight, we will see that the band contracts abruptly (Fig. 44-1). So it is true that when we have a rubber band it pulls, and this fact is definitely related to the fact that when we release the tension of it, it coils.

The internal machinery of rubber that causes these rather surprising complications. We will describe it from a molecular point of view to some extent although our main purpose in this chapter is to understand the relationships of these effects independently of the molecular model. In particular, we can see from the molecular model that the effects are directly related, this way to understand the behavior of rubber, it is necessary that this substance consists of an enormous tangled web of long chains of molecules, a kind of "molecular spaghetti," with one end, exemplified between the electric lamp are cross-links. The spaghetti that is sometimes welded together where it crosses and in place of spaghetti, a good example. When we pull on such a tangled network of the chains tend to pull on each other, the direction of the pull. At the same time, the chains are in their natural form, so they try and pull one firmly. It follows that such a net, if stretched, would not be like a normal stretched, because it would be held down the sides by the other chains and other electrons, and would tend to kink up again. So the real reason why a rubber band tends to contract is this: when one pulls it out, the chains are lengthened, and the thermal agitation of the molecules on the sides of the chains tend

44-2 Heat engines: the first law

44-3 The second law

44-4 Reversible engines

44-5 The thermodynamic temperature

44-6 Entropy

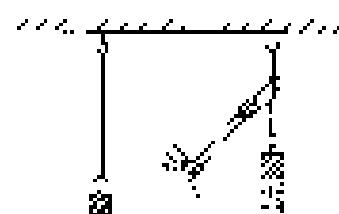


Fig. 44-1. The heated rubber band.

buckle the chains up and make them shorter. One can then appreciate that if the chains are held tautened and the temperature is increased, with the effect of the bombardment on the sides of the chains, there is an increase in the chains tends to pull in, and it is possible to put a stronger weight when loaded. In other terms, material for a time, a rubber band is allowed to relax and chain becomes soft, and the molecules walking in. You are by as they pass and from the relaxing chain. See the very good figure 1.

We have seen how these two effects, contraction when heated and stretching during relaxation, can be related by the binomial theory, but it would be a tremendous challenge to determine from the theory the precise relationship between the two. We would have to know how many different thermal expansion coefficients there are, what the chains look like, and we would have to be account of all kinds of other complications. The detailed mechanism is no example, but we can find by kinetic theory, pretty certainly exactly what happens at a definite interface between the two effects which occur to an extent without knowing anything about the internal machinery!

The whole subject of thermodynamics has already been fully up to the covering a kind of science different from another one is "stranger" to angle. Temperature is often to be at lower temperatures, it is not to be possible to lift weights, and to move them around, and thus you burn with heat. In fact, we have already seen experimentally that a twisted rubber band can lift a weight. The study of the way that one does work with heat is the beginning of the science of thermodynamics. Can we make an engine which uses the heating effect on a rubber band to work? One can cause a silly looking engine to do this, as the "Foucault's" bicycle wheel. You can make of the wheel with a lot of heat bands, then because "a longer" than the rubber bands on the other side. The center of gravity of the wheel will be pulled away from the bearing so that the wheel turns. At the same, a hot rubber band will expand the heat, and the bands move away from the heat and cool, so that the effect will be damped as long as the heat is applied. The efficiency of this engine is very nearly low. Four hundred watts of power pour into the thermometer, but it is just possible to lift 4 kg. With such an engine! An interesting question, however, is whether we can get heat to do the job more in more efficient ways.

In fact, the science of thermodynamics began with an analysis, by the great engineer Sadi Carnot, of the problem of how to built the best and most efficient engine, and this was just one of the few famous cases to which engineering has contributed significantly to physical theory. Another example that comes to mind is the more recent analysis of information theory by Claude Shannon. These two are you, interestingly, two can be closely related.

Now the way a steam engine, originally invented, is built is to pump some water, and the steam is to the engine and passes on a piston which makes a heat pressure, so the steam pushes the piston. And, the "Carnot heat" which he just a simple way to compute the cycle would be to let the steam escape into the air. If you do not stop supplying water, it is cheaper, more efficient to let the steam pass directly out, where it is combined by air, water, and then pass the water back to the boiler, or heat exchanger continuously. Heat is thus supplied to the engine and converted into work. Now would it be better to use alcohol? What property of fuel a substance have on than it makes the best possible engine? That was the question to which Carnot addressed himself, and one of his by-products was the discovery of the law of adiabaticity that we have just examined above.

The results of Carnot's engine are still considered important in certain applications, although statements called the law of thermodynamics. At the time when Carnot wrote, the first law of thermodynamics, the conservation of energy, was not known. Carnot's arguments were so carefully drawn, however, that they are said even though the first law was not known in his time. Some time later, Clausius' theory came to similar conclusions, which were understood more easily than Carnot's very elaborate reasoning. But it turned out that Clausius' work, and the famous 44-2

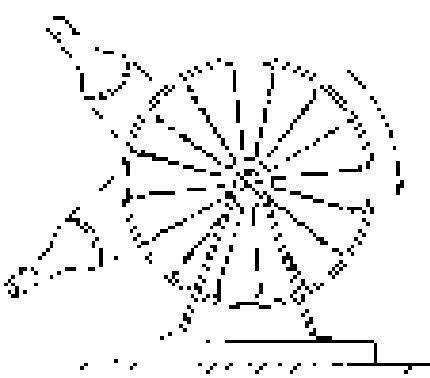


Fig. 44-2. The rubber-band bicycle.

on of energy in general, but that was not done and according to the colour theory, which was later shown to be false. So it has often been said that Carnot's logic was wrong, but his logic was quite correct. Only Carnot's simplified system, with everything read, was incorrect.

The so-called second law of thermodynamics was first discovered by Carnot before any first law. It would be interesting to give Carnot's argument, that did not use the first law, but we shall not do so because we want to be in physics, not history. What I mean by first law from the start, in spite of the fact that a great deal can be done without it,

Let us begin by stating the first law. If we increase the energy of one has a system and put heat into it, and they work out, then the energy is increased by the heat plus the work done. We can write this as follows. The heat Q put into a system plus the work W done on the system is the increase in the energy E of the system; the latter energy is sometimes called the internal energy.

$$\text{Change in } E = Q + W. \quad (44.1)$$

This change in E can be represented as taking a little heat dQ and doing a little work dW .

$$dE = dQ + dW, \quad (44.2)$$

which is a different form of the same law. We know that very well. Now to continue.

44.2 The second law

Now what about the second law of thermodynamics? We know that if we do work and the energy, the work added to us is equal to the heat added. If we do work at a constant temperature T , and we do the work slowly enough, the temperature does not change much, and the converted work and the given amount of heat will be equal. Is it possible to convert the heat back into work, at a given temperature? Is it possible to convert the heat back into work, at a given temperature? The second law of thermodynamics says that it is not. It would be very convenient to be able to convert heat into work merely by reversing a process like this. If we consider only the conservation of energy, we might think that just energy comes out in the vibrations of a molecule, might provide a good example of this. But Carnot assumed that it is impossible to extract the energy of heat at a single temperature. In other words, if the whole world were at the same temperature, one could not convert all this heat entirely into work, while the process of making work go into heat can take place at a given temperature, one cannot reverse it to get the work back again. Specifically, Carnot assumed that heat cannot be taken from a certain temperature and converted into work with no other change in the system or the surroundings.

The last point is very important. Suppose we have a lot of compressed air at certain temperature, and we let the air expand. If we do work, it can make increases in, for example, "Laws of" like in the expansion, but if we try to get the like expansion, at a given temperature, it does not do. We could warm it up again. So the heat added to the air, and we have done work with the compressed air. But Carnot was very wrong, because he did not take into this as if that, if we recompress the air, that we let expand, we will find we can do some work, and when we are finished we will discover that we can only get no work out of the system at temperature T , but we actually put some in. We must necessarily occur vibrations at which the result of the whole process is established, and we can't get more work just as the net result of the process of doing work against friction is to take work and convert it into heat. If we move a car, we will bring the system back precisely to its starting point, and the result the net work against friction and produced heat. Can we reverse the process? It is impossible, so that everything goes backwards, on the return, there won't be any, and ends the cycle. According to Carnot, no. So let us suppose that this is impossible.

If it were possible it would mean, among other things, that we could take heat out of a cold body and you "freeze" it. Only at present is it possible. Now we know it is natural that a hot thing can warm up a cold thing if we simply put a hot body and a cold one together. A discharge heating coil, our experience assures us, it is not going to happen next the hot one gets hotter, and the cold one gets colder. But if we could obtain work by extracting the heat out of the ocean, say, or from anything else at a single temperature, then that work could be converted back into heat by friction at some other temperature. For instance, the other arm of a working machine could be rubbing something that is already hot. The net result would be no work done. Thus a "heat" body, the ocean, could be put . . . into a hot body. Now, the hypothesis of Carnot, now known as of the second law, is sometimes stated as follows: heat cannot be used. Now Comte added to a hot object. But, as we have just seen, these two statements are quite different, but one cannot do what a process which only results to convert heat to work, at a single temperature, and secondly, heat can convert heat back by itself from a cold to a hot place. We shall merely use the first form.

Carnot's analysis of heat engines is quite similar to the argument that we gave above regarding engines in our discussion of the conservation of energy in Chapter 11. In fact, that argument was patterned after Carnot's argument about heat engines, and so the present lecture will follow very much the same.

Suppose we build a heat engine that takes "heat" Q_1 at a temperature T_1 . A certain heat Q_2 is taken from the boiler, the steam engine does some work W , and it then delivers some heat Q_3 into a "condenser" at another temperature T_2 (Fig. 44-3). Carnot did not say how much heat Q_2 was lost; he did not know the first law, and he did not use the first law. Heat Q_3 was equal to Q_2 , because he did not believe in it. Although everybody thought it so, according to the caloric theory, the heats Q_1 and Q_3 would have to be the same. Carnot did not say they were the same—but that is part of the weakness of his argument. If we do use the first law, we find that the net thermal Q_1 is the heat Q_3 , but we put in more heat if that was done:

$$Q_3 = Q_1 - W. \quad (44.5)$$

(If we have some kind of cyclic process where water is pumped back into the boiler after it is condensed, we will say that we have used Q_1 absorbed and sent W down. During each cycle, the a certain amount of water can goes around the cycle.)

Now we shall build another engine, and see if we cannot get more work from the same amount of heat being added at the temperature T_1 , with the condenser still at the temperature T_2 . We start out the same amount of heat, Q_1 , from the boiler, and we shall try to get more work than we did out of the steam engine—perhaps by using some new fluid, such as alcohol.

44-3 The reversible engine

Now we must analyze our engine. One thing is clear: we will not get something of the engine's power because in which there is friction. The best engine will be a frictionless engine. We assume, then, the same hypothesis that we did when we studied the conservation of energy. Let it a perfectly frictionless engine.

We must also consider the analog of free molecular motion, "frictionless" heat transfer. If we push hot object at a high temperature against a cold object, we can be heat flows, then it is not possible to make frictionless flow in opposite direction by a very small change in the temperature of either object. So when we have a heat engine operating, it can move it with a little force one way, "push hot way," and if we push it with a little force the other way, it goes the other way. We need to find the analog of molecular motion: heat transfer when direction we can move it only a little. If the difference in temperature is small, that is impossible, but if you make going the heat flows always the same two things, i.e., essentially, the same temperatures, will just not introduce any difference to make it flow in the desired direction. You, however, will be forced to heat source (Fig. 44-4). If we had

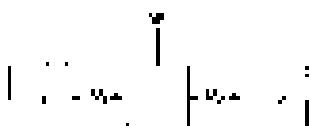


Fig. 44-3. Heat engine.

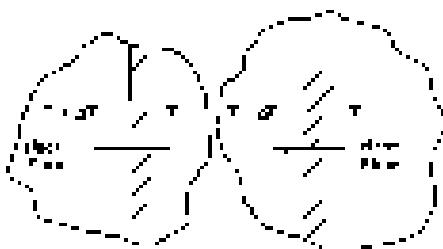


Fig. 44-4. Reversible heat engine.

The objection can be left a little, heat will flow to the gas; if we cool it a little, heat will flow out. But we find that the ideal engine is a *reversible* reversible machine, in which every process is reversible in the sense that, by minor changes, infinitesimal changes we can make the engine go in the opposite direction. This means that *nowhere* in the machine must there be any appreciable friction, and nowhere in the machine must there be any place where the heat of compression, or the flame of the boiler, is in direct contact with something definitely cooler or warmer.

Let us now consider an idealized engine in which all the processes are reversible. To show that such a thing is possible in principle, we will give an example of an engine cycle which may or may not be practical, but which is at least conceivable, in the sense of *Conceivable*. Suppose that we have a gas in a cylinder equipped with a free piston & piston. The gas is not necessarily a perfect gas. The fluid does not even have to be a gas, but to be specific let us say we have a perfect gas. Also, suppose that we have two heat pads, T_1 and T_2 , which may be things like the infinite temperature ∞ , T_1 and T_2 . We will suppose in this case that T_2 is higher than T_1 . Let's first cool the gas and then compress it, which is in contact with the heat pad at T_1 . As we do this, pushing the piston out very slowly so the heat flows from the gas, we will make sure that the temperature of the gas will be less than T_1 , and that the process will not be quite reversible, but if we pull it in slowly enough, the temperature of the gas will never depart much from T_1 . On the other hand, if we push the piston back slowly, the temperature would be more infinitesimally higher than T_1 , and the heat would flow back. We see that such an adiabatic compression is *approximately* reversible, that slowly and gently enough is a reversible process.

To make it more clear, we are doing what is called a *decompression*, or the pressure of the gas against its volume. As the gas expands, the pressure falls. The curve marked (1) tells us how the pressure and volume change if the temperature is kept fixed at the value T_1 . For an ideal gas this curve would be $PV = \text{constant}$. During the isothermal expansion the pressure falls as the volume increases until we stop at the point A. At the same time, a certain heat Q_1 enters the gas from the reservoir T_1 if the gas were insulated without a piston, it would, with the same work, it would cool off, so we *add* heat. Having continued the isothermal expansion, stopping at the point B, let us take the cylinder away from the reservoir and continue the expansion. This time we permit no heat to enter the cylinder. Again we perform the expansion slowly, so there is no reason why we cannot remove it and again it remains there, no friction. The gas continues to expand and the temperature falls since there is no heat gain/heat entering the cylinder.

We let the gas expand, following the curve marked (2), until the temperature falls to T_2 at the point C, but in this kind of expansion, made without adding heat, is called *adiabatic expansion*. Even on the gas, we find easily know that curve (2) lies below the curve $P^{\gamma} = \text{constant}$, where γ is a constant greater than 1, so that the adiabatic curve has a more negative slope than the isothermal curve. The gas cylinder has now reached the temperature T_2 , so that if we put it on the heat pad at temperature T_2 there will be no reversible change. Now we slowly compress the gas while it is in contact with the reservoir at T_2 , following the curve marked (3); (Fig. 44-2, Sec. 3). Because the cylinder is in contact with the reservoir, the temperature is constant, i.e., heat lost is flow from the cylinder into the reservoir at the temperature T_2 . Having compressed the gas adiabatically along curve (1) to the point D, we remove the cylinder from the heat pad at temperature T_2 and compress it still further without adding any heat, flow now. The temperature will rise and the pressure will follow the curve reversed (2). If we carry out each step properly, we can return to the point A at temperature T_1 where we started and repeat the cycle.

So we have, in Fig. 44-2, shown we have carried the gas through a complete cycle, and during one cycle we took up Q_1 at temperature T_1 , and have released Q_2 at temperature T_2 . Now the point is that this cycle is reversible in that we could represent it in steps the other way around. We could have gone backwards instead of forwards; we could have started at point A at temperature T_1 , expanded

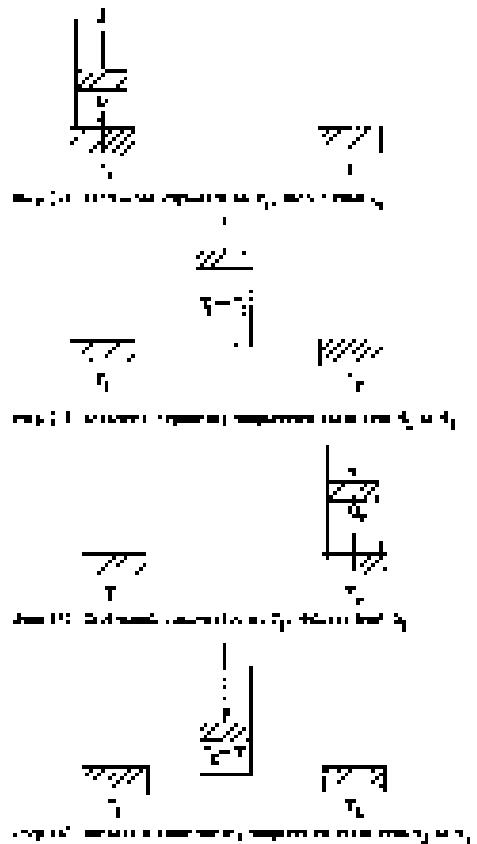


Fig. 44-3. Step in Carnot cycle.

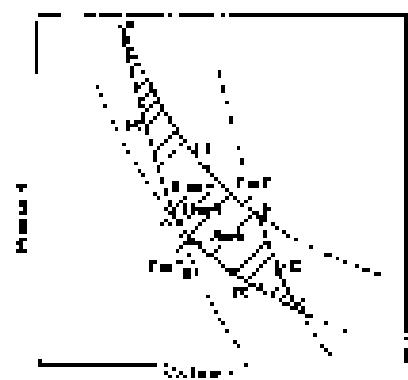


Fig. 44-4. The Carnot cycle.

along the curve. (4), expanded directly to temperature T_2 , absorbing heat Q_2 and $+W'$, along curve 4-5-6-7 along the cycle back to 1. If we run the cycle in one direction, we make the work on the gas; if in the other direction, the gas does work $+W'$.

Technically, it is easy to find out what the total amount of work is, because the work during expansion is the pressure times the change in volume, if δV . On this particular diagram, we know that P vertically and V horizontally, so if we call the vertical distance x and the horizontal distance y , that is $\int y dx$ in other words, the area under the curve. So the area under each of the numbered curves is a measure of the work done by or on the gas in the corresponding step. It is easy to see that the net work done is the signed area of the figure.

Now that we have given a single example of a reversible machine, we shall suppose that other sorts of machines also possible. That is, we assume that we have a reversible engine which takes (Q_1) at T_1 , does work W' , and delivers some heat to T_2 . Now let us suppose we have any compressor B made by means of a flywheel which is designed to absorb an amount of heat $Q_1 + W'$, and rejects the heat at the lower temperature T_2 (Fig. 44-7). Assume the engine B does no work. Then we shall show that B is acting as a heat reservoir. To see this, we can do more work than a reversible one. What? Suppose that, indeed, B works like this B' . Then we could take the heat Q_1 out of the reservoir at T_2 , and work engine B would do work W' and deliver some heat to the reservoir B . To see we do more work than a reversible one. That is, suppose that B' would do work W'' and deliver some heat to the reservoir B . Then, we do more work than B' ; we could use a part of it, W' , to drive the remainder, $W'' - W'$, for useful work. With this work W' we could run engine B backwards because it is a reversible engine. It will absorb some heat from the reservoir at T_2 and deliver Q_1 back to the reservoir at T_2 . After this double cycle, the net result would be that we would have put the system back the way it was before, and we would have done some excess work, namely $W'' - W'$, and all we would have done would be to extract energy from the reservoir at T_2 . We were careful to assume the heat Q_1 is the reservoir at T_2 . So that reservoir can be small and "instantaneous" and immobile machine A + B , whose net effect is therefore to remove a net heat $W'' - W'$ from the reservoir at T_2 , and convert it into work. But to obtain useful work from a reservoir at a single temperature with no other changes is impossible according to Carnot's principle; it cannot be done. Therefore no engine which absorbs a given amount of heat from a higher temperature T_2 and delivers it at the same temperature T_2 can do more work than a reversible engine operating under the same temperature conditions.

Now suppose that engine B is not reversible. Then, of course, not only must W' be not greater than W'' , but now we can reverse the argument and show that W' cannot be greater than W'' . So, if these engines are reversible they must both do the same amount of work, and with this come to Carnot's brilliant conclusion that A + B whose net effect is therefore to remove the net heat $W'' - W'$ from the reservoir at T_2 , and convert it into work. But to obtain useful work from a reservoir at a single temperature with no other changes is impossible according to Carnot's principle; it cannot be done. Therefore no engine which absorbs a given amount of heat from a higher temperature T_2 and delivers it at the same temperature T_2 can do more work than a reversible engine operating under the same temperature conditions.

If we could find out what the law is that determines how much work we obtain when we absorb the heat Q_1 at T_1 and we give back T_2 , this quantity would be a universal thing, independent of the substance. Of course if we knew the properties of the Coulomb interaction, we could work it out and thereby determine the substance must give the same amount of work in a reversible engine. That is the key idea, the clue by which it is possible to find the relationship between how much work is obtained when the field contracts when we heat it, and how much it costs when we let it contract. Imagine that we can turn rubber band into a reversible machine, and that we make it go around a reversible cycle. The net result, the total amount of work done, is then a universal function, that is to say, function which is independent of substance. So we see that a substance's properties must be limited in a certain way; one

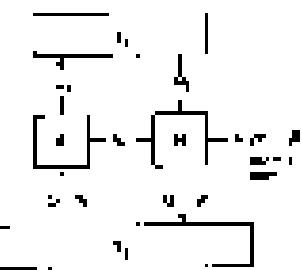


FIG. 44-7. Reversible engine A being driven backwards by engine B .

cannot make up anything he wants, or he would be able to live a life which he could use to produce more than the maximum amount of useful work when he carried it around a reversible cycle. This principle, this law, is the only law of nature that concerns us of the thermodynamics.

44.4 The efficiency of an ideal engine

Now we shall try to find the law which determines the heat \dot{Q}_1 as a function of T_1 , T_2 , and P_2 . It is clear that \dot{Q}_1 is proportional to \dot{Q}_2 , for if we consider two reversible engines in parallel, both working together and both doing the same, the combination is also a reversible engine. If each one does heat \dot{Q}_2 , the two together do $2\dot{Q}_2$, and the work done is also doubled, and so on. So it is not at all surprising that \dot{Q}_1 is proportional to \dot{Q}_2 .

Now the next important step is to find this relation law. We can, and will, do so by studying a reversible engine with the one particular substance whose laws we know, a perfect gas. It is also possible to obtain the rule by a purely logical argument, using our particular substance at ∞ . There is one of the very beautiful aspects of learning in physics and we are reluctant not to show it to you, but for those who would like to see it we said "there it is just a minute". The first we shall see the much less abstract and simpler method of direct calculation for a perfect gas.

We need only obtain formulas for \dot{Q}_1 and \dot{Q}_2 , for \dot{W} is just $\dot{Q}_2 - \dot{Q}_1$, the heat exchanged with the reservoir during the isothermal expansion or contraction. For example, how much heat \dot{Q}_1 is absorbed from the reservoir at temperature T_1 during the isothermal expansion (isentropic) from the initial point (P_1, V_1) to some point (P_2, V_2) , temperature T_2 , to point (P_2, V_2) pressure P_2 , volume V_2 , and the same temperature T_2 ? For a perfect gas each molecule has an energy that depends only on its temperature, and since the temperatures and the number of molecules are the same at (P_1, V_1) and at (P_2, V_2) , the total energy is the same. There is no change in total energy due to the work done by the gas.

$$\dot{W} = \int_{V_1}^{V_2} p dV,$$

during compression is energy \dot{W} , taken from the reservoir. During the expansion, $\dot{W} = \dot{m}RT_1$, or

$$p = \frac{\dot{m}RT_1}{V^2}$$

or

$$\dot{Q}_1 = \int_{V_1}^{V_2} p dV = \int_{V_1}^{V_2} \dot{m}RT_1 \frac{dV}{V} \quad (44.1)$$

and

$$\dot{Q}_1 = \dot{m}RT_1 \ln \frac{V_2}{V_1}$$

is the heat taken from the reservoir at T_1 . In the same way, for the compression of \dot{V}_2 (curve 12) of Fig. 44-6) the heat delivered to the reservoir at T_2 is

$$\dot{Q}_2 = \dot{m}RT_2 \ln \frac{V_2}{V_1} \quad (44.2)$$

To finish our analysis we must only find a relation between P_1/V_1 and P_2/V_2 . This we do by noting that a free, adiabatic expansion from \dot{V}_1 to \dot{V}_2 with pV^γ is a constant. Since $dV = \dot{m}RT$, we can write this as $(pV^\gamma)^{1/\gamma} = \text{constant}$, in terms of T and V , as $TV^{\gamma-1} = \text{constant}$,

$$T_1 V_1^{\gamma-1} = T_2 V_2^{\gamma-1} \quad (44.3)$$

Likewise, since $(pV)^\gamma$, the expansion from \dot{V}_2 to \dot{V}_1 , is also adiabatic, we find

$$T_1 V_1^{\gamma-1} = T_2 V_2^{\gamma-1} \quad (44.4)$$

If we divide this equation by the previous one, we find that \dot{W}_1/\dot{W}_2 , and \dot{V}_1/\dot{V}_2 , in the ratios (44.4) and (44.5) are equal, and that

$$\frac{\dot{Q}_1}{\dot{V}_1} = \frac{\dot{Q}_2}{\dot{V}_2}. \quad (44.7)$$

This is the relation we were seeking. Although proved for a perfect gas engine we know it must be true for any reversible engine at all.

Now we shall see how this reversible law could also be obtained by logical argument, without knowing the complete theory of any specific substance, as follows. Suppose that we have three engines and their temperatures, let us say T_1 , T_2 and T_3 . Let one engine absorb heat \dot{Q}_1 from the temperature T_1 and do a certain amount of work \dot{W}_{12} , and let it deliver heat \dot{Q}_2 at the temperature T_2 (Fig. 44-8). Let another engine run backwards between T_2 and T_3 . Suppose that we let the second engine be of such a size that it will absorb the same heat \dot{Q}_2 , and deliver the heat \dot{Q}_3 . We will have input exergy amount of work, \dot{W}_{23} , and its negative, because the engine is running backwards. When the first machine goes through a cycle it absorbs heat \dot{Q}_1 and delivers \dot{Q}_2 at the temperature T_2 ; then the second machine takes the same heat \dot{Q}_2 out of the reservoir at the temperature T_2 and delivers it into the reservoir at temperature T_3 . Therefore the net result of the two machines is to demand \dot{W}_{12} to take the heat \dot{Q}_1 from T_1 and deliver \dot{Q}_3 at T_3 . These two machines are thus equivalent to a third one, which absorbs \dot{Q}_1 at T_1 , does work \dot{W}_{12} , and delivers heat \dot{Q}_3 at T_3 , because $\dot{W}_{12} = \dot{W}_{12} - \dot{W}_{23}$. Let us now easily show from the first law, as follows:

$$\dot{W}_{12} - \dot{W}_{23} = (\dot{Q}_1 - \dot{Q}_2) - (\dot{Q}_2 - \dot{Q}_3) = \dot{Q}_1 - \dot{Q}_3 = \dot{W}_{13}. \quad (44.8)$$

We can now obtain the loss which enters the efficiencies of the engines, because this clearly must be some kind of relationship between the efficiencies of engines running between the temperatures T_1 and T_3 , and between T_2 and T_3 , and between T_1 and T_2 .

We can make the argument very clear in the following way. We have just seen that we can always relate the heat absorbed at T_1 to the heat delivered at T_3 by finding the heat delivered at some other temperature T_2 . Therefore we can get all the engines' properties if we know, or a standard temperature, everything with that standard temperature. In other words, if we know the efficiency of an engine running between certain temperatures T_1 and a certain arbitrary standard temperature, then we can work out the efficiency for any other difference in temperatures. Because we know, because using only reversible engines, we can work from the initial temperature down to the standard temperature and back up to the final temperature again. We can define the standard temperature arbitrarily as one degree. We shall then adopt a special symbol for the heat which is delivered at this standard temperature, which shall call it \dot{Q}_s . In other words, when a reversible engine absorbs the heat \dot{Q}_1 at temperature T_1 it will deliver, at the unit temperature, a heat \dot{Q}_s . If our engine absorbing heat \dot{Q}_1 at T_1 delivers the heat \dot{Q}_s at one degree, and if an engine absorbing heat \dot{Q}_2 at temperature T_2 will also deliver the same heat \dot{Q}_s at one degree, then it follows that an engine which absorbs heat \dot{Q}_1 at temperature T_1 will deliver heat \dot{Q}_s if it runs between T_1 and T_2 , as we have already proved by considering engines running between these temperatures. So all we really have to do is to find the standard heat \dot{Q}_s , we need to put in at the temperature of T_1 in order to deliver a certain amount of heat \dot{Q}_s at the unit temperature. If we discover that, we have everything. The heat, \dot{Q}_s , of course, is a function of the engine, i.e., T_1 . It is easy to see that the heat must increase as the temperature increases, for we know that it takes work to run an engine backwards and deliver heat at a higher temperature. It is also easy to see that the heat \dot{Q}_s must be proportional to \dot{Q}_1 . Our important law is now quite simple. For a given amount of heat \dot{Q}_s delivered at one degree from an engine running at temperature T_1 degrees, the heat \dot{Q}_s delivered must be that amount \dot{Q}_s times unity increasing linearly of the temperature:

$$\dot{Q}_s = Q_s(T). \quad (44.9)$$

44-5 The thermodynamic temperature

At this stage we are not going to try to find the formula for the above increasing function of the temperature in terms of our familiar centigrade temperature scale, but instead we shall define temperature by a new scale. At one time "the temperature" was defined empirically by dividing the expansion of water into even degrees of a certain size. But when one compares expansion with a mercury thermometer, one finds that the degrees are no longer even. But now we can make a definition of temperature which is independent of any particular substance. We can use the function $\beta(T)$ which does not depend on what substance we use, because the behavior of these reversible engines is independent of their working substances. Since the function β is constant with temperature, we will define the function $\theta(T)$ as the temperature, measured in units of the standard atmosphere temperature, as follows:

$$\theta = \beta T, \quad (44-10)$$

where

$$\theta_0 = \beta \cdot T_0. \quad (44-11)$$

This means that the heat lost on object is by forcing our standard heat, is also lost by a reversible engine working between the temperature of the object and the unit temperature (Fig. 44-6). If seven times more heat is lost out of a boiler than is delivered to a condenser, the temperature of the boiler will be only seven degrees, and so forth. By measuring how much heat is absorbed at the unit temperature, we determine the temperature. This is just what is done in this very simple absolute thermometric temperature, and it is independent of the substance. We shall use this definition exclusively from now on.¹

Now we see that when we have one engine, one working between T_1 and one degree, the odds making between T_1 and one degree defining the same heat θ_1 , or a temperature, then the heat absorbed must be called θ_1 .

$$\frac{\theta_1}{T_1} = \beta = \frac{\theta_2}{T_2}. \quad (44-12)$$

This also means that if we have a single engine running between T_1 and T_2 , then the result of the above analysis (by analogy) is that θ_1 is to T_1 as θ_2 is to T_2 . If the engine absorbs energy θ_1 at temperature T_1 and gives off heat θ_2 at temperature T_2 . Whenever the engine is reversible, the ratio between the heat absorbed to heat given off is the center of the ratios of the thermodynamic.

If this is all there is to thermodynamics, why is it considered such a difficult subject? In doing a problem involving a given mass of some substance, the condition of the substance at any moment can be described by telling what its temperature is and what its volume is. If we know the temperature and volume of a substance, and the substance is some function of the temperature and volume, then we know the internal energy. One could say, "This is not very difficult way. Tell me the temperature and the pressure, and I will tell you the volume." I can think of the volume as a function of temperature and pressure, and the internal energy is a function of temperature and pressure, and so on." That is why thermodynamics is hard, because everyone uses a different approach. If we would only sit down once and decide on our variables, and let's do here, it would be fairly easy.

Now we start to make distinctions. Just as $T = \theta_0$ is the center of the universe in mechanics, and it goes on and on and on after that, in the same way the principle just found is all there is to thermodynamics. They can one make contributions to it?

¹ We have previously defined our scale of temperature in a different way, namely by saying that the mean kinetic energy of a molecule is proportional to the temperature, and that the performance of an engine is proportional to T . Is this new definition equivalent? Yes, since the first result (44-11) derived from the first law is the same as that given here. We shall discuss this point again in the next chapter.

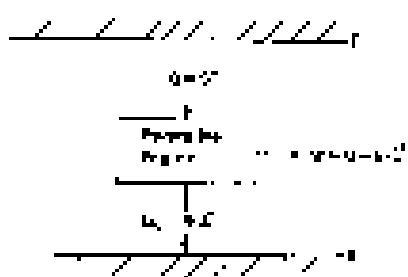


Fig. 44-6. Absolute thermodynamic temperature.

We begin. To obtain our first conclusion, we shall combine heat, law, the law of conservation of energy and this law which relates the heats Q_1 and Q_2 , and we can easily obtain the efficiency of a reversible engine. From the first law, we have $W = Q_1 - Q_2$. According to the new principle,

$$Q_1 = \frac{T_1}{T_2} Q_2$$

so the work becomes

$$W = Q_1 \left(1 - \frac{T_2}{T_1}\right) = Q_1 \frac{T_1 - T_2}{T_1} \quad (44.10)$$

which tells us the efficiency of the engine. The total work we get out of the mass, however. The efficiency of an engine is proportional to the difference in the temperatures between which the engine runs, divided by the higher temperature.

$$\text{Efficiency} = \frac{W}{Q_1} = \frac{T_1 - T_2}{T_1} \quad (44.11)$$

The efficiency of most engines at their normal operating temperatures is around 20% to 30%, absolute zero. So, since T_1 must be positive, the efficiency is always less than unity. That is our first conclusion.

4.1.6 Entropy

Equation 44.7 or 44.10 can be interpreted in a special way. Working always with reversible engines, a heat Q_1 at temperature T_1 is "reciprocated" to Q_2 at T_2 if $Q_1/T_1 = Q_2/T_2$, in the sense that one is absorbed the other is released. This suggests that if we call Q/T remaining, we can say: "a reversible process as much Q/T is associated to it liberates, there is no gain or loss of Q/T . $T_1 > Q/T$ is called entropy, and we say "heat is nothing but a change in entropy in a reversible system." If $Q/T = 0$, then the entropy is Q/T or, as we symbolize it, $Q/T^0 = S$. Actually, S is the letter usually used for entropy, and it is numerically equal to the heat (which we have called Q) of unit area. However, (entropy is not itself a heat, it's heat divided by a temperature, because it's measured in joules per degree).

Now it's interesting that besides the pressure, which is a function of the temperature and the volume, and the internal energy, which is a function of temperature and volume, we have four variables, all of which is a function of the condition, i.e., the entropy of the substance. Let us try to explain how we denote S_1 and S_2 : We mean what we call a "function of the condition." Consider the system in two different conditions, a and b , we find in the experiment where we did the adiabatic and isothermal expansions (thermodynamically, there is no need that the engine does only these processes, it could have three or four different expansions at which it takes in and delivers heat, and so on). We can move around on a $p-V$ diagram a , over the cycle, and go from one condition to $a \rightarrow b \rightarrow a$. In other words, we could do the same in a certain condition a , and then in going over to some other condition b , and we will require that this transition, made known to be reversible. Now suppose that all along the path from a to b we have to overcome at different temperatures, so that the heat dQ removed from the substance at each little step is converted to each reservoir of the temperature corresponding to that point on the path. Then let us assume all these reservoirs, by whatever means, to a range between the two temperatures. When we are finished carrying the substance from a to b , we still bring it to the case we took off their original condition. Any heat dQ that has been released from the substance at temperature T has now been converted by a reversible machine, and a certain amount of entropy dS has been added at the unit temperature as follows:

$$dS = dQ/T \quad (44.12)$$

Let us determine the total amount of entropy which has been released. The entropy difference is the entropy needed to get S out of S_1 by this path.
 44-10

reversible transformation, is the total entropy, the total of the entropy taken out of the system, minus, and different at the unit temperature:

$$S_e - S_u = \int_{T_1}^{T_2} \frac{dQ}{T}. \quad (44.16)$$

The question is, does the entropy difference depend upon the path taken? There is more than one way to go from a to b . Remember that in the Carnot cycle we could go from a to b (Fig. 44-10) by first expanding isothermally and then adiabatically, or we could first expand adiabatically and then isothermally. So the question is whether the entropy change, which occurs when we go from a to b in Fig. 44-10 in the same OR ONE CYCLE AS IT WOULD ANOTHER. It would be reasonable because if we went $a \rightarrow b$ via around the cycle, going forward then $a \rightarrow b$ and backward on another, we would have a reversible engine, and there would be no loss of heat to the reservoir at unit temperature. In a truly reversible cycle, no heat may be taken from the reservoir at the unit temperature; so the entropy needed to go from a to b is the same over one part as it is over another. It is independent of path and depends only on the endpoints. We can, therefore, say that entropy is a *state function*, which we shall mean simply by *state*, thus depending only on the condition, i.e., only on the volume and temperature.

We can find a function $S(T)$ which has this property—the “law” implies the change in entropy, as the substance is moved along any reversible path, in terms of its heat removed at unit temperature, then

$$\Delta S = \int \frac{dQ}{T}, \quad (44.17)$$

where dQ is the heat removed from the substance at temperature T . This total entropy change is the difference between the entropy calculated at the initial and final points:

$$\Delta S = S(T_2, T_1) - S(T_1, T_2) = \int_{T_1}^{T_2} \frac{dQ}{T}. \quad (44.18)$$

This expression does not completely define the entropy, but rather only the difference of entropy between two different temperatures. Only if we can calculate the entropy for one special condition can we truly define a *temperature*.

For a long time it was believed that absolute entropy meant nothing, that only differences could be defined—but finally Nernst proposed what he called the *Absolute Entropy*, which is also called the *Third Law of Thermodynamics*. It is very simple. We will see what it is, but we will not explain why it is true. Nernst's postulate states simply that the entropy of any object at zero temperature is zero. We know of one case at $T = 0$ K, namely $S = 0$, where S is zero; and so we can set the entropy at any value point.

To give an illustration of the reverse of this idea, let me take up a perfect gas. Let us isothermal (and therefore reversible) expansion, $\int dQ/T = 0/J$, since T is constant. Therefore (from 44.17) the change in entropy is

$$S(T_2, T_1) - S(T_1, T_2) = \theta R \ln \frac{V_2}{V_1}$$

so $S(T, V) = \theta R \ln V$ is some function of V only. Now does S depend on T ? We know that for a free volume adiabatic expansion, S does not increase. Thus entropy does not change even though T changes, provided that V changes such that $TV^{\gamma-1} = \text{constant}$. Can you see that this implies that

$$S(V, T) = \theta k \left[\ln T + \frac{1}{\gamma-1} \ln V \right] + c,$$

where c is a constant, independent of both V and T . θ is called the *absolute constant*. It depends on the gas in question, and may be determined experimentally. From the Carnot Cycle, by measuring the heat liberated in cooling and condensing

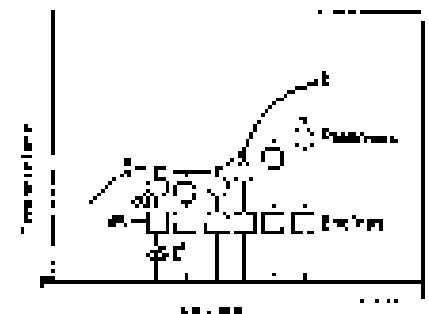


Fig. 44-10. Change in entropy during a reversible isothermal cycle.

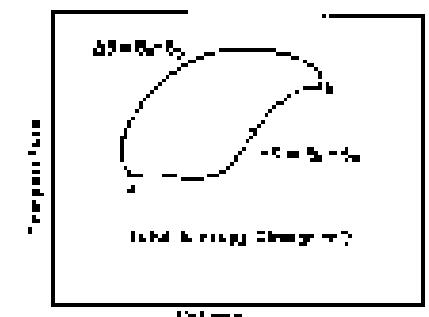


Fig. 44-11. Change in entropy in a non-reversibly reversible cycle.

the gas, and it is brought to a point (or for liquid, a liquid) at T , by expending $\delta Q/T$. It can also be demonstrated theoretically by means of Planck's constant and quantum mechanics, but we shall not study it in this course.)

Now we shall review some of the properties of the entropy of things. We first remember that if we do Q of reversible work on a system by δQ , then the entropy of the substance will change by $\Delta S = \delta Q/T$. And we remember that as we go along the path, the entropy—the heat delivered at temperature—increases according to the rule $dS = dQ/T$, where $dQ > 0$ the heat we remove from the substance when its temperature is T .

We already know that if we have a reversible cycle, the total entropy of everything is not changed, because the heat (Q) absorbed at T_1 and the heat (Q) released at T_2 correspond to equal and opposite changes in entropy, so that the net change in the entropy is zero. So for a reversible cycle there is no change in the entropy of anything, including the universe. This rule may not be the conservation of energy again, but it is odd; it applies only to reversible cycles. If we include irreversible cycles there is no law of conservation of entropy.

We shall give two examples. First, suppose that we do irreversible work on an object by friction, passing Q and Q' on some object at temperature T . The entropy is increased by Q/T . The heat Q is used in the work, and thus when we do a certain amount of work by friction against an object whose temperature is T , the entropy of the whole world increases by Q/T .

Another example of irreversibility is this: If we put together two objects that are at different temperatures, say T_1 and T_2 , a certain amount of heat will flow from one to the other by itself. Suppose, for instance, we put a hot stone in cold water. Then when a certain heat ΔQ is transferred from T_1 to T_2 , how much does the entropy of the hot stone change? It decreases by $\Delta Q/T_1$. How much does the water entropy change? It increases by $\Delta Q/T_2$. The heat will, of course, flow only from the higher temperature T_1 to the lower temperature T_2 , so that ΔQ is positive $\geq T_1 > T_2$. So the change in entropy of the whole world is positive, and it is the difference of the two first laws:

$$\Delta S = \frac{\Delta Q}{T_2} - \frac{\Delta Q}{T_1}. \quad (44.19)$$

So the following proposition is true: In any process that is irreversible, the entropy of the whole world is increased. Only in reversible processes does the entropy remain constant. Since no process is truly reversible, there is always at least a little gain to the universe, a reversible process is an idealization of which we have nothing to gain or entropy-minimal.

Unfortunately, we are not going to enter into the field of thermodynamics very far. Our purpose is only to illustrate the principal ideas involved and the reasons why it is possible to make such arguments. You will find that thermodynamics very often in this course. Thermodynamics is used very often by engineers and particularly by chemists. So we must learn some thermodynamics in physics in chemistry or engineering. Because it is not *mathematics*, complicating everything, we shall just give basic conclusions of the origin of the theory, rather than teach all the general applications.

The two laws of thermodynamics are often stated this way:

First law: the energy of the universe is always constant

Second law: the entropy of the universe is always increasing

This is not a very good statement of the second law; it does not say, for example, that in a reversible cycle the entropy stays the same, and it does not say exactly what the entropy is. It is just a close approximation to the two laws, but it does not really tell us exactly what it means. We have summarized the laws discussed in this chapter in Table 44-1. In the next chapter we shall apply these laws in disease. The relationship between the heat generated in the expansion of a rubber band, and the entropy increase when it is heated.

Table 44.1
Summary of the laws of thermodynamics

First law:

Heat added to a system \rightarrow work done on a system \rightarrow increase in internal energy of the system:

$$\delta Q + \delta W \approx dU.$$

Second law:

A common usage (and net result) is to take heat from a reservoir and convert it to work is impossible:

No heat engine taking heat Q_1 from T_1 and delivering heat Q_2, T_2 can do more work than a reversible engine, for which

$$\Delta U = Q_1 - Q_2 = Q_1 \left(\frac{T_2 - T_1}{T_1} \right).$$

The entropy of a system is often (but not always)

(a) If heat ΔQ is added reversibly to a system at temperature T , the increase in entropy of the system is $\Delta S = \Delta Q/T$.

(b) $dS = 0$ if $T = 0$ (absolute zero)

In a reversible change, the total entropy of all parts of the system (including reservoirs) does not change.

In irreversible change, the total entropy of the system always increases.

Illustrations of Thermodynamics

45-1 Internal energy

Thermodynamics is a rather difficult and complex subject when we want to apply it, and it is not appropriate for us to go very far into the applications at this time. The subject is of very great importance, however, to engineers and chemists, and those who are interested in the subject can learn about the requirements in physical chemistry or in engineering thermodynamics. There are also good reference books, such as Zemansky's *Heat and Thermodynamics*, where one can learn more about the subject. In the Encyclopedia Britannica, four volumes, one can find excellent articles on thermodynamics and thermodynamics, and in *Physical Chemistry*, there are four chapters on physical chemistry, vapor law, liquefaction of gases, and so on.

The subject of thermodynamics is complicated because there are so many different ways of describing something. If we wish to describe the behavior of a gas, we can say that the measure depends on the temperature and on the volume, or we can say that the volume depends on the temperature and the pressure. Or, were we to refer to the internal energy U , we might say that it depends on the temperature and volume, if those are the variables we have chosen. But we might also say that it depends on the temperature and the pressure, or the pressure and the volume, and so on. In the last chapter we discussed a certain function of temperature and volume, called the entropy S , and we can of course construct as many other functions of these variables as we like: T . S is a function of temperature and volume. We see here a large number of different quantities which can be functions of many different combinations of variables.

To keep the subject simple in this chapter, we shall stick at the earliest and most basic and reduce to the independent variables. Because we know, as said previously, that it is easier to measure and control in chemical experiments, here we shall use temperature and volume throughout this chapter, except in one place where we shall see how to make the transformation into the common system of variables.

We shall first then, consider only one system of independent variables: temperature and volume. Secondly, we shall discuss only two dependent functions: the internal energy and the pressure. All the other functions can be derived from these, so it is not necessary to discuss them. With these limits, however, thermodynamics is still a fairly difficult subject, but it is not quite so impossible.

First we shall review some mathematics. If a quantity is a function of two variables, the idea of the derivative of the quantity receives a little more emphasis than for the case where there is only one variable. What do we mean by the derivative of A with respect to the variable T ? The pressure changes accompanying a change in the temperature depends partly, of course, on what happens in the volume while T is changing. We must specify the change in V before the concept of a derivative with respect to T has a precise meaning. We might ask, for example, for the rate of change of P with respect to T if V is held constant. This is not just the ordinary derivative that we usually write as dA/dT . We systematically use a special symbol, $\partial A/\partial T$, to remind us that A depends on several variables P and V is on T , and that A 's other variable is held constant. We do not only use the symbol ∂ to call attention to the fact that the other variable is held constant, but we also use the notation that A held constant as a value $(\partial A/\partial T)_P$. Since we have only two independent variables, this notation is redundant, but it will help us keep our wits about us in the thermodynamic jungle of partial derivatives.

45-2 Entropy

45-2 Applications

45-3 The Clausius-Clapeyron equation

Let us suppose that the function $f(x, y)$ depends on two independent variables x and y . By $\partial f/\partial x$, we mean simply the ordinary derivative, obtained in the usual way, if we treat y as a constant:

$$\left(\frac{\partial f}{\partial x}\right)_y = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}.$$

Similarly, we define

$$\left(\frac{\partial f}{\partial y}\right)_x = \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}.$$

For example, if $f(x, y) = x^2 + xy$, then $(\partial f/\partial x)_y = 2x + y$, and $(\partial f/\partial y)_x = x$. You can no longer refer to simple derivatives: $\partial^2 f/\partial x^2$ or $\partial^2 f/\partial y^2$. The latter symbol indicates that we first differentiate f with respect to x , treating y as a constant, then differentiate the result with respect to y , treating x as a constant. The overall order of differentiation is immaterial: $\partial^2 f/\partial x \partial y = \partial^2 f/\partial y \partial x$.

We will need to compute the change Δf in $f(x, y)$ when we change $x \rightarrow x + \Delta x$ and $y \rightarrow y + \Delta y$. We assume throughout the following that Δx and Δy are infinitesimally small:

$$\begin{aligned} \Delta f &= f(x + \Delta x, y + \Delta y) - f(x, y) \\ &= \underbrace{f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y)}_{\Delta x \left(\frac{\partial f}{\partial x} \right)_y} + \underbrace{f(x, y + \Delta y) - f(x, y)}_{\Delta y \left(\frac{\partial f}{\partial y} \right)_x}. \end{aligned} \quad (45.1)$$

The last equation is the **differential relation** that expresses Δf in terms of Δx and Δy .

As an example of the use of this relation, let us calculate the change in the internal energy $U(T, V)$ when the temperature changes from T to $T + \Delta T$ and the volume changes from V to $V + \Delta V$. Using Eq. (45.1), we write

$$\Delta U = \Delta T \left(\frac{\partial U}{\partial T} \right)_V + \Delta V \left(\frac{\partial U}{\partial V} \right)_T. \quad (45.2)$$

In our last chapter we found a similar expression for the change ΔH in the enthalpy when a quantity of heat ΔQ was added to the gas:

$$\Delta H = \Delta T + P \Delta V. \quad (45.3)$$

In comparing Eqs. (45.2) and (45.3) one might be led to think that $P = (\partial H/\partial V)_T$, but this is not correct. To obtain the correct relation, let us first suppose that we add a quantity of heat ΔQ to the gas while keeping the volume constant, so that $\Delta V = 0$. With $\Delta V = 0$, Eqs. (45.2) and (45.3) reduce to $\Delta U = \Delta Q$, and Eqs. (45.2) and (45.3) reduce to $\Delta H = (\partial U/\partial T)_V + \Delta Q$, so that $(\partial U/\partial T)_V = \Delta Q/P$. The ratio $\Delta Q/P$, the amount of heat one must put into a substance in order to change its temperature by one degree with the volume held constant, is called the **specific heat at constant volume** and is designated by the symbol C_V . By this argument we have shown that

$$\left(\frac{\partial U}{\partial T}\right)_V = C_V. \quad (45.4)$$

Now let us again add a quantity of heat ΔQ to the gas, but this time we will hold T constant and allow the volume to change by ΔV . The analysis in this case is more complex, but we can calculate ΔH by the argument of Carnot, and find the result of the Carnot cycle we introduced in the last chapter:

The pressure-volume diagram for the Carnot cycle is shown in Fig. 45.1. As we have already seen, the total amount of work done by the gas in a reversible cycle is $\Delta Q(T_f/V_f)$, where ΔQ is the amount of heat energy added to the gas as it expands adiabatically at the temperature T from volume V_i to $V_f = \Delta V$, and $T_f = T_i$. The final temperature reached by the gas is, in accordance with the first law of thermodynamics, given by the equation

the shaded area in Fig. 45-1. In some circumstances, the work done by the gas is $P dV$, and its positive when the gas expands and negative when the gas is compressed. If we plot $P \times V$, the variation of $P \times V$ is represented by a curve which gives the value of P corresponding to a particular value of V . As the volume changes from one value to another, the work done by the gas (in ergs) in dV is the area under the curve connecting the initial and final values of P . When we apply this idea to the Carnot cycle, we see that as we go around the cycle, owing attention to the sign of the work done by the gas, the net work done by the gas is just the shaded area in Fig. 45-1.

Now we want to evaluate the shaded area geometrically. The cycle we have used in Fig. 45-1 differs from that used in the previous chapter in that we now suppose the ΔT and ΔQ are "infinitesimally small". We are working between adiabatic lines and isothermal lines that are very close together, and the "steps" described by the broken lines in Fig. 45-1 will approach a parallelogram as the increments dT and dQ approach zero. The area of this parallelogram is just $dP dV$, where dV is the change in volume as energy dQ is added to the gas at constant temperature, and dP is the change in pressure as the temperature changes by dT at constant volume. One can easily show that the shaded area in Fig. 45-1 is given by $dP dV$ by recognizing that the shaded area is equal to the area enclosed by the dashed lines in Fig. 45-2, which is just dV times the rectangle bounded by dP and dV only by the addition and subtraction of the shaded triangular areas in Fig. 45-2.

Now let us summarize the main result of the arguments we have developed so far.

$$\left. \begin{aligned} \text{Work done by the gas} &= \text{shaded area} = \Delta Q \left(\frac{\Delta T}{T} \right) \\ \text{or} \\ &= \frac{\Delta T}{T} \cdot \text{heat needed to change } V \text{ by } dV \text{ at } T \\ \text{or} \\ &= \Delta P \cdot \text{change in } T \text{ when } T \text{ changes by } dT \text{ at } V \\ \text{or} \\ &= \frac{1}{2} V \cdot \text{heat needed to change } T \text{ by } dT, \quad T \left(\frac{\partial U}{\partial T} \right)_V \end{aligned} \right\} \quad (45.7)$$

Equation (45.7) expresses the essential result of Carnot's argument. The whole of thermodynamics can be deduced from Eq. (45.7) and the First Law, which is stated in Eq. (45.3). Equation (45.7) is essentially the Second Law, although it was originally deduced by Carnot in a slightly different form, since he did not use our definition of temperature.

Now we are prepared to calculate $(\partial U / \partial T)_V$. By how much would the internal energy change if we changed the volume by dV ? First, $dU = -dQ + dW$. This is first, and second, dQ comes because work is done. The heat term is

$$dQ = T \left(\frac{\partial F}{\partial T} \right)_V dT,$$

according to Eq. (45.5), and the work done is $-F dV$. Therefore

$$dU = T \left(\frac{\partial F}{\partial T} \right)_V dT - F dV. \quad (45.8)$$

Dividing both sides by dV , we get for the rate of change of U with V at constant T

$$\left(\frac{\partial U}{\partial V} \right)_T = T \left(\frac{\partial F}{\partial V} \right)_T = P. \quad (45.9)$$

In most thermodynamics, in which F and P are the only variables and P and V are the only functions, Eqs. (45.6) and (45.9) are the basic equations from which all the rest of the subject can be deduced.

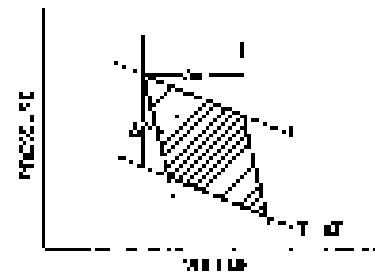


Fig. 45-1. Pressure-volume diagram for a Carnot cycle. The curve meeting T and $T - \Delta T$ are isothermal lines; the steeper curves are adiabatic lines. ΔV is the volume change at heat ΔQ is added to the gas at constant temperature T . ΔP is the pressure change in constant volume at the gas temperature is changed from T to $T - \Delta T$.



Fig. 45-2. Shaded area = area enclosed by dashed lines = area of rectangle = $dP dV$.

45.2 Applications

Now let us discuss the meaning of Eq. (45.7) and see why it answers the questions which we proposed in our last chapter. We considered the following problem: at constant density, the volume of a rubber band depends on temperature, because of the thermal movements of the atoms of a rubber. For the same pressure, suppose when we let the piston move back, that is when out of the gas and, in order to keep the temperature constant, heat will have to be put back in. The question is when it expands, and the pressure rises what is needed? There must be some connection between these two phenomena, and this connection is given explicitly by Eq. (45.7). If we hold the volume fixed and increase the temperature, the pressure rises at a rate $(\partial P / \partial T)_V$. Related to that fact is this: if we increase the volume, the gas will cool unless we put some heat in; in accordance with the temperature condition, and only ΔU tells us the amount of heat needed to raise the temperature. Equation (45.7) expresses the fundamental interrelationship between these two effects. That's where we predicted we would find when we discussed the laws of thermodynamics. Without knowing the internal mechanism of the gas, and knowing only that we cannot make perpetual motion of the second type, we can deduce the relationship between the amount of heat needed to maintain constant temperature when the gas expands, and the pressure change when the gas is cooled.

Now that we have Eq. (45.7), we want to apply it to determine the rubber band. When we stretch it, that is, when we heat it, its temperature falls, and when we heat a rubber band, we find that it pulls itself in. What is the equation that gives the same relation for a rubber band as Eq. (45.7) gives for gas? For a rubber band equilibrium will be something like this: when $\Delta U = 0$, i.e., the internal energy is changed by ΔU and ΔU is zero work is done. The only difference will be that the work done by the rubber band is $-F dL$ instead of $P dV$, where F is the force on each end, and L is the length of the band. The value F is a function of temperature and of length of the band. Replacing ΔU in Eq. (45.7) by $-F dL$, we get

$$\Delta U = -F dL + \Delta U_{\text{ext}} \quad (45.8)$$

Comparing Eqs. (45.7) and (45.8), we see that the rubber band equation is obtained by a direct substitution of one term for another. Furthermore, if we substitute $-F$ for P , and $-L$ for V , in our discussion of the Carnot cycle applies to the rubber band. We can immediately deduce, for instance, that the heat ΔU needed to stretch the band by dL is given by the expression $(\partial U / \partial L)_{T,P} dL = F(T, P, L) dL$. This equation tells us that if we keep the length of a rubber band fixed and heat the band, we can calculate how much the force will increase in terms of the heat, needed to keep the temperature constant when the band is stretched a little bit. So we see that the same equation applies to both gas and a rubber band. In fact, if one can write $\Delta U = -F dL + \Delta U_{\text{ext}}$, where F and ΔU_{ext} represent different quantities, force and length, pressure and volume, the result may apply to heat conduction. For example, substituting dA and θ for dL and T . For example, consider the electric potential ϕ between the "voltage," V in a battery and the "charge," Q , that moves through the battery. We know that the work done in passing the charge, Q , through the battery, is QV . (Since we include a QV term in the work, we require that our battery maintains a constant voltage.) Of course, what thermodynamics can tell us about the performance of a battery. If we substitute dA for dL and θ for T in Eq. (45.8), we obtain

$$\frac{\Delta U}{dA} = -F \left(\frac{\partial V}{\partial Q} \right)_A \theta \quad (45.9)$$

Equation (45.9) says that the internal energy U is changed when a charge dQ flows through the cell. What is $\Delta U/dQ$ not simply the voltage V ? In battery? The answer is that a real battery gets warm when charge moves through the cell. The internal energy of the battery is changed, first, because the battery did some work in the external circuit, and second, because the battery is heated. The nu-

workable thing is that the standard free enthalpy can be expressed in terms of the way in which the battery voltage changes with temperature. Fortunately, when the charge moves through the electron circuit, reaction $\text{A} + \text{B} \rightarrow \text{C}$, Eq. (45.9) says precisely the way of measuring the amount of energy required to produce a chemical reaction. All we need do is measure a cell that, when on the reaction, measure the voltage, and measure how much the voltage changes with temperature when we take an charge from the battery!

Now we have another! But the volume of the battery can't be an infrared constant, since we have obtained the $P\Delta V$ term when we do the work done by the battery equal to E_cell . That is not true; it is technically quite difficult to keep the volume constant. It is much easier to keep the total Δn constant. This approach is preferable. For this reason, the chemists do not use any of the equations we have written above; they prefer equations which do not contain Δn under constant pressure. We chose at the beginning of this chapter, after Eqs. 45.4 and 45.5, independent variables. The chemists prefer P and T , and we will now consider how the results we have obtained up to now are transformed into a new kind of system of variables. Remember that in the following treatment oxygen is a unity set in parentheses, so we are shifting gears from T and P to T and P .

We started in Eq. 45.9 with $\Delta H^\circ = \Delta Q - P\Delta V$; $P\Delta V$ may be replaced by $P\Delta T + P\Delta S$. If we could somehow replace the last term, $P\Delta S$, by ΔH° , then we would now interchanged P and T , and the chemists would be happy. We can do this if we notice that the difference of the terms in Eq. 45.4 is $\Delta H^\circ + P\Delta V$, and if we add this equality to Eq. 45.5, we obtain

$$\frac{\Delta H^\circ = P\Delta V + P\Delta T}{\Delta H^\circ = \Delta Q - P\Delta S} = P\Delta T$$

$$\frac{\Delta H^\circ - \Delta Q}{\Delta H^\circ - P\Delta V} = P\Delta S$$

In order that our result look like Eq. (45.5), we define $\Delta = P\Delta T$ to be something new, called the *voltage*, Δ , and we write $\Delta = \Delta Q - P\Delta V$.

Now we are ready to transform our variables into standard language with the following rules: $\text{P} \rightarrow P$, $\text{J} \rightarrow -P$, $\text{V} \rightarrow P$. For example, the fundamental relationship that chemists would use instead of Eq. (45.7) is

$$\left(\frac{\partial f}{\partial P}\right)_T = -f\left(\frac{\partial f}{\partial T}\right)_P = 1.$$

It would now be clear how one transforms to the chemists' variables P and T . We now go back to our original variables; for the remainder of this exercise, P and T are the independent variables.

Now let us apply the results we have obtained to a number of physical situations. Consider first an ideal gas. From kinetic theory we know that the internal energy of a gas depends only on the mass of the molecule and the number of molecules. The internal energy depends on T , but not on P . If we change P but keep T constant, E is not changed. Therefore in $\partial Q/\partial T = 0$, and Eq. 45.7 reduces to the, for an ideal gas,

$$-\left(\frac{\partial f}{\partial T}\right)_P = P = 0. \quad (45.10)$$

Equation 45.10 is a differential equation for f , or something about P . We take account of the partial derivatives in the following way: Since the partial derivative is to constant P , we will replace the partial derivative by an ordinary derivative and we explicitly, to remind us, constant P . Equations 45.10) then becomes

$$\frac{\Delta f}{\Delta T} = P = 0; \quad \text{constant } P. \quad (45.11)$$

which we can integrate to get

$$\ln P = \ln T + \text{const.} = \text{const. } K, \quad (45.12)$$

$$P = \text{const. } \times T = \text{const. } K. \quad (45.12)$$

We know that for an ideal gas the pressure is equal to:

$$P = \frac{RT}{V} \quad (45.13)$$

which is consistent with (45.2), since V and R are constants. Why did we bother to go through this calculation if we already knew the results? Because we have been using the "absolute" definition of temperature. At this stage we assumed that the kinetic energy of the molecules was proportional to the temperature, an assumption that defines one scale of temperature which we will call the "ideal gas scale." The T in Eq. (45.2) is based on this gas scale. We also say a temperature measured on the gas scale must be the same. Further, we defined the temperature in a way which was completely independent of the substance. From large tables based on the Simeon's law we defined T_0 . We might call the "grand thermodynamic absolute temperature" T , the T that appears in Eq. (45.13). What we learned here is that the pressure of an ideal gas (defined as one for which the internal energy does not depend on the volume) is proportional to the T and the molecular weight of the molecule. We can know that the pressure is proportional to the temperature measured on the gas scale. Therefore when we divide total molecular temperature by proportionality constant we get the "grand thermodynamic absolute temperature," that means of course, but if we were sensible we would make two scales agree. In this instance, at least, the two scales agree because chosen so that they actually are proportional by constant has been imposed to us. Most of the time man chooses himself for himself, but in this case he made them equal.

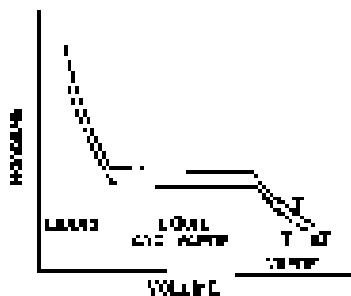


Fig. 45-9. Isothermal lines for a condensable vapor compression in a cylinder. At the left, the substance is in the liquid phase. At the right, the substance is vaporized. In the center, both liquid and vapor are present in the cylinder.

45-7 The Clausius-Clapeyron equation

The vaporization of a liquid is another application of the results we have derived. Suppose we have some liquid in a cylinder, such that we can compress it by pushing on the piston, and we ask ourselves, "If we keep the temperature constant, how does the pressure vary with volume?" In other words, we want a curve on isothermal line in the $P-V$ diagram. The substance in the cylinder is not an ideal gas but as mentioned earlier, since it may be in the liquid or the vapor phase, we both may be present. If we apply sufficient pressure, the substance will condense to a liquid. Now if we squeeze still harder, the volume changes very little, and we increase the pressure, finally with decreasing volume, as shown at the left in Fig. 45-9.

If we increase the volume, by pulling the piston out, the pressure drops until we reach the point at which the liquid starts to boil, and then vapor starts to form. If we pull the piston out further, all that happens is that more liquid evaporates while more liquid and part vapor in the cylinder. The two phases are in equilibrium. Liquid is evaporating and vapor is condensing at the same rate. If we make more room for the vapor more vapor is needed to maintain the pressure, so a little more liquid evaporates, but the pressure remains constant. On the flat part of the curve in Fig. 45-9 the pressure does not change, and the value of the press., as here is called the vapor pressure at temperature T . As we continue to increase the volume, there comes a time when there is no more liquid to evaporate. At this juncture, if we expand the cylinder further, the volume will fall as for an ordinary gas, as shown at the right of the $P-V$ diagram. The lower curve in Fig. 45-9 is the iso-enthalpic line at slightly lower temperature $T - \Delta T$. The pressure in the liquid phase is slightly reduced because liquid expands with α . Increase in temperature (for most substances, but not for water near the freezing point) and, of course, the vapor pressure is lower at the lower temperature.

We will now make a cycle out of the two isotherm lines by connecting them together by adiabatic lines between the ends of the flat sections, as shown in Fig. 45-4. The little jagged on the lower right-hand corner of the figure will indicate the difference and we will neglect. We are going to use the argument of Carnot, which tells us that the heat added to the substance is hanging it from a liquid and again is related to the work done by the substance as it goes around the cycle. Let



Fig. 45-10. Pressure-volume diagram for a Carnot cycle with a condensable vapor in the cylinder. At the left, the substance is in the liquid state. A quantity of heat Q_1 is added at temperature T_0 to vaporize the liquid. The vapor expands adiabatically as T changes to $T = T_0$.

us call L the heat needed to vaporize the substance in the cylinder. As in the experiment immediately preceding Eq. (45.13), we know that $L(2T_0) = \text{work done by the substance}$. As before, the work done by the substance is $L \times \text{vaporized area}$, which is approximately $\Delta P(T_0) \cdot V_0$, where $\Delta P(T_0)$ is the difference in vapor pressure at the two temperatures $T_0 < T < 2T_0$, V_0 is the volume of the gas, and V_0 is the volume of the liquid, both volumes measured at the vapor pressure. Setting these two expressions for the areas equal, we get $L(2T_0) = \Delta P(V_0 - V_0)$.

or

$$\frac{L}{V_0(V_0 - V_0)} = (\partial P/\partial T)_0. \quad (45.14)$$

Thus Eq. (45.14) gives the relationship between the rate of change of vapor pressure with temperature and the amount of heat required to vaporize the liquid. This relationship was deduced by Carnot, but it's called the Clausius-Clapeyron equation.

Now let us compare Eq. (45.14) with the results deduced from kinetic theory. Here T_0/V_0 is very much larger than T_0 . So $V_0 - V_0 \approx V_0 = RT/P$ per mole. If we further assume that L is a constant, independent of temperature, and if we good approximation for ΔP we have $\Delta P/T = L/V_0^2 T^2$. The solution of this differential equation is

$$P = C T e^{-L/V_0 T}, \quad (45.15)$$

Let us compare this with the power law variation with temperature that we deduced earlier from kinetic theory. Kinetic theory indicated the possibility, at least roughly, that the number of molecules of vapor above a liquid would be

$$n = \left(\frac{V}{V_0}\right) e^{-L/V_0 - 2E_0/RT}, \quad (45.16)$$

where $E_0 - V_0$ is the internal energy per mole in the liquid. Since the internal energy per mole in the gas, i.e., the energy needed to vaporize a mole of liquid. Equation (45.15) from thermodynamics and Eq. (45.16) from kinetic theory are very closely alike, because the pressures are the same, but they are not exactly the same. However, they will turn out to be exactly the same if we assume $L = T_0$, or equivalently, if $L = \text{const}$. If we ignore $L - E_0$, a constant independent of temperature, then the argument leading to Eq. (45.15) will produce Eq. (45.16).

This comparison shows the similarities and discrepancies of thermodynamics and kinetic theory. First of all, the agreement is exact, while Eq. (45.16) can only be approximated, for instance, if L is really constant and it the mole. \approx right. Second, we may not understand correctly how the gas becomes liquid, nevertheless, Eq. (45.15) is right, while (45.16) is only approximate. Third, although non-linearity applies to a gas condensing into a liquid, the argument is true for any other change of state. For instance, the solid-liquid transition has the same kind of curve as that shown in Figs. 45.3 and 45.4. Introducing the latent heat L resulting from the latent heat change in Eq. (45.14) there is $(\partial P_{\text{gas}}/\partial T)_0 = \Delta P/V_0^2 T_0 = L/V_0^2$. Although we may not understand the kinetic theory of the melting process, we nevertheless have a correct equation. However, when we compare with the kinetic theory, we have another difference. Equation (45.15) is only a *Calculus* approximation, and we have no way of obtaining the constants of integration, i.e., the kinetic theory can obtain the constants also if we have a good model that describes the phenomenon completely. So there are advantages and disadvantages to each. Why? Knowledge is weak, and the situation is complex, thermodynamical relations are only the most general. When the situation is very simple and a theoretical calculation is made, then it is better to try to go more in to mathematical theoretical analysis.

One more example: blackbody radiation. We have discussed a box containing radiation and nothing else. We have looked at the equilibrium between the radiation and the radiation. We also found that the pressure hitting the wall of the box would over the pressure P and we found $PV = U/T$, where $T > 0$.

and energy of all k -photons $\omega \in V$ is the volume of the box. If we substitute $\Omega = \partial V$ in the box Eq. (45.7), we find

$$\left(\frac{\partial \Omega}{\partial V}\right)_V = \Omega \quad \Rightarrow \quad \left(\frac{\partial \Omega}{\partial V}\right)_V = S. \quad (45.10)$$

Since the volume of our box is constant, we can replace $(\partial \Omega / \partial V)_V$ by $dV/dV = 1$ to get $S = \text{const} \propto V$. The pre-factor of V varies as the fourth power of the temperature and the energy content of the radiation, $S/V = P/V$, also varies as T^4 . It is useful to write $S/V = (4\pi/3)T^4$, where c is the speed of light and ϵ is constant. It is not possible to get ϵ from thermodynamics alone. There is a prior example of its power, and its limitations. We know that S/V goes as T^4 is a good idea, but to know how big S/V actually is at any temperature requires that we integrate the function ϵ itself. But only a complete theory can supply such backgound radiation we have since a theory and we can obtain an expression for the constant ϵ in the following manner.

Let $d\omega$ be the frequency distribution, i.e. energy flow through $\epsilon \text{ m}^2$ in one second with frequency between ω and $\omega + d\omega$. The energy density distribution = energy/volume = $E(\omega)/V$ is

$$\begin{aligned} \frac{E}{V} &= \text{total energy density} \\ &= \int_{-\infty}^{\infty} \text{energy density } \epsilon \text{ and } \omega + d\omega \\ &= \int_{-\infty}^{\infty} \frac{E(\omega) d\omega}{c}. \end{aligned}$$

From our car's discussions, we know that

$$E(\omega) = \frac{\hbar \omega^3}{\pi^2 c^3 (e^{h\omega/kT} - 1)}.$$

Substituting this expression for $E(\omega)$ in our equation for E/V , we get

$$\frac{E}{V} = \frac{1}{\pi^2 c^3} \int_1^{\infty} \frac{\hbar \omega^3 d\omega}{e^{h\omega/kT} - 1}.$$

If we substitute $x = h\omega/kT$, the expression becomes

$$\frac{E}{V} = \frac{(2\pi)^3}{h^3 c^3 k^3} \int_1^{\infty} \frac{x^3 dx}{e^x - 1}.$$

This integral is just some number that we can get, approximately, by drawing a curve and taking the area by counting squares. It is roughly 6.5. The most interesting thing we can see is that the index x is exactly ≈ 4 . Comparing this expression with $S/V = (4\pi/3)T^4$, we find

$$\sigma = \frac{8\pi^3}{90h^3 c^3} = 5.67 \times 10^{-5} \frac{\text{W/m}^2 \text{K}^4}{\text{steradiance}} \text{ (SI units)}$$

* Since $x^n = 1/x^{n-1} = 1/(x-1) + \dots$ the integral is

$$\sum_{n=1}^{\infty} \int_1^{\infty} x^{n-1} dx.$$

But $\int_1^{\infty} x^{n-1} dx = 1/n$, and differentiating with respect to n there gives $\int_1^{\infty} x^{n-1} dx = n/(n+1)$, on the interval $[1, \infty)$. $\int_1^{\infty} x^{n-1} dx$ is a good estimate except from adding the first few terms. In Chapter 20 we will find a way to show that the sum of the first N terms fourth powers of the integers is $\sim N^5/5$.

If we make a small hole in our box, how much energy will flow per second through the hole of unit area? To go from energy density to energy flow, we multiply the energy density E/V by c . We also multiply by $\frac{1}{2}$, which can be seen following three factors by 4, because only the energy above the mean contributes; and second, by the factor $\sqrt{2}$, because $c/\sqrt{2}$ is what approximates the hole at the edge to the source, in less effectively filling the right-hand hole by a linear factor. The average value of the cosine is $\frac{1}{2}$. It is clear now why we wrote $E/c^2 = \text{Joules/m}^3$; so that we can ultimately say that the flux from a $4\pi r^2$ ball is 1.7×10^{24} joules/sec.

Stated *without proof*

46-1 How a ratchet works

In the chapter we discuss the ratchet and pawl, a very simple device which allows a shaft to turn only one way. The possibility of doing something turn only one way requires some detailed and careful analysis and there are some very interesting implications.

The plan of the discussion is as follows: attempting to derive a elementary explanation from the molecular or statistical point of view, for the fact that there is a maximum amount of work which can be extracted from a heat engine. Of course we have seen the version of Carnot's argument, but it would be worthwhile to re-examine it which is elementary in the sense that we can say what is happening physically. Now, here is a simplified mathematical demonstration which follows from Newton's laws to demonstrate that we can get only a certain amount of work, when heat flows from one place to another, but there is great difficulty in converting this into an elementary demonstration. To start we do not understand it, although we can follow the mathematics.

In Carnot's argument, the fact that more than a certain amount of work cannot be extracted in going from one temperature to another is deduced from simple axiom, which is that everything is at the same temperature, heat cannot be converted to work by means of a cyclic process. But, let us back up and try to see, in at least one elementary example, why this simpler statement is true.

Let us try to have a device which will violate the Second Law of Thermodynamics, but is a ratchet which will generate work, i.e. a heat reservoir with everything at the same temperature. Let us say we have a gas at a certain temperature, and inside there is a wheel with spokes at θ . (see Fig. 46-1) further, $T_1 = T_2 = T_{\text{constant}}$. Because of the thermal motion of gas molecules on the wheel, the spokes will rotate. All we have to do is to fasten the other end of the spokes to a wheel which can only one way, the ratchet and pawl. Then when the shaft tries to joggle one way, now, the pawl will catch it against the slot, it will turn. Then the wheel will slowly turn, and perhaps we might even tie a thin string hanging from a slot on the shaft and catch it. Now let us ask if this is possible. According to Carnot's hypothesis, it is impossible. But if we just look at it, we see, without doubt, that it seems quite possible. So we must look more closely. Indeed, if we look at the mechanism now, we see a number of complications.

First, our ratchet and pawl is as complex as possible, but even so, it uses a pawl, and there must be a spring in the pawl. For now, think about after running off a bath, or trapping it, and so on.

A larger feature of this ratchet and pawl, and shown in the diagram, is quite peculiar. Suppose the device were made of perfectly elastic parts. At first the pawl is lifted off the end of the teeth and is turned back by the spring, it will bounce against the wheel and continue to bounce. Then, when enough vibration stops, the wheel could turn the other way, because the pawl would get under the teeth during the moment when the pawl was up! The other important part of the incompleteness of our wheel is a damping or decaying mechanism which stops the bouncing. When the damping happens, of course, the energy that was in the pawl goes into the wheel, and gives up heat. Since it loses heat, the wheel will get hotter and hotter. To make the thing stop, we can put a gear around the wheel to take up all the heat. Anyway, let us say the gas keeps rising in temperature, going with the wheel. Will it go on forever? Not! The gas will heat to some temperature

46-2 How a ratchet works

46-3 The ratchet as an engine

46-4 Reversibility in mechanics

46-5 Irreversibility

46-6 Order and entropy



Fig. 46-1. The ratchet and pawl mechanism.

T , also have Brownian motion. This motion is such that, every once in a while, by accident, the pawl lifts itself up and over a tooth just at the moment when the Brownian motion on the wheel is trying to turn the wheel backwards - and as things get worse, this happens more often.

So, this is the reason this device does not work in perpetual motion. When it happens get asked, sometimes the pawl lifts up and goes over the end. After some time, when it tries again to lift it's way, the pawl has already lifted due to the fluctuation and it's reflections off the wheel side, and the wheel goes back the other way! The net result is oscillation. It is not hard to demonstrate that when the temperature on the wheel is very low, there will be no net average motion of the wheel. Of course the wheel will do a lot of jiggling this way, one way, then another way.

Let us look at the reason. It is necessary to do work against the spring in order to lift the pawl to the top of a tooth. Let us call this energy e , and let θ be the angle between the teeth. The chance that the system can accumulate enough energy, e , to go the pawl over the top of the tooth, is $e^{-\theta/e}$. But the probability that the pawl will accidentally, for up to down, $e^{-\theta/e}$. So the number of times that the pawl is up and the wheel turns in one direction freely is equal to the number of times the pawl has enough energy to turn it forward when the pawl is down. We thus get a balance, and the wheel will not go anywhere.

46-2 The ratchet sees an engine

Let us now go further. Take the example where the temperature at the bottom is T_1 and the temperature of the wheel, or pawl, is T_2 , and T_2 is less than T_1 . Because the wheel is cold and the oscillations of the pawl are also less frequent, it will be very hard for the pawl to gain an energy e . Because of the high temperature T_1 , the pawl will often drain the energy e , so our engine will not run unless it is designed.

You would now like to see if it can lift weights. On to the drum is the end of the string, and puts a weight, such as you have, on the string. We let L be the length from the weight. If L is not too great, our machine will lift the weight because the pawl is flat; certain parts of it move easily to receive in one direction than the other. We want to find how much weight it can lift. How fast it goes, currently, and so on.

First we consider a general motion, the usual way one designs a ratchet mechanism. In order to make one step forward, I now must energy losses be borrowed from the total body. We must have an energy e to lift the pawl. The wheel turns through an angle θ again: $\theta \propto e / E$, so we can read the $e = k \theta / M$. The total amount of energy E and we have to burn e is thus $e = k \theta / M$. The probability that we get this energy is proportional to $e^{-k \theta / M}$. Actually, it is not only a case that of getting the energy, but we also would like to know the number of times the pawl has the energy. The probability per second is proportional to $e^{-k \theta / M}$, and you shall see it be proportionality constant k_1 . It will depend on the end anyway. When a forward step happens, the work done on the weight is Ld . The energy taken from the wire is $e + Ld$. The spring gets wound up with energy e , but it goes farther, easier, longer, and it is energy e to move the L . The energy taken out goes off the weight, and in doing so, the weight falls back and gives loss. In the other case,

Now we look at the opposite case, which is harder, if you can. Why happens here? To get the wheel to go forward again, we have to do a step up the energy e to lift the pawl high enough so that the tooth will slip. This is still energy e . The probability per second for the pawl to lift this high is now $(k_1 e)^{-\theta/e}$. The proportionality constant is the same, but this time $k_1 e$ shows up because of the cold environment. When this happens, the work is released because the wheel slips backwards. It loses potential, and it releases work in. The energy taken from the ratchet system is e , and the energy given to the gas at T_1 at the wire side is $(k_1 e)^{-\theta/e}$. It takes a little thinking about the reason for this. Suppose the pawl has turned itself up upside down by a fluctuation. Then, when it falls back and the spring e .

Table 46-1

Economy of operation of calculator and pen.

Method	Number of steps	ΔT	Time spent	Rate	Efficiency
Calculator	26	1	1 min	26/min	100%
Does not work	12	1	1 min	12/min	100%
Calculator + pen	1	1	1 min	1/min	100%

Method	Number of steps	ΔT	Time spent	Rate	Efficiency
Calculator	26	1	1 min	26/min	100%
Does not work	12	1	1 min	12/min	100%
Calculator + pen	1	1	1 min	1/min	100%

If calculator does not work, then:	$\frac{\Delta T}{T_1} = \frac{1}{2}$
Calculator rate: Does not work	$\frac{1}{2} \cdot \frac{T_1}{\Delta T} = \frac{T_1}{2\Delta T}$

Calculator rate: Does not work	$\frac{1}{2} \cdot \frac{T_1}{\Delta T} = \frac{T_1}{2\Delta T}$	However	$\frac{T_1}{\Delta T} = \frac{T_1}{\Delta T}$
-----------------------------------	--	---------	---

pushes it down against the earth. There is a force trying to turn the wheel, whereas the wheel is pushing out at a constant speed. This force is doing work, and so a net force due to the weight. So out, together cause no net force, and all the energy which is slowly released appears at the same end as heat. Of course it must be used rather of energy, but one must be careful to this, the units though! We notice that all these energies are exactly the same, however we add. So, depending upon which of these two rates is greater, the weight is either slowly lifted or slowly lowered. Of course, it is obviously going around, picking up both up and down T_1 & T_2 , which, as we are talking about the average behavior.

Suppose that for a particular weight the rates happen to be equal. Then we exert an infinitesimal weight to the system. The weight goes slowly up down, and work will be done on the machine. Let us say it takes 1 unit of weight given to the system. Then when we take off a little bit of weight, then the infinitesimal is the corresponding. For example, if 1 unit of heat is taken from the wheel and put into the wheel. So we have the components of Carnot's reversible cycle, provided that the weight is just such that the two are equal. This means that it is unlikely that $T_1 = 1.000T_2$, etc., let us say that the machine is slowly lifting the weight. Energy is released from the system and change the heat added to the wheel, and heat removed is ΔT in the cycle. If $T_1 = T_2$ then, if we are lowering the weight, we also have $T_1/\Delta T = T_2/\Delta T$. Thus (Table 46-1) we have

$$T_1/\Delta T = T_2/\Delta T.$$

Furthermore, the work we get out is to the energy taken from the same is ΔT is $-T_1 + T_2$ & hence is $(T_1 - T_2)/T_1$. We see that our theory cannot release more work than the Carnot process does. This is the result that we expected from Carnot's experiment, and the main result of this lecture. However, we can now consider the mechanics of a lot of other phenomena, even out of equilibrium, and therefore beyond the range of thermodynamics.

Let us immediately think about some very device we'd like if every thing were in the same temperature and we bring a weight on the system. If we pull very, very hard, of course, there are all kinds of complications. The pen also uses the number of the spine breaking or something. But suppose we pull gently enough that everything works nicely. In these circumstances, the above analysis is right for the probability of the weight going forward and backward. If we assume

that the two tensions are not equal. In each step the angle β is obtained, so the angular velocity γ is β times the amplitude of one of these ratios per second. It goes forward with probability $p = 1/(p_0 e^{-\beta} + p_1 e^{\beta})$ and backward with probability $(1-p)/e^{-\beta}$. So for future angular velocity we have:

$$\begin{aligned}\omega &= (p_1/p_0 e^{-\beta} + p_0 e^{\beta}) \omega \\ &\cdot (1, e^{-\beta}, e^{\beta}, 1, \dots) \quad (46.1)\end{aligned}$$

If we plot ω against β , we get the curve shown in Fig. 46-2. We see that it makes a great difference whether β is positive or negative. If β increases in the positive range, which happens when we try to drive the wheel backward, the backward-velocity approaches a constant. As β becomes negative, ω really "takes off" forward, since in a tremendous power is very great!

The angular velocity law we obtain from different forces is thus very unsymmetrical. Going one way it is easy; we get a lot of angular velocity for a little force. Going the other way, we can put on a lot of force, and yet the wheel hardly goes around.

We did the same thing for a electrical rectifier. Instead of the force, we have the electric field, and instead of the angular velocity, we have the electric current. In the case of a rectifier, the voltage is not proportional to resistance, and the situation is unsymmetrical. The same analysis can, we think for the mechanical rectifier as I also work for an electrical rectifier. In fact, the kind of formula we obtained above is typical of the current-voltage capacities of transistors as a function of their voltage.

Now let us take off the weight away, and look at the original machine. If T_2 were less than T_1 , the racket would go forward as anybody would believe. But when I tried to believe, at first, it was opposite. If T_2 is greater than T_1 , the racket goes around the opposite way. A dynamic racket will hit its heel on its own tail end, the heel of the racket, just by running. If the racket, for a moment, is on the saddle somewhere, pushes the saddle plane sideways. But it is always pushing on the saddle plane, because it .. happens to lift up high enough to just past the peak of a tooth, then the saddle plane slides by, and it comes down again on an inclined plane. So a hot racket and ball is really built to go around in a direction exactly opposite to that for which it was originally designed.

In spite of the cleverness of Captain George, if the two temperatures are exactly equal there is no more propensity to go one way than the other. The moment we look at it, it may be running one way or the other, but at the long run, it goes nowhere. The fact that it goes nowhere is really the fundamental deep principle on which all of thermodynamics is based.

46-3 Reversibility in mechanics

What does mechanical irreversibility tell, in the long run, if the temperature is kept the same everywhere? A gadget will turn neither to the right nor to the left. We evidently have a fundamental proposition that there is no way to keep a machine which, left to itself, will go more with to the left than to the right after a long enough time. We must try to see how this follows from the laws of mechanics.

The laws of mechanics go something like this: we must find the acceleration \ddot{x} the force, and the force on each particle is some complicated function of the positions of all the other particles. There are other situations in which forces depend on velocity, such as in magnetism, but let us not consider that now. We take a simple case, such as gravity. These forces depend only on position. Now suppose that we have solved our set of equations and we have a certain motion $x(t)$ for each particle. In a complicated enough system, the solutions are very complicated, and what happens with this seems not to be very surprising. If we write down any arrangement of places for the particles, we will see this arrangement actually occur if we wait long enough! That's how our solution for a long

enough terms, it does everything that it can do, so it stuck. This is not a pathology however - in the simplest device, but when you get something *exactly* with enough terms, it happens. Now there's something else the solution can do. If we solve the equations of motion, we may get certain functions such as $r + r^2 + \dots$. We claim that one term would be $-r + r^2 - \dots$. In other words, if we substitute $-r$ everywhere for r , then up to the entire solution, we will have again a solution of the same equation. This follows from the fact that if we substitute $-r$ for r in the original differential equation, nothing is changed, since only second derivatives with respect to r appear. This means that if we have a solution, then the exact opposite motion is also possible. In the complete confusion which comes if we wait long enough, it finds itself going one way sometimes, and it finds itself going the other way sometimes. There is nothing more than the random walk of the motion that that other side is impossible to design a machine which, in the long run, is more likely to be going one way than the other, if the machine is sufficiently complicated.

You might think up an example for which this is obviously untrue. If we take a wheel, for instance, and spin it at high speed... will go the same way forever? That's a common condition, like the conservation of angular momentum, which might not always implement - the just requires that the argument be made with a little more care. Perhaps the wheel like all the angular momentum, or something similar, on that we have no special conservation law. Even, if the system is complicated enough, the argument is true. It is based on the fact that the laws of mechanics are reversible.

For my next invention, we would like to repeat on a device invented by Maxwell, who had worked out the dynamical theory of gases. We suppose the following situation. We have two boxes of gas at the same temperature, with a little hole between them. At one side sits a little demon (who has been invented by the author). He watches the molecules coming from the left. Whenever he sees a fast molecule, he opens the door. Whenever he sees a slow one, he leaves it closed. If we want him to be an extra special demon, he can have eyes on the back of his head, and see the signs of the molecules from the other side. He lets the slow ones through to the left, and the fast ones through to the right. Pretty soon, the left side will get cold and the right side hot. Then, is the idea of thermodynamics violated because we could have such a demon?

It turns out that he has a limited-sized demon, for the demon himself goes in reverse. He comes out very well after a while. The simplest possible demon, as an example, would be a trap door held over the hole by a spring. A fast molecule comes through, because it is able to lift the trap door. The slow molecule cannot get through, and becomes fast. This big thing is nothing but one ratchet and pawl in another form, and ultimately, the ratchet will heat up. If we assume that the quantity N of the demon is not infinite, it must heat up. It has but a finite number of internal gears and wheels, so it cannot get rid of the extra heat that it gets from overdriving the mechanism. Even if it is shaking from vibration, for example, that a ratchet will continue it is moving or going, much less whether the molecules are coming or going, so it does not work.

46-4 Irreversibility

Are all the laws of physics reversible? Obviously not! Just try to make a billiard ball stop. Run a moving picture backwards, and it looks only a few minutes for everybody to start to split. The may be the characteristic of all phenomena is even obvious irreversibility.

What does irreversibility come from? In the end, one of our Newtonian laws. If we claim that the behavior of everything is naturally to be understood in terms of the laws of physics, and if it also turns out that all the equations have the *conservative* property, but if we put $v = -v$ we have another equation, then every phenomenon is reversible. Now then, does it come about in nature on a large scale that things are not like this? Of course there must be some law, some

obscure but fundamental question: perhaps it is causality, maybe it is previous physics, in which it does not do which way time goes.

Let us discuss this question now. We already know one of these laws, which says that the entropy is always increasing. If we have a hot thing and a cold thing, the heat goes from hot to cold. An arrow of entropy is well defined. But, we expect to understand the law of entropy from the point of view of mechanics. In fact, we have just been successful in this in our first few consequences of the argument that heat comes from the motion by itself from just mechanical arguments, and we thereby obtained an understanding of the Second Law. Apparently we can get irreversibility from irreversibility of time. Let me tell why a mechanical argument fails miserably. Let me look into it more closely.

Since our question has to do with the entropy, our problem is to try to find a microscopic description of entropy. If we say we have a certain amount of gas in a container like a gas, then we can get a microscopic picture of it, and say each atom has a certain energy. All these energies added together give us the total entropy. Similarly, maybe every atom has a certain enthalpy. If we add everything up, we would have the total enthalpy. It does not work out well; let us see what happens.

As an example, let's calculate the entropy difference between a gas at a certain temperature at one volume, and a gas at the same temperature at another volume. We remember from Chapter 44, that we have for the change in entropy,

$$\Delta S = \int \frac{dQ}{T}.$$

In the present case, the energy of the gas is the same whether it is also expanding. Since the last part does not change, we have to add enough heat to equal the work done by the gas, or, for each little change of volume,

$$dQ = P dV.$$

Putting this in for dQ , we get

$$\begin{aligned}\Delta S &= \int_{V_1}^{V_2} \frac{P dV}{T} = \int_{V_1}^{V_2} \frac{RT dV}{V T} \\ &= R \ln \frac{V_2}{V_1},\end{aligned}$$

as we obtained in Chapter 44. For instance, if we expand the volume by a factor of 2, the entropy change is $R \ln 2 \approx 1.39$.

Let us now consider another little interesting topic. Suppose we have a box with a barrier in the middle. One molecule is moving ("heat" molecule), and on the other, sugar ("solvent" molecule). Now we take out the barrier, and let them mix. How much has the entropy changed? It is possible to imagine, for most of the barrier we have a piston, with piston rings that pass the water through but not the sugar, and another kind of piston which is the other way around. If we move one piston to the side, and we see that, for each gas, the problem is like the one we just solved. So we get an entropy change of 1.39 . In 2, which means that the entropy was increased by π in 2 per molecule. The 2 has to do with the π to mean that the molecule has, which is rather peculiar. It is not a property of the molecule itself, but of how much room the molecule has to run around in. This is a strange situation, where the two molecules mix, where everything has the same temperature and the same energy. The only thing that is changed is that the molecules are distributed differently.

We still know that it is not well. We know that everything will go round up after a long time due to the collisions. Collisions, the banging, and so on. Every now and then a water molecule goes toward a black end, a black end is covered a while, and maybe they pass. I, personally, do not care; their is no, by accident, across into the space of black, and the blacks were their. And, by accident, into the space of white. If we wait long enough we get a mixture.

Clearly, there is an irreversible process in the real world, and ought to be also in the laws of the universe.

Now we have a simple example of an irreversible process which is completely composed of reversible events. Every time there is a collision between any two molecules, they go \rightarrow in certain directions. If we took a moving picture of a collision it would show nothing wrong with the picture. In fact, one kind of collision is just as likely as another. So no mixing is completely reversible, and yet it is irreversible. You know that if we start with white and black, separated, we would get a mixture within a few minutes. If we wait and looked at it for several more minutes, it would not separate again and go to its original. So we have an irreversibility which is based on reversible situations. But we also see the reason why. We started with an arrangement which is far from equilibrium. During the cleavage of the C_6H_6 sites, it becomes disordered. But the change from an ordered arrangement to a disordered arrangement which is the source of the irreversibility.

It is true that if we took a motion picture of this, and showed it backwards, we would see it gradually become ordered. Some one would say, "That is against the laws of physics!" So we would run the film forwards again, and we would look at every collision. Every one would be perfect, and every one would be obeying the laws of physics. The reason, however, is that every molecule's velocities are just right so if the plates are all followed back, they get back to their original condition. But that's a very unlikely circumstance to have. If we start with the gas in an excited state (permanently white and black), it never gets back.

46-5 Order and entropy

So we now have to talk about what we mean by disorder. And what we mean by order. It is not a question of pleasant order or unpleasant disorder. What is different in our mind and in mind of others is the following. Suppose we divide the space into little volume elements. Now, how many small black molecules, how many white could we distribute them among the volume elements so the white is on one side, and black on the other? On the other hand, how many ways could we distribute them with no restriction on which goes where? That is, there are many more ways to arrange than in the latter case. We measure "disorder" by the number of ways that arrangements can be arranged, + that from the outside it looks the same. The logarithm of that number of ways is the entropy. The number of ways in the segregated case is less, so the entropy is less, or less "disorder" is less.

So with the above rather abstract definition of disorder we can understand the proposition. First, the entropy measures the disorder. Second, the universe always goes from "order" to "disorder," so entropy always increases. Order is the order in the sense that we pick up arrangement, but in the sense that the number of different ways we can look it up, and still have it look the same from the outside, is relatively increased. In the case where we reversed our initial problem of the gas burning, there was not as much regularity now. Tonight, those single atoms had exactly the same speed and direction, so count out right! The entropy was not high after all, even though it appeared so.

What about the reversibility of the other physical laws? When we talked about the electric field which comes from an accelerating charge, we said that we must take the retarded field. At a time t , and at a distance r from the charge, we take the field due to the acceleration at a time $t - r/c$, not $t + r/c$. So it goes to zero, as if the law of electricity is not reversible. Very strangely, however, the laws we used come from a set of equations called Maxwell's equations, which are, in fact, reversible. Furthermore, it is possible to argue that if we were to use only the retarded field, the field due to the state of affairs at $t - r/c$, and do it absolutely consistently in a completely arbitrary spatial, everything happens exactly the same way as if we had used both! This apparent irreversibility in electricity, at least in an enclosure, is thus not an irreversibility at all. We have some feeling for that already, because we know that when we have an oscillating charge which generates fields which are reflected from the walls of an enclosure we ultimately

you in an equilibrium at which there is no inequalities. The number of free up objects is only a very small fraction in the mass of a solution.

So it has been known all the fundamental laws of physics, like Newton's equations, are reversible. Even where does irreversibility come from? It comes from order giving to disorder. In fact it is irreversibility that we know the origin of irreversibility. Why is that the processes we find ourselves in every day are always out of equilibrium? One possible explanation is the following. Consider again our box of mixed air and blue molecules. Now it is possible if we wait long enough by sheer, grossly improbable, but possible accident that the distribution of molecules gets to be mostly white on one side and mostly blue on the other. After that, as time goes on and we wait longer, they get more mixed up again.

The one possible explanation of the high degree of order in the present-day world is that it is just a question of luck. Because our universe happened to have had a "lucky" big bang in the past, at which the present equivalent age is 13.7 and now they are running back together again. This kind of theory is not irreversibility, because we could still when the separate gas looks like water a little in the future or a little in the past. In either case, across a gray area, is the interface, because the molecules are mixing again. No matter which way we run time, the gas is gas. But this theory would say the irreversibility is just one of the accidents of it.

We would like to argue that this is not the case. Suppose we do not look at the whole box, but only a piece of the box. That is a certain moment, suppose we discover a certain amount of order. In this little piece, white and black are separate. What if and what does happen between two places where we look and yet "order"? If we truly believe that the order came from something disorder by a fluctuation, we must surely take the same likely fluctuation which could produce it and the most likely conclusion is not that the rest of it has also become disordered! Therefore, from the hypothesis that the world is a "statistic" all of the predictions are right. If we look at a part of the world we have never seen before, we will find it mixed up, not like the piece we just looked at. If our model is random no fluctuation, we would not expect order anywhere but where we have just looked at.

Now we claim the argument is because the past of the universe was really ordered. It is not due to a fluctuation, but the whole thing used to be white and black. This theory now predicts that there will be order at older places - the order is not due to fluctuation, but due to a much higher order at the beginning of time - then we would expect to find order in places where we have not yet looked.

The astronomers, for example, have only looked at some of the stars. They may have sent their telescopes to other stars, and to new stars observing the same things as the other stars. We therefore conclude that our universe is not a fluctuation, and the "background" is a memory of conditions when things are hot. This is not to say that we understand the logic of it. For some reason, in the sense at one time had a very low entropy, very strong constraint and since then the entropy has increased. So entropy goes toward the future. That is the origin of all irreversibility, that is what makes the process of growth and decay, that makes us consider the past and not the future, remember the things which are closer to the minimum in the history of the universe when the order was higher than now, and why we are not able to measure things where the disorder is bigger than now, which we call the figure. So, as we demonstrated in an earlier lecture, the entire universe is like a glass of wine. If we look at it closely enough, in this case the glass of wine is complex, because there's water and glass and light and everything else.

A central concept of our subject of physics is that even simple and idealized things like the wheel and pulley work only because they are part of the universe. The wheel and pulley works in only one direction because it has some interaction contact with the rest of the universe. The wheel and pulley were in a box and isolated for some sufficient time, the wheel would not rotate and there would be no friction. But because we pull up the string and let the right up, because

we cool off on the earth and go hot from the sun, the nucleus and gamma that we make can turn one way. This one wayness is interrelated with cause and effect that is part of the universe. It is part of the universe not only in the sense that it obeys the physical laws of the universe, but its one-way behavior is tied to the one-way behavior of the entire universe. It cannot be completely understood until the mystery of the beginning of the history of the universe are released with further speculations to scientific understanding.

Chapter 47: The wave equation

47-1 Waves

In this chapter we shall discuss the phenomenon of waves. This is a phenomenon which appears in many contexts throughout physics, and therefore our attention should be concentrated on it not only because of the particular example considered here, which is sound, but also because of the most widespread application of the ideas in all branches of physics.

It was pointed out when we studied the harmonic oscillator that there are many mechanical examples of oscillating systems but, clearly, not as well. Waves are related to oscillating systems, many other wave oscillations appear not only as time oscillations at one place but propagate in space as well.

We have met only optical waves. When we studied light in learning about the properties of waves in that subject, we paid particular attention to the interference of waves from several sources at different locations and all at the same frequency. There are two important wave phenomena that we have not yet discussed which occur in light, i.e., electromagnetic waves, as well as in any other form of waves. The first of these is the phenomenon of interference or cancellation in space. If we have two sources of sound which have slightly different frequencies and if we listen to both at the same time, then sometimes the waves come with the crests together and sometimes with the crest and the pit together (see Fig. 47-1). The strong amplifying of one wave that results is the phenomenon of beats; it is called interference in time. The second phenomenon involves the wave pattern which results when the waves are confined within a given volume and reflect back and forth from walls.

These effects could have been discussed of course for the case of electromagnetic waves. The reason for my having done it only that my first non-example would not generate difficulties and we are actually learning about more in the sub-subject at the same time. I made the emphasis the general applicability of wave theory; electrodynamics, we consider here a different example, in particular sound waves.

Other examples of waves are water waves consisting of long waves that we see coming to the shore, or the smaller ripples consisting of surface tension ripples. As another example, there are two kinds of elastic waves in solids, a transverse (or longitudinal) wave in which the particles of the solid vibrate back and forth along the direction of propagation of the wave (so-called waves in a gas are of this kind), and a shear wave in which the particles of the solid oscillate in a direction perpendicular to the direction of propagation. Longitudinal waves contain shear waves of both kinds, just as by a magnet a source due to the current's effect.

Still another example of waves is found in modern physics. These are waves which give the probability amplitude of finding a particle at a given place. The "matter waves" which we have already mentioned. Their energy is proportional to the energy and their wave number is proportional to the momentum. They are the waves of quantum mechanics.

In this chapter we do consider only waves in which frequency is independent of the wavelength. This is, for example, the case for light in a vacuum. The speed of light is then the same for audiences. $V = C/\lambda$, green light, blue light, for any other wavelength. Because of this behavior, when we began to study the wave phenomena we did not look at first directly at wave propagation. Instead, we said that if a charge is moved at one place, the electric field will change a wave

47-1 Waves

47-2 The propagation of sound

47-3 The wave equation

47-4 Solutions of the wave equation

47-5 The speed of sound

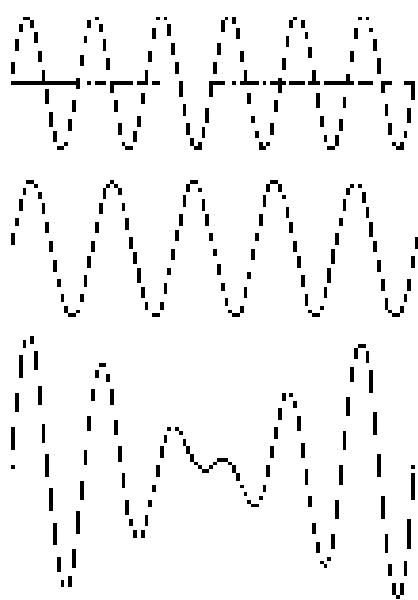


Fig. 47-1. Interference in time of two sound waves with slightly different frequencies, resulting in beats.

proportional to the acceleration, not at the time t , but at the earlier time x/c . Therefore if we were to picture the electric field in space as static instead of linear, as in Fig. 47-2, the electric field at a place x later would have started the distance $c t$, as indicated in the figure. Mathematically, we can say that in the case mentioned example we are taking the electric field as a function of $x - ct$. We see that at $x = 0$, it is zero because of x . If we consider a later time, we need only increase x somewhat to get the same value of the electric field. For example, if the maximum field occurred at $x = 0$ at time t_0 , now to find the new position of the maximum field at time t we need

$$x - ct = 0 \quad \text{or} \quad x = ct + x_0$$

We see that this kind of function represents the propagation of a wave.

Such a function, $f(x - ct)$, then represents a wave. We may summarize this propagation of a wave by saying simply that

$$\hat{f}(x - ct) \quad \hat{f}(x + ct - ct) = \hat{f}(x)$$

when $ct = c\Delta t$. There is, of course, another possibility, i.e., that instead of a source to the left as indicated in Fig. 47-2, we have a source on the right so that the wave propagates toward negative x . Then the wave would be described by $\hat{g}(x - ct)$.

There is the additional possibility that more than one wave exists in space at the same time, and in the double pendulum the sum of the two fields, even one propagating independently. The behavior of electric fields may be described by saying that if $f_1(x - ct)$ is a wave, and if $f_2(x - ct)$ is another wave, then their sum is also a wave. This is called the principle of superposition. The same principle is valid in sound.

We are familiar with the fact that if sound is produced, we hear will exemplify fidelity the same sequence of sounds we generated. If we had high frequencies travelling faster than low frequencies, a short, sharp noise would be heard as a succession of musical sounds. Similarly, if red light travelled faster than blue light, a beam of white light would appear first as red, then as yellow, and finally as blue. We are familiar with the fact that this is not the case. Both sound and light travel with a speed in air which is very nearly independent of frequency. Examples of wave propagation for which this independence is not true will be considered in Chapter 48.

In the case of light (electromagnetic waves) we give a rule which summarizes the electric field at a point as a result of the acceleration of a charge. One might expect here that we should do the same for the velocity since quality of the air, say the pressure, is determined by a given distance from a source in terms of the source motion, caused by the travel time of the sound. In the case of light this procedure was conceivable because all that we know was that a charge at one place exerted force on another charge at another place. The details of propagation from one place to the other were not absolutely essential. In the case of sound, however, we know that it propagates through the air between the source and the hearer, and it is certainly a natural question to ask what, in any given instant, the pressure of the air is. We would like, in addition, to know exactly how the air moves. In the case of electricity we could accept a rule, since we could say that we do not yet know the laws of electricity, but we cannot make the same inference with regard to sound. We would not be satisfied with a rule stating that the sound pressure increases through the air because the pressure ought to be understandable as a consequence of the laws of mechanics. In short, sound is a branch of mechanics and so it is to be understood in terms of Newton's laws. The propagation of sound from one place to another is merely a consequence of mechanics and the properties of gases, that propagates to a gas or of the properties of liquids or solids if it propagates through such materials. Later we shall derive the properties of light and its wave propagation in a similar way from the laws of electrodynamics.

47-2 The propagation of sound

We shall give a derivation of the properties of the propagation of sound between the source and the receiver as a consequence of Newton's laws, and we shall not consider the interaction with the source and the receiver. Ordinarily we expect to have a result rather than a particular derivation of it. In this chapter we take the opposite view. The point here, is a certain sense, is "the derivation itself". This problem of explaining wave phenomena in terms of what goes on when we have the laws of the real world, is perhaps the greatest task of mathematical physics. The mathematical physicist has two main kinds of tools at his disposal, given the experiments, and the other is to find the equations which describe a new phenomenon. The derivation here is an example of the second kind of problem.

We shall take the simplest example here—the propagation of sound in one dimension. To carry out such a derivation it is necessary first to have some kind of understanding of what is going on. Fundamentally what is involved is this: if an object is moved at one place in the air, we observe that there is a disturbance which travels out from the air. If we ask what kind of disturbance, we would say that we would expect that the reaction of the object is a change of pressure. Of course, the object is moved, and there is only "local motion", but that is not enough; we must say that it is a rapid motion so that there is not sufficient time for such a flow. Then, with the motion, the air is compressed and a change of pressure is produced which pushes on additional air. This air is in turn compressed, which leads again to an even greater pressure, and a wave is generated.

We now want to formulate such a process. We have to decide what variables we need. In our previous work— we would need to know how much the air has moved, so that the air displacement at the source wave is certainly one relevant variable. In addition we would like to describe how fast it "velocity" changes as it is displaced. The air pressure is another, or this is another variable of interest. Last, of course, the air does nothing, so that we shall have to describe the velocity of the air particles. The air, before it is forced to vibrate, tends to be at rest; many variables are even needed that the velocity and acceleration would be known if we knew how the air displacement varies with time.

As we said, we shall consider the wave in one dimension. We can do this if we are sufficiently far from the source that what we call the wavefronts are very nearly planes. We thus make our argument simpler by taking the least complicated example. We shall then be able to say that the displacement ψ depends only on x and t , and not on y and z . Before the description of $\psi(x,t)$ is given by us, is it

In this description complete? It would appear to be far from complete. Do we know now of the details of how the air molecules are moving? They are moving in all directions, and this state of affairs is certainly not described by means of the function $\psi(x,t)$. From the point of view of kinetic theory, if we have a large density of molecules at one point and a lower density elsewhere, in that place, the molecules would move away from the region of higher density to the one of lower density, we would expect this. Apparently we could not get such a situation and there would be no air out. What is necessary to get the sound wave is this condition: as the molecules drift out of the region of higher density and take pressure, they give momentum to the molecules in the adjacent region of lower density. In order to be generated, the regions over which the density and pressure change must be much larger than the distance the molecules travel between collisions. This distance is the mean free path, and the distance between successive crests and troughs must be much larger than this. Otherwise the molecules would move right from the crest to the trough and immediately return to the wave.

It is clear that we are going to require this sort of behavior on a scale large compared with the mean free path, and so the properties of the gas will not be described in terms of the individual molecules. The displacement, for example, will be the displacement of the center of mass of a small element of the gas, and the pressure or density will be the pressure or density in the region. We shall call the pressure P and the density ρ , and they will be functions of x and t . We must keep in mind that this description is an approximation which is valid only when these quantities are changing slowly with distance.

47-4 The wave equation

The physics of the propagation of sound waves that involves mass densities:

- I. The expansion and change in density.
- II. The change in density corresponds to a change in pressure.
- III. Pressure inequalities generate gas motion.

Let us consider at first the case of a liquid, or a solid, the pressure is some function of the density. Before the sound wave arrives, we have equilibrium, with a pressure P_0 and a corresponding density ρ_0 . A pressure P in the medium is connected to the density by some characteristic relation $P = f(\rho)$ and, in particular, the equilibrium pressure P_0 is given by $P_0 = f(\rho_0)$. The change of pressure in sound from the equilibrium value is denoted by ΔP . A convenient unit for measuring pressure is the kilobars (1 bar = 10^5 newton/m²). The pressure of dry air at sea level is very nearly 1 bar; 1 atm = 1.0133 bars. In sound we use a logarithmic scale of intensities, since the sensitivity of the ear is roughly logarithmic. This scale is the decibel scale, in which the acoustic pressure level for the pressure amplitude is defined as

$$\text{Decibel pressure level} = 20 \log_{10}(P/P_0) \text{ db}, \quad (47.1)$$

where the reference pressure $P_0 = 1 \times 10^{-14}$ bar. A pressure amplitude of $P = 10^2 P_0 = 2 \times 10^{-12}$ bar corresponds to a moderately intense sound of 30 decibels. Notice that the pressure change is usually extremely small compared with the equilibrium, or mean, pressure of 1 atm. The displacement and the density changes are also correspondingly extremely small. In explosions we do not have such small changes; the excess pressure amplitude can be greater than 1 atm. These large pressure changes lead to new effects which we shall consider later. In sound we do not often consider acoustic intensity levels over 100 db, though it is a condition painful to the ear. Therefore, for sound, if we write

$$P = P_0 + \delta P, \quad \rho = \rho_0 + \delta \rho, \quad (47.2)$$

we shall always have the pressure change δP , very small compared with P_0 , and the density change $\delta \rho$, very small compared with ρ_0 . Then

$$P_0 - P_1 = f(\rho_0 - \delta \rho) - f(\rho_1) + \delta P/f(\rho_0), \quad (47.3)$$

where $f'(\rho_0)$ and $f'(\rho_1)$ denote the derivatives of $f(\rho)$ evaluated at $\rho = \rho_0$. We can take the second step to this equality only because $\delta \rho$ is very small. We find in this way that the excess pressure δP is proportional to the excess density $\delta \rho$, and we may call the proportionality factor α

$$\delta P = \alpha \delta \rho, \quad \text{where } \alpha = f'(\rho_0) = (\partial P / \partial \rho)_0. \quad (47.4)$$

The result we wanted for α is the very simple one:

Let us now consider I. We shall suppose that the position of a portion of air initially by the sound wave is x and the displacement of the air due to the sound is $\delta(x, t)$, so that its new position is $x - \delta(x, t)$ as in Fig. 47-5. Now the initial local position of a nearby particle $\delta(x, t)$ is $x + \delta x$, and its new position is $x - \delta x - \delta(x, t)$. We can now find the density change in the following way. Since we are limiting ourselves to plane waves, we can take a unit area perpendicular to the direction in which is the direction of propagation of the sound wave. The amount of air, per unit area, in Δx is then $\rho_0 \Delta x$, where ρ_0 is the undisturbed, or equilibrium, air density. The air, when displaced by the sound waves, now lies between $x + \delta(x, t)$ and $x + \delta x - \delta(x, t)$, so that we have the same number of this interval that was in Δx when undisplaced. If ρ is the new density, then

$$\rho_0 \Delta x = \rho(x + \delta x + \delta(x, t) + \delta x, t) - \rho(x + \delta(x, t)) \quad (47.5)$$

* With this value of ρ_0 , the δ is not a real pressure in the sound wave but the "root-mean-square" pressure, which is $(\rho)^{1/2}$ times the real pressure.

Since Δx is small, we can write $\rho(x + \Delta x, t) = \rho(x, t) + (\partial\rho/\partial x)\Delta x$. This derivative is a partial derivative, since x depends on t , just as well as on x . Our equation then is

$$\rho_0 \ddot{x} = -\rho \left(\frac{\partial v}{\partial x} + \dot{v} \right) \quad (47.6)$$

or

$$\rho_0 \ddot{x} = \rho_0 \left(1 - \rho_0 \frac{\partial^2 v}{\partial x^2} \right) + \rho_0 \dot{v} \quad (47.7)$$

Now in our previous chapter we saw that $\rho_0 > \rho$, so that $\rho_0 - \rho$ is small, and $\partial\rho/\partial x$ is also small. Therefore in the second term we can just forget,

$$\rho_0 \ddot{x} = \rho_0 \frac{\partial^2 v}{\partial x^2} - \rho_0 \frac{\partial v}{\partial x} \quad (47.8)$$

and we neglect $\rho_0 \partial v / \partial x$ compared with $\rho_0 \partial^2 v / \partial x^2$. Thus we get the relation we needed for 1.

$$\rho_0 \ddot{x} = -\rho_0 \frac{\partial^2 v}{\partial x^2} \quad (47.9)$$

This equation is what we would expect classically. If the displacement v varies with x , then there will be density changes. The sign is also right: if the displacement v increases with x , so that $\partial v / \partial x$ is positive, then the density ρ must decrease, as shown.

We now need the third equation, which is the equation of the motion produced by the pressure. If we know the relation between the force and the pressure, we can then get the equation of motion. If we take a thin slab of air of length Δx and of unit area perpendicular to x , then the mass of air in this slab is $\rho_0 \Delta x$ and it has the mass in $\text{kg}/\text{m}^2 \Delta x^2$, so the mass times the acceleration of the slab is $\rho_0 \Delta x \cdot 10^3 \text{kg}/\text{m}^2 \Delta x^2$. (It makes no difference to small Δx whether the acceleration $\partial^2 v / \partial t^2$ is evaluated at the edge of the slab or at some intermediate position.) If now we find the force on this slab due to a pressure P perpendicular to x , it will turn out to be $\Delta x \cdot 10^3 P \text{N}$. We have the force in the x -direction at x , of amount $P(x, t)$ per unit area, and we have the force in the opposite direction, at $x + \Delta x$, of amount $P(x + \Delta x, t)$ per unit area (Fig. 47-4).

$$P(x, t) - P(x + \Delta x, t) = -\frac{\partial P}{\partial x} \Delta x = -\frac{\partial P}{\partial x} \Delta x, \quad (47.10)$$

since Δx is small and since the only part of P which changes is the surface pressure P . We have then 1:

$$\rho_0 \frac{\partial^2 v}{\partial x^2} = -\frac{\partial P}{\partial x}, \quad (47.11)$$

and we now have enough equations to incorporate strings and reduce down to one variable, say v in X . We can eliminate P from 1,2 by using 1, so that we get

$$\rho_0 \frac{\partial^2 v}{\partial x^2} = -v \frac{\partial^2 v}{\partial t^2}. \quad (47.12)$$

and here we can use 1 to eliminate v_t . In this way we find that v cancels out and that we are left with

$$\frac{\partial^2 v}{\partial x^2} = \frac{1}{c_p^2} \frac{\partial^2 v}{\partial t^2}. \quad (47.13)$$

We shall call $c_p^2 = c_p$, and then we can write

$$\frac{\partial^2 v}{\partial x^2} = \frac{1}{c_p^2} \frac{\partial^2 v}{\partial t^2}. \quad (47.14)$$

This is the wave equation which describes the behavior of sound in matter.

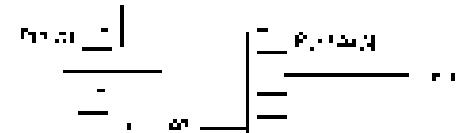


Fig. 47-4 The net force in the positive x -direction is reduced by the pressure owing on one area perpendicular to x in $-(\partial P / \partial x) \Delta x$.

47.4 Solution of the wave equation

We must ask whether this equation really does describe the essential properties of sound waves in matter. We want to consider now a second pulse, or disturbance, that moves with a constant speed. We want to verify that two different pulses can move through each other, the principle of superposition. We also want to verify that the initial disturbance is the right one in the left. All these properties should be contained in this one equation.

We have remarked that any plane-wave disturbance which moves with a constant velocity v has the form $f(x - vt)$. Now we have to see whether $x(t, y) = f(y - vt)$ is a solution of the wave equation. When we calculate $\partial_x^2 x$, we get the derivative of the function, $\partial_x f(x) = f'(x - vt)$. Differentiating once more, we find

$$\frac{\partial^2 x}{\partial t^2} = f''(x - vt). \quad (47.12)$$

The differentiation of this same function with respect to t gives $-v$ times the derivative of the function, or $\partial_t^2 x = -vf'(x - vt)$, and the second time derivative is

$$\frac{\partial^2 x}{\partial y^2} = v^2 f''(x - vt). \quad (47.13)$$

It is evident that $f(x - vt)$ will satisfy the wave equation provided the wave velocity v is equal to v_0 .

We find, therefore, from the law of propagation that any sound disturbance propagates with the velocity v_0 , and in addition we find that

$$c_s = v^{1/2} = (kT/m)^{1/2},$$

and so we have reduced the wave velocity to a property of the medium.

If we consider a wave travelling in the opposite direction, so that $x(t, y) = f(x + vt)$, it is easy to see that such a disturbance also satisfies the wave equation. The only difference between such a wave and one travelling from left to right is in the sign of v , but when we have $x + vt$ or $x - vt$ as the argument in the function does not affect the sign of $\partial^2 x / \partial t^2$, since it involves only v^2 . It follows that we have a solution for waves propagating in either direction with speed v_0 .

An extremely interesting question is related to that of superposition. Suppose one solution of the wave equation has been found, say x_1 . This means that the second derivative of x_1 with respect to x is equal to $(\partial^2 x_1 / \partial t^2)$, times the second derivative of x_1 with respect to t . Now any other solution x_2 has this same property. If we superpose these two solutions, we have

$$x(t, y) = x_1(x, t) + x_2(x, t), \quad (47.17)$$

and we wish to verify that $x(t, y)$ is also a wave, i.e., that x satisfies the wave equation. We can easily prove this result, since we have

$$\frac{\partial^2 x}{\partial y^2} = \frac{\partial^2 x_1}{\partial y^2} + \frac{\partial^2 x_2}{\partial y^2}, \quad (47.18)$$

and, in addition,

$$\frac{\partial^2 x}{\partial t^2} = \frac{\partial^2 x_1}{\partial t^2} + \frac{\partial^2 x_2}{\partial t^2}. \quad (47.19)$$

It follows that $\partial^2 x / \partial x \partial t = (\partial^2 x_1 / \partial x \partial t) + (\partial^2 x_2 / \partial x \partial t)$, so we have verified the principle of superposition. The proof of the principle of superposition follows from the fact that the wave equation is linear in x .

We can now expect that a plane longitudinal wave propagating in the x -direction, polarized so that the electric field is in the y -direction, will satisfy the wave equation,

$$\frac{\partial^2 E_y}{\partial x^2} = \frac{1}{c_s^2} \frac{\partial^2 E_y}{\partial t^2}. \quad (47.20)$$

where c is the speed of light. This wave equation is one of the consequences of Maxwell's equations. The equations of electrodynamics will lead to the wave equation for light just as the equations of mechanics lead to the wave equation for sound.

47.5 The speed of sound

The deduction of the wave equation for sound has given us a formula which connects the wave speed with the ratio of change of pressure with the density at the normal pressure:

$$c_s^2 = \left(\frac{dp}{\rho} \right)_s. \quad (47.21)$$

In order to bring the \rightarrow 's of change, it is assumed to know how the temperature varies. In a sound wave, we would expect that in the region of compression the temperature would be raised, and that, in the region of rarefaction, the temperature would be lowered. Newton was the first to calculate the rate of change of pressure with density, and he supposed that the temperature remained unchanged. He argued that the heat was conducted from one region to the other so rapidly that the temperature could not rise or fall. This argument gives the "athermal" speed of sound, and it is wrote. The correct deduction was given later by Laplace, who put forward the opposite idea—that the pressure and temperature change adiabatically in a sound wave. The heat flow from the compressed region to the rarefied region is negligible in time, so the wavelength is long compared with the mean free path. Under this condition, the slight amount of heat flow in a sound wave does not affect the speed, although it gives a small absorption of the sound energy. We can expect correctly that this absorption increases as the wavelength approaches the mean free path, but these wavelengths are smaller by factors of about a million than the wavelengths of audible sound.

The actual variation of pressure with density is, except water, the one that allows no heat flow. This corresponds to the adiabatic variation, which we found to be $pV^\gamma = \text{const}$, where V was the volume. Since the density ρ varies inversely with V , the adiabatic connection becomes p and ρ is

$$\rho = \text{const}/p^\gamma. \quad (47.22)$$

From which we get $(dp)/p = -\gamma p dV/V$. We then have for the speed of sound the relation

$$c_s^2 = \frac{\gamma p}{\rho}. \quad (47.23)$$

We can also write $c_s^2 = \gamma p/\rho^\gamma$ and make use of the relation $pV = \text{const}$. Further, we see that ρV is the mass of gas, which can also be expressed as μn , or as μ , where n is the mass of 1 molecule and μ is the molecular weight. In this way we find that

$$c_s^2 = \frac{\gamma p}{n} = \frac{\gamma \mu}{\mu}, \quad (47.24)$$

from which it is evident that the speed of sound depends only on the gas temperature and not on the pressure or the density. We also have observed that

$$\delta V = \beta n (\bar{v}^2), \quad (47.25)$$

where \bar{v}^2 is the mean square of the speed of the molecules. It follows that $c_s^2 = (\delta V/\delta p) \bar{v}^2$, or

$$c_s = \left(\frac{\delta V}{\delta p} \right)^{1/2} \bar{v}_{\text{av}}. \quad (47.26)$$

This equation states that the speed of sound is some number which is roughly $1/(\beta)^{1/2}$ times some average speed, \bar{v}_{av} , of the分子 of the square root of the

and negative velocity. In other words, the speed of sound is of the same order as the mean free path of the molecules, and is normally somewhat less than this average speed.

Of course we could expect such a result, because a disturbance like a change in pressure is after all propagated by the motion of the molecules. However, such an argument does not tell us the precise propagation speed; it could have turned out that sound was carried primarily by the fastest molecules, or by the slowest molecules. It is reasonable and satisfying that the speed of sound is roughly $\frac{1}{2}$ of the average molecular speed v_{av} .

Beats

45-1 Adding two waves

Some time ago we discussed in considerable detail the properties of light waves and their interference. There, the effects of the superposition of two waves from different sources. In all those days we assumed that the frequencies of the sources were *not* the same. In this chapter we shall discuss some of the phenomena which result from the interference of two sources which have different freq. ratios.

It is very easy to see what is going on here. Proceeding in the same way as we have done previously, suppose we have two equal oscillating sources of the same frequency whose phase is unadjusted, i.e., that the signals arrive in phase at some point P . At the point of P is light, the light is very strong, E_H is small. It is very low; or if it is electrons, none of them arrive. On the other hand, if the amplitudes were 100% out of phase, we would get no signal at P , because the two amplitudes there is then a minimum. Now suppose that one source has the "phase shift" α of one of the sources and change the phase. A look and for the eye, this making it 2π and then 180° , and so on. Of course, we would then have variations in the net signal strength. Now we also see that if one phase of one source is slowly changing relative to that of the other it is possible, with our source, starting at zero, going up to ten, twenty, thirty, forty degrees, and so on, then what we would measure is it would be a series of strong and weak "permutations." Because when the phase shifts through 360° the amplitude returns to a maximum. Of course, to say that one source is shifting its phase relative to another is a uniform way of thinking by saying that the number of oscillations per second is slightly different for the two.

Now, however, if we have two sources of slightly different frequencies we should find, as a net result, an oscillation with a slowly oscillating intensity. That is, of course, is the subject.

It is very easy to formulate this in mathematics also. Suppose, for example, that we have two waves, and that we do not work for the moment about all the specific relations but simply analyze who, or how, in P . For convenience let us say we want to assume, one from the other source, one, say, where the two ω 's are not exactly the same. Of course the amplitudes may not be the same either, but we can solve the general problem later. Let us first take the case where the amplitudes are equal. Then the total amplitude A is the sum of the two changes. If we plot the amplitudes of the waves against the time, as in Fig. 45-1, we see that

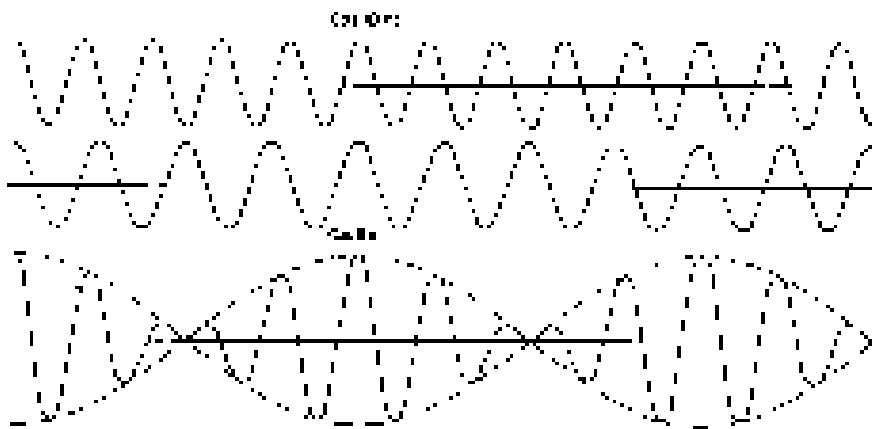


Fig. 45-1. The superposition of two acoustic waves with frequencies in the ratio 8:10. The periodic repetition of the pattern within each "beat" is not typical of the general case.

45-2 Adding two waves

45-2 Beat notes and modulation

45-3 Side bands

45-4 Localized wave zones

45-5 Probability amplitudes for particles

45-6 Waves in three dimensions

45-7 Normal modes

where the waves interfere we get a strong wave, and where a trough and crest coincide we get practically zero, and then when the waves interfere again we get a strong wave again.

Mathematically, we need only to add two cosines and determine the result somehow. There exist a number of useful relations among cosine which is not difficult to derive. Of course we know that

$$e^{i(\omega_1 t + \phi_1)} = \cos \omega_1 t + i \sin \omega_1 t, \quad (48.1)$$

and that $e^{i\theta}$ has a real part, $\cos \theta$, and an imaginary part, $\sin \theta$. If we take the real part of $e^{i(\omega_1 t + \phi_1)}$, we get $\cos(\omega_1 t + \phi_1)$. If we multiply out:

$$e^{i\theta_1} e^{i\theta_2} = (\cos \theta_1 + i \sin \theta_1)(\cos \theta_2 + i \sin \theta_2),$$

we get $\cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2$, plus some imaginary parts. But we now need only the real part, so we have

$$\cos(\theta_1 + \theta_2) = \cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2. \quad (48.2)$$

Now if we change the sign of θ_2 , since the cosine does not change sign while the sine does, the same equation, for negative θ_2 , is

$$\cos(\theta_1 - \theta_2) = \cos \theta_1 \cos \theta_2 + \sin \theta_1 \sin \theta_2. \quad (48.3)$$

If we add these two equations together, we see the signs and we learn that the product of two cosines is half the cosine of the sum, plus half the cosine of the difference:

$$\cos \theta_1 \cos \theta_2 = \frac{1}{2} \cos(\theta_1 - \theta_2) + \frac{1}{2} \cos(\theta_1 + \theta_2). \quad (48.4)$$

Now we can also reverse the formula and find a formula for $\cos \theta_1 + \cos \theta_2$ if we simply let $a = \theta_1 + \theta_2$ and $b = \theta_1 - \theta_2$. That is, $a = 2(\theta_1 + \theta_2)$ and $b = 2(\theta_1 - \theta_2)$, so that

$$\cos \theta_1 + \cos \theta_2 = \frac{1}{2} \cos \frac{a}{2} (\theta_1 + \theta_2) + \frac{1}{2} \cos \frac{b}{2} (\theta_1 - \theta_2). \quad (48.5)$$

Now we can analyze our problem. The sum of $\cos \omega_1 t$ and $\cos \omega_2 t$ is

$$\cos \omega_1 t + \cos \omega_2 t = 2 \cos \frac{1}{2}(\omega_1 + \omega_2) t \cos \frac{1}{2}(\omega_1 - \omega_2) t. \quad (48.6)$$

Now let us suppose that the two frequencies are nearly the same, so that $\frac{1}{2}(\omega_1 + \omega_2)$ is the average frequency, and is more or less the same as either ω_1 or ω_2 if ω_1 is much smaller than ω_2 , or ω_2 because, as we suppose, ω_1 and ω_2 are really equal. That means that we can ignore the evolution by saying that there is a high-frequency noise or the more or less like the ones we started with, but that its "size" is slowly changing. Its "size" is pulsating with a frequency which appears to be $\frac{1}{2}(\omega_1 - \omega_2)$. But is this the frequency at which the beats are heard? Although (48.6) says that the amplitude goes as $\cos \frac{1}{2}(\omega_1 - \omega_2) t$, what it is really telling us is that the high-frequency oscillations contained between two opposed cosine curves is shown dotted in Fig. 48-1. On this basis one could say that the amplitude varies at the frequency $\frac{1}{2}(\omega_1 - \omega_2)$, but if we are talking about the intensity of the wave we must think of it as having twice this frequency. That is, the pulsation of the amplitude, in the sense of the strength of its intensity, is at frequency $\omega_1 - \omega_2$, although the formula tells us that we multiply by a cosine wave at half that frequency. The technical basis for this difference is that the high-frequency wave has a little different phase relationship to the second harmonic.

Ignoring this small complication, we may conclude that if we add two waves at frequency ω_1 and ω_2 we will get a resulting wave of average frequency $\frac{1}{2}(\omega_1 + \omega_2)$ which oscillates in strength with a frequency $\omega_1 - \omega_2$.

If the two amplitudes are different, we can do, all over again, by multiplying the cosine by different amplitudes A_1 and A_2 , and do a lot of mathematics, rearranging, and so on, using equations like (48.2)-(48.5). However, there are other, easier ways of doing the same analysis. For example, we know that it is much

easier to work with exponents than with sines and cosines and we can represent it, however, in real part of $A_1 e^{i\omega t}$. The other sum would similarly be the real part of $A_2 e^{i\omega t}$. If we add the two, we get $A_1 e^{i\omega t} + A_2 e^{i\omega t}$. If we then factor out the average frequency, we have

$$A_1 e^{i\omega t} + A_2 e^{i\omega t} = e^{i\omega t}(A_1 \cos(\omega t) + A_2) = A_0 \cos(\omega t - \phi) \quad (45.7)$$

Again we have the high frequency wave with an oscillation of the direct intensity.

45.2 Beat notes and modulation

If we are now asked for the intensity of the wave of (45.7), we can either take the sum of the squares of the left side, or of the right side. That is to say, the left side. The frequency then is:

$$\nu = A_1^2 + A_2^2 + 2A_1 A_2 \cos(\phi) = \nu_0 \nu \quad (45.8)$$

We see that the intensity grows and falls at a frequency $\omega_0 = \pi/\tau$, varying between the limits $(A_1 + A_2)^2$ and $(A_1 - A_2)^2$. If $A_1 > A_2$, the maximum intensity is not zero.

One might try to represent this idea by means of a drawing, like Fig. 45-2. We draw a vector of length A_1 , rotating at a frequency ω_0 , to represent one of the waves in the complex plane. We draw another vector of length A_2 , going around at a frequency ω_0 , to represent the same wave. If the two frequencies are exactly equal, their resultant is of fixed length ν_0 and keeps revolving, and we get a definite fixed intensity from the two. But if the frequencies are slightly different, the two complex vectors go around at different speeds. Fig. 45-3 shows what the situation looks like relative to the moving object. We see that A_1 is moving slowly away from A_2 , and so the amplitude ν_0 is not big, because the two vectors are getting farther apart. As it gets to the left extreme position the resultant ν_0 goes to zero. At the right extreme position the two vectors go around, the amplitude of the sum wave gets exaggerated bigger, and the intensity is a pulsator. It is a relatively simple idea, and there are many different ways to represent the same thing.

The effect is very easy to observe experimentally. In the case of sound, we may arrange two oscillators driven by two separate oscillators, one for each loudspeaker, so that they never make a tone. We can receive one note from one source and a different note from the other source. If we make the frequencies exactly the same, the resulting effect will have a constant strength of a given square location. If we then change them a little bit, we have some variation in the intensity. The further they are detuned, the more rapid are the variations of sound. The ear has been built following analogous more solid rules than we are now aware.

We may also see the effect on an oscilloscope which simply displays the sum of the currents to the two speakers. If the frequency of passing is relatively low, we simply see a sinusoidal wave that wave amplitude pulsates, but as we make the oscillations more rapid we see the kind of wave shown in Fig. 45-3. As we go to greater frequency differences, the "beams" move closer together. Also, if the amplitudes are not equal and we take one signal a range than the other, even we get a wave whose amplitude does not go because zero, just as we expect. Even though it works the way it should both acoustically and electrically.

The opposite phenomenon occurs in radio transmission using so called amplitude modulation (AM). The sound is broadcast by the station as follows: the same transmitter sends an electromagnetic wave which is at a very high frequency, for example 400 kHz/cycles per second, in the broadcast band. If this carrier signal is turned over, it can generate across a wave which is of uniform amplitude of 800,000 times its a second. If it was the "information" is transmitted, the address and of information about what kind of car to buy, is that when someone talks into a microphone, the amplitude of the car in sight is changing then along with the vibrations of sound entering the microphone.

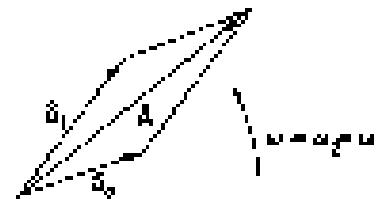


Fig. 45-2. The resultant of two complex vectors of equal frequency.

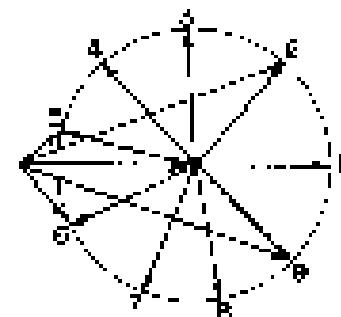


Fig. 45-3. The resultant of two complex vectors of unequal frequency, as seen in the rotating frame of reference of one vector. Note successive positions of the slowly rotating vector A_1 , above.

If we take as the simplest mathematical case the situation where a siren is singing a perfect note, with perfect sinusoidal oscillations of the vocal cords, then we get a signal whose strength is changing as shown in Fig. 4B-4. The instantaneous information is then recovered at the receiver, we get rid of the noise wave and just look at the envelope which represents the oscillations of the vocal cords, or the sound of the singer. For the speaker, then makes an oscillating vibration at the same frequency as the siren, and the listener is then sensitive enough to feel the difference in frequency. Because of a number of formidable and often subtle effects it is in fact impossible to tell whether we are listening to a siren or to a note, except; otherwise the idea is as ridiculous above.

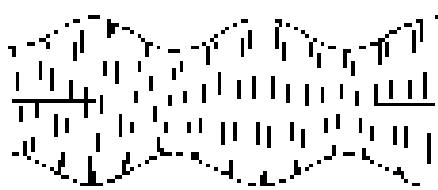


Fig. 4B-4. A modulated carrier wave. In the schematic notation, ω_0/T is the signal frequency, $\omega_0/\omega_m = 100$.

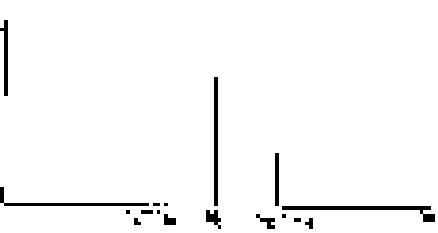


Fig. 4B-5. The frequency spectrum of a carrier wave as modulated by a single side wave ω_m .

4B-3 Side bands.

Mathematically, the modulated wave described above would be expressed as

$$S = C + \text{modulation} \quad (4B-9)$$

where C , represents the frequency of the carrier wave, ω_0 , the frequency of the modulating wave. Again we use all the theorems about the sounds, or we can use $e^{j\theta}$; it makes no difference as to how with ω_0' , but it is the same thing. We can get

$$S = \cos \omega_0 t + j \sin \omega_0 t \omega_m (\omega_0) \quad \text{given } (\omega_0 - \omega_m) \quad (4B-10)$$

From another point of view, we can say that the output wave of the system consists of three waves added in superposition: first, the original wave at the frequency ω_0 , that is, a carrier frequency, and then two new waves at two new frequencies. One is the carrier frequency plus the modulation frequency and the other is the carrier frequency minus the modulation frequency. In therefore, we make some kind of pair of the frequency being generated by the wave due to a function of frequency, as we wanted a low intensity at the frequency of the carrier, naturally, but when a signal starts having we want to have also the intensity proportional to the strength of the carrier, S , at frequencies $\omega_0 + \omega_m$ and $\omega_0 - \omega_m$ as shown in Fig. 4B-5. These are called side bands when there is a modulated signal here, the carrier, or, then, the side bands. If there is more than one tone at two different frequency ω_0 and ω_1 , there are two instances of passing, or if there is any other complicated wave here, then, of course, we can see from the mathematics that we get some more waves that correspond to the frequencies $\omega_0 + \omega_1$, etc.

Therefore, when there is a demodulated amplitude, this can be represented as the sum of many waves. That is, that the detector is measuring out, a range of frequencies, namely the carrier frequency plus or minus the modulation frequency that are modulated signal components.

Although at first we might believe that a radio transmitter transmits only at the one frequency of the carrier, since there are big, super stable crystal oscillators in there, and everything is allowed to be at precisely 500 K cycles, the moment someone measures the radio set, it 3.0 hertzcycles, he modulates the 500 kilocycles, and so they can no longer precisely at 500 kilocycles. Suppose the transmitters are so built that they are able to transmit over a good range of the radio's sensitivity (that is, from up to 20,000 cycles per second, but usually radio transmitters and receivers do not work very at 10,000), so we do not have the highest part, that is, when the main signals are within very narrow frequencies running up, etc., to 10,000 cycles, so the transmitted is modulating frequencies ...

* A different remark: In what circumstances can a curve be approximated by a lot of straight segments. In all ordinary circumstances except for certain cases the mathematics are difficult. Of course, if you must have only one value at a given point, you may not be able to draw a curve which joins an infinite number of lines, or informed curves, or something like that. But inside them such continuous approximation can be done that a siren is going to be able to make by sketching on one family of curves or compounded by adding others more together.

which may range from 790 to 810 megacycles per second. Now if we have another station at 795 because there would be a lot of interference. And, if we make our receiver so sensitive that it picks up only 800, one did not pick up the 795 megacycles on either side, we would not hear what the man was saying, because the information would be on the side which frequencies. Therefore it is absolutely essential to stop the picture at certain intermediate points, so that the side bands do not overlap and, also, the receiver must not be so selective that it does not permit reception of the side bands as well as of the radio central frequency. In the case of sound, the problem does not really come much into play. We can hear over a 20 kilocycle range, and we have usually from 300 to 1000 kilocycles in the broadcast band, so there is plenty of room for lots of stations.

The reception problem is more difficult. As the electron beam goes across the face of the picture tube, there are various little spots of light and dark. Let "light" and "dark" be the "signal." Now suddenly the beam scans over the whole picture. Suppose, approximately, in a vertical column. Let us assume that the resolution of the picture vertically and horizontally is more or less the same, so that there are the same number of dots per inch along a scan line. We want to be able to distinguish dark from light, dark from light, dark from light, etc., 300 lines. In order to be able to do this with colors we have the choice a wavelength needed that corresponds to a wavelength, from ~~maximum to minimum~~, of one 750th of the green wave. So we have $300 \times 300 \times 30$ pieces of information per second. The highest frequency that we are going to carry, therefore, is thus: 4 megacycles per second. Actually, to keep the television stations apart, we have to use 3 bits of information from this, a total of 12 bits; part of it is used to carry the sound signal. And other information. So, resolution demands are 6 megacycles per second with. It certainly would not be possible to transmit TV on the 400 kilocycles center, since we cannot measure at a higher frequency than the carrier.

At any rate, the television wave starts at 4 megacycles. The first transmission channel, which is channel 2 (U), has a frequency range from 51 to 60 megacycles, which is 9 megacycles wide. "But," one might say, "we have just proved that there were side bands on both sides, and therefore it is 18 megacycles wide." I think not. Just the radio engineers are rather clever. If we realize the modulation signal using only sine waves, but having nothing to do with the physical mechanics, we can see that there is a definite, invariant relationship between the side band on the high-frequency side and the side band on the low-frequency side. What we prove is that there is no new information on that other side band. So what is done is to suppress one side band, and the receiver is fitted with that the information which is missing is reconstructed by looking at the single side band and the carrier. Single-side-band transmission is a clever scheme for decreasing the band width needed to transmit information.

4.4 Identified wave forms

The next subject we'd like discuss is the waveforms of waves in both space and time. Suppose that we have two waves travelling in space. We know, of course, that we can represent a wave travelling in space by $e^{i(kx-\omega t)}$. This might be, for example, the displacement in a sound wave. This is a solution of the wave equation provided $\omega^2 = k^2 c^2$, where c is the speed of propagation of the wave. In this case we can write it as $e^{i(kx-\omega t)}$, modulus of the general form $f(x - ct)$. Therefore this must be a wave which is travelling at this velocity, with, and that is all right.

Now we want to add two such waves together. Suppose we have a wave that is travelling with one frequency, and another wave travelling with another frequency. We have to remember to consider the case where the amplitudes are different; it makes no real difference. Let us write it as $A_1 e^{i(k_1 x - \omega_1 t)}$. We can add this by the same kind of mathematics we used when we added signal waves. Of course, if ω_1 is the same as ω_2 , this is easy, since it is the sum of what we had before:

$$e^{i(k_1 x - \omega_1 t)} + e^{i(k_2 x - \omega_2 t)} = e^{ik_1 x} [e^{i\omega_1 t} + e^{i\omega_2 t}] \quad (15.11)$$

except that $\phi = \omega_0 t$ is the variable instead of t . So we get the same kind of wavefunctions, naturally, but we see, of course, that these modulations are moving along $\rho \neq 0$. In other words, if we added two waves, but these waves were not just oscillating but a moving in space, then the resultant wave would move along also, at the same speed.

Now we would like to generalize this to the case of waves in which the relationship between the frequency and the wave number is not so simple. For example, consider using an index of refraction. We have already studied the theory of the index of refraction in Chapter 11, where we found that we could write $k = n\omega/c$, where n is the index of refraction. As an interesting example, let us say we know that the index is

$$n = 1 + \frac{\partial \phi^2}{2c_0 \omega^2}. \quad (48.10)$$

We actually derived a more complicated formula in Chapter 11, but this one is good enough for our example.

Incidentally, we note that even when n and k are not linearly proportional, the ratio n/k is constant, the speed of propagation for the particular frequency and wave number. When it is the speed of propagation, the speed at which the phase, or the nodes of a sinusoidal wave, would move. Let us

$$v_p = \frac{\omega}{k}. \quad (48.11)$$

This phase velocity, for the case of waves in glass, is given by the speed of light in vacuum ($c = 3 \times 10^8$ cm/sec) times n (as from Eq. 1), and that is a bit higher, because we do not think we can send speeds faster than the speed of light.

What we are going to determine is the interference of two waves in which ω_1 and ω_2 have a definite formula relating them. Use above formula for v_p , Eq. 1, is given as a definite function of ω . This quantity in this particular problem—the formula for k in terms of ω is

$$k = \frac{\omega}{c} - \frac{n}{v_p}, \quad (48.12)$$

where $n = \omega_0^2/(\omega_0 + \omega_1 + \omega_2)$. At any rate, for each frequency there is a definite wave number, and we want to add two such waves together.

Let us do it just as we did in Eq. (48.9);

$$\begin{aligned} e^{i(\omega_1 t + k_1 x)} &= e^{i(\omega_0 t + k_0 x)} + e^{i(\omega_0 t + k_0 x + \omega_1 t + \omega_1 x)} \\ &\times \{e^{i(\omega_0 t + \omega_1 t + k_0 x + k_1 x)} - e^{i(\omega_0 t + \omega_1 t + k_0 x + k_1 x + 2\pi)}\}. \end{aligned} \quad (48.13)$$

So we have a modulated wave again, a wave which travels with the mean frequency and the mean wave number, but whose k and ω is varying with x in which the amplitude of the difference frequency is $\omega_1 = \omega_1 - \omega_0$ wave number.

Now let us make the case that the difference between the two waves is relatively small. Let us suppose that we are adding ω_1 to ω_0 where ω_0 and ω_1 are nearly equal; then $(\omega_1 + \omega_0)/2$ is probably the same as either one of the ω 's, but certainly for $(\omega_1 + \omega_0)/2$. Thus the speed of the wave, the "group velocity," the index is not necessarily n . But back to the point of propagation where modulation is not the same! How to reduce we have to change k to account for a certain constant of ω . The speed of this modulation wave is the rule:

$$v_g = \frac{\omega_1 - \omega_0}{k_1 - k_0}. \quad (48.14)$$

The speed of modulation is sometimes called the group velocity. If we take the case that the difference in frequency is relatively small, and the difference in wave number is also relatively small, then this expression approaches us, in the limit,

$$v_g = \frac{\omega_1}{k_1}. \quad (48.15)$$

In other words, for one source, modulation, the signal beats, there is a definite speed at which they travel which is not the same as the phase speed of the waves—what a mysterious thing!

The group velocity is the derivative of ω with respect to k , and the phase velocity is v/c .

Let us see if we can understand why. Consider two waves, again of slightly different wavelength λ , as in Fig. 45-1. They are out of phase, $\pi/2$ phase, out of phase, and so on. Now these waves represent, really, the waves in space travelling with slightly different frequencies also. Now because the phase velocity, the velocity of the crests of these two waves, is not necessarily the same, something new happens. Suppose we ride along with one of the waves and look at the other one; if they both travel at the same speed, then the other wave would stay right where it was relative to us, as we ride along on this wave. We find, in fact, that now and right opposite us we see a beat. If the two velocities are equal the crests stay on top of each other but if it is not we find the two velocities are really equal. There is only a small difference in frequency and therefore only a $\sim \frac{1}{\lambda}$ difference in velocity, but because of that difference in velocity, as we ride along the other wave moves slightly forward, say, or nothing, relative to our wave. So as in a guitar, what happens? If we move one wave just a shade forward, the wave travels forward (or backward) considerably distance. That is, the sum of these two waves has an envelope, and as the waves travel along the envelope rises or falls at a different speed. The group velocity is the speed at which modulated signals would be transmitted.

If we move a signal, i.e., some kind of charge in the wave that one could recognize when he listened to it, a kind of modulation, then this would then travel at the group velocity, provided that the amplitudes were relatively small. (When they are large, it is much more difficult to analyze.)

Now we may show (not long ago), that the speed of propagation of γ -rays in a block of carbon is very greater than the speed of light, although the phase velocity is greater than the speed of light. In order to do that we must find $d\omega/dk$, which we get by differentiating (45.14) $d\omega/dk = v/c - n\omega/c$. The group velocity, therefore, is the reciprocal of this quantity,

$$v_g = \frac{c}{1 + \omega^2 n^2} . \quad (45.18)$$

which is smaller than c . So although the phases are travel faster than the speed of light, the modulating signals travel slower, and that is the result of the apparent paradox. Of course, if we have the simple case that $n = 1$, then v_g will be also c . So when all the photons have the same velocity, certainly the group has the same velocity.

46-5 Probability amplitudes for particles

Let us now consider one more example of the phase velocity which is extremely interesting. It has to do with quantum mechanics. We know that the amplitude to find a particle at a place and, in some circumstances, vary in space and time, let us say in one dimension, in the manner:

$$\psi = A e^{i(kx-\omega t)} , \quad (46.19)$$

where ω is the frequency, which is related to the charge, e , of the charge through $E = Pv$, and k is the wave number, which is related to the momentum through $p = \hbar k$. We should say that you could not find a definite momentum p if the wave number were exactly k , but it is k plus more which goes on with the same amplitude everywhere. Equation (46.19) gives the amplitude, and if we take the absolute square, we get the relative probability for finding the particle as a function of position and time. This is a constant, which means that the probability is always the same a particle anywhere. Now suppose, instead, that we have a situation where we know that the particle is more likely to be at one place than at another. We would

represented more accurately by a wave which has a maximum and decays on either side (Fig. 48-6). It is more realistic than a wave-like (48.7) which has a series of maxima, but it is possible, by adding several waves of nearly the same ω and k together, to get rid of all but one maximum.

Now at different instants, since the square of (48.19), m^2/c^2 , is the square of having a particle somewhere, we know that at a given instant the particle is most likely to be near the center of the "lump," where the amplitude of the wave is maximum. If we wait a few moments, the waves will interact, and after some time the "lump" will reseparate like Fig. 48.7. If we know that the particle originally was scattered somewhere, obviously we can expect that it would later be elsewhere as a matter of fact, because it has moved, after all, with some velocity. The quantum theory, then, will go into the correct classical theory for the relationship of momentum, energy, and velocity only if the group velocity, the velocity of propagation, is equal to the velocity that the particle drifts classically for a particle of the same momentum.

It is very necessary, when discussing this subject, to note the case. According to the classical theory, the energy is related to the velocity through an equation like

$$E = \frac{mv^2}{\sqrt{1 - v^2/c^2}}. \quad (48.19)$$

Similarly, the momentum is

$$p = \frac{mv}{\sqrt{1 - v^2/c^2}}. \quad (48.20)$$

To get the classical theory, one does a very simple part of the classical theory by eliminating v ; we can show that

$$E^2 - p^2c^2 = m^2c^4.$$

This is the four-dimensional general result that we have talked and talked about, that $p_{\mu}p^{\mu} = m^2$; that is the relation between energy and momentum in the classical theory. Note that again, since (48.19) is the group velocity, we must add E^2 , by substitution, $E^2 = E^2 + p^2c^2$. So that for quantum mechanics it is no mystery that

$$\frac{E^2c^2}{c^2} = p^2c^2 + m^2c^2. \quad (48.21)$$

This then is the relationship between the frequency and the wave number of a quantized longitudinal amplitude wave representing a particle of mass m . From this equation we can deduce that ω is

$$\omega = c\sqrt{k^2 + m^2/c^2}.$$

The group velocity, $c\sqrt{k^2}$, is necessarily larger than the speed of light.

Now let us look at the group velocity. The group velocity equals $c\sqrt{k^2}$, or the speed to which the modulations move. We have to differentiate a square root, *i.e.*, a square root times a square root, which is not very difficult. The derivative is

$$\frac{dk}{dt} = \frac{\partial k}{\partial t} = \frac{\partial k}{\partial \omega} \frac{\partial \omega}{\partial t} = \frac{\partial k}{\partial \omega} \frac{\partial}{\partial t} \left(\sqrt{\omega^2 - p^2c^2} \right).$$

Now the square root is, after all, ω , so we could write this as $d\omega/dt = \pm k/m$. In other words, ω/E is

$$\omega = \frac{ck}{E}.$$

In both (48.20) and (48.21), $c\sqrt{k^2}$ is the velocity of the particle, according to classical mechanics. So we see that whereas the fundamental quantum-mechanical relationship $E = h\nu$ and $p = \hbar k$, for the combination of ω and k with the classical E and p , only produces the equation $\omega^2 - p^2c^2 = m^2/c^2$, now we also understand the relationships (48.20) and (48.21) which connect E and ω to k .

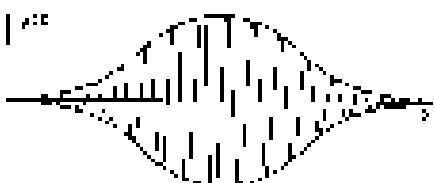


Fig. 48-6. A localized wave train.

to the velocity. Of course the group velocity must be the value v_g of the particle if the interpretation is going to make sense. If we think the particle is moving at constant speed, and then the particle has to think it is over there, as the quantum mechanics said, the distance traversed by the "lump," caused by the disturbance, can't be classically the velocity of the particle.

48.6 Waves in three dimensions

We start now in our discussion of waves to a third with a few general remarks about the wave equation. These remarks are intended to give some idea of the future, and that we can understand something much later on, but rather to see what things are going to look like when we get to waves in three dimensions. Fig. 48.17, the wave equation for waves in one dimension was

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2},$$

where c is interpreted as above, the wave to run in the case of sound, is the sound speed, in the case of light, is the speed of light. We showed that for a forced wave the displacement would propagate through space at a certain speed. But the excess pressure also propagates at a certain speed and so does the excess density. So we should expect that the pressure would satisfy the same equation, or indeed it does. We shall revert to the basic wave picture again. First, ρ is proportional to the rate of change of ψ with respect to x . Therefore if we differentiate the wave equation with respect to x , we will immediately discover that $\delta\rho/\delta x$ satisfies the same equation. That is, we can multiply the same equation. If ψ is proportional to ρ , and therefore $\delta\rho/\delta x$ does too. So the pressure, the displacement, everything, satisfy the same wave equation.

Usually one sees the wave equation for sound written in terms of pressure instead of in terms of displacement. This is because pressure is a scalar and has no direction. But the displacement is a vector and has direction, and it is easier to analyze the pressure.

The next problem we discuss is to do with the wave equation in three dimensions. We know that the sound wave solution in one dimension is $e^{i(kx-\omega t)}$, and $\omega = kc_n$, but we also know that in three dimensions a wave would be represented by $e^{i(\sqrt{k^2 + \omega^2}x - \omega t)}$, where, in this case, $k^2 = k_x^2$, which is, of course, $(\partial/\partial x)^2 + (\partial/\partial y)^2 + (\partial/\partial z)^2$. Now what we want to do is to guess what the correct wave equation in three dimensions is. Naively, for the case of sound this can be obtained by going through the same dynamic argument in three dimensions that we made in one dimension. But we shall not do that; instead we just write down what comes out, the equations for the pressure (or displacement, ψ , according) x

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{\partial^2 \rho}{\partial x^2} = \frac{\partial^2 P}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2}. \quad (48.23)$$

That this is correct can be verified by differentiating $e^{i(\sqrt{k^2 + \omega^2}x - \omega t)}$. Clearly, every time we differentiate with respect to x , we multiply by $-ik_x$. If we differentiate twice, we require first to multiply by $-k_x^2$, so the first term would become $-k_x^2 e^{i(\sqrt{k^2 + \omega^2}x - \omega t)}$; the second term becomes $-k_x^2 e^{i(\sqrt{k^2 + \omega^2}x - \omega t)}$, and the third term becomes $-k_x^2 e^{i(\sqrt{k^2 + \omega^2}x - \omega t)}$. On the right, we get $-(\omega^2/c^2)e^{i(\sqrt{k^2 + \omega^2}x - \omega t)}$. Then, if we take away the first and change the sign, we get all the relation, and between variables is however it is wanted.

Moving to quantum theory, we can start writing down the general equation which corresponds to the dispersion equation (48.12) for quantized mechanical waves. If ψ represents the amplitude for finding a particle's position x, y, z , at the time t , then the great equation $\psi_{n,k}$ is given again, its for 2D particles is this

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} = \frac{\omega^2 \psi}{c^2} - \frac{m^2 \psi^2}{E^2} x. \quad (48.24)$$

In other words, the only difference with this expression is suppressed by the appearance

of x , y , z and t of the nice combination continuity wave improves. Second, it's a wave equation which, if we try a plane wave, would produce an expression that $-k^2 + \omega_0^2 k^2 = m^2 c^2 / N$, which is the right relationship for quantum mechanics. There is still another great thing contained in the wave equation: the fact that any superposition of waves is also a solution. So this equation contains all of the quantum mechanics and the relativity that we have been discussing so far, at least so long as it deals with a single particle in empty space with no external potentials in there at all.

46-7 Normal modes

Now let's do another example of the phenomenon of beats, which is rather curious and a little different. Imagine two equal pendulums which have, between them, a rather weak spring connection. They are made as nearly as possible the same length. If we put one with some energy it moves back and forth, and it pulls on the connecting spring so it moves back and forth, and even more is a machine for generating a force which has the natural frequency of the other pendulum. Therefore, is a consequence of the theory of resonance, which we studied before: when we put a force of something at just the right frequency, it will drive it. So, sure enough, one pendulum moving back and forth drives the other. However, is this the circumstance? This is a new thing *suppose*, because the total energy of the system is finite, so when one pendulum takes its energy out of the other to drive it, it finds itself gradually losing energy until, if the driving is just right along with the speed, it loses all its energy and is reduced to a stationary condition. There, of course, is the other pendulum but that has all the energy and the first one which has more and as time goes on we can see it works also in the opposite direction, until finally the energy is passed back into the first self: this is a very interesting and amusing phenomenon. We said, however, for this is related to the theory of beats, and we must now explain how we can analyze this motion from the point of view of the theory of beats.

We note that the motion of either of the two bobs is an oscillation which has an amplitude which changes steadily. Therefore the motion of one of the bobs is approximately analytic in a different way, in that it is the sum of two oscillations, present at the same time but having two slightly different frequencies. The idea is again to be possible to find two other motions in this system and to claim that what we saw was a superposition of the two motions, because this is of course a linear system. Indeed, it is easy to find two ways that we could start the motion, each one of which is a perfect single-frequency motion—absolutely periodic. The motion that we started with before was not strictly periodic, since it did not last, since one bob was passing energy to the other and so it's going to oscillate but there are ways of starting the motion so that nothing happens and, of course, as soon as we see it we understand why. For example, if we made both pendulums go together, then, since they are of the same length and the spring is not carrying anything, they will of course continue to swing like that for all time, according to Hooke and that everything is perfect. On the other hand, there is another possible motion which also has a definite frequency, that is, if we made the pendulums approach, pulling them with exactly equal distances, then again they would be in absolutely periodic motion. We can imagine that the spring just adds little to the restoring force that the spring supplies the overall and the system just keeps oscillating at a slightly higher frequency than in the first case. Why? Because the spring is pulling, in addition to the gravitation, and it makes the system a little "tighter," so that the frequency of this motion is just a shade higher than the other.

Thus this system has two ways in which it can oscillate with undamping amplitude: it can either oscillate in a manner in which both pendulums go the same way and neither all the time at one frequency, or they could go in opposite directions at a slightly higher frequency.

Now the actual motion of the ring, because the system is linear, can be represented as a superposition of the two. (This is left as an exercise, remember, at 1b.)

(- the effect of adding two motions with different frequencies.) So think what would happen if we considered these two motions. First - if the two motions were to start with equal amplitude and in the same phase, the sum of the two motions means that one half, having been increased one way by the first motion and the other way by the second motion, is zero while the other half, having been displaced the same way in both motions, has a large amplitude. As time goes on, however, the two halves remain "independent," and a plateau like the top of the wave is slowly shifting. Then comes, even though a sufficiently long time, when the time is enough that one motion could have gone "180°" oscillations, while the other were only "90°." The relative phase would be just reversed with respect to what it was before. That is, the large amplitude section will have fallen to zero and in the meantime, of course, the initially stationary half will have acquired full strength.

So we see that we could analyze this complicated motion either by the law that there is a transient and that one gains energy at the other, or else by the superposition of two counter-amplitude motions at two coherent frequencies.

Waves

49-1 The reflection of waves

This chapter will consider some of the remarkable phenomena which may result in conflicting waves in some finite regions. We will be led to the subject by particular facts about vibrating strings, for example, and then the generalization of these facts will give us a principle which is probably the most far-reaching selectable of mechanics and physics.

Our first example of conflicting waves will be to consider waves at one boundary. Let's take the simple example of a one-dimensional wave on a string. One could equally well consider waves in one dimension against a wall, or the situation is similar nature, but the example of a string will be sufficient for our present purposes. Suppose the string is held at one end, for example by fastening it in an "indefinitely solid" wall. Then we can visualise mathematically by saying that the displacement y of the string at the position $x = 0$ must be zero, because the end does not move. Now if it were not for the wall, we know that the general solution for the string is the sum of two vibrations, $y(t) = y_1 + y_2$, the first representing a wave travelling to the left in the string, and the second a wave travelling the other way in the string:

$$y = F(x - ct) + G(x + ct) \quad (49-1)$$

is the general solution for any string. If we have now to satisfy the condition that the string does not move at one end, if we put $x = 0$ in Eq. (49-1), and consequently setting y to zero of t we get $y = F(0 - ct) + G(0 + ct)$. Now this is to be zero for all time, it means that the function y must be $-F(-ct) - G(ct)$. In other words, if we multiply y by $e^{i\omega t}$, F and G are the same thing. If this result is put back into Eq. (49-1), we find that the solution for the problem is

$$y = F(x - ct) - F(x + ct). \quad (49-2)$$

It is easy to check that we still get $y = 0$ if we set $x = 0$.

Fig. 49-1 shows a wave travelling to the right at speed c near $x = 0$, and a hypothetical wave travelling in the other direction, also set in motion on the other side of the origin. The string is fixed at the side of the origin. The total motion of the string is to be regarded as the sum of these two waves in the region of zero x . As they reach the origin, they will always meet at $x = 0$, and truly *cancel each other* (destructively interfere) by the time one has gone positive x and it will, of course, be travelling in the negative x region. These results are analogous to the following statement: if a wave meets a rigid wall and reflects off, it will be reflected with a change in sign. Since a reflection can always be understood by inverting that part of the motion of the string comes out upside down from behind the wall. In short, if we assume that the string is infinite and that we have one wave going one way we have another one going the other way with the same amplitude, the displacement at $x = 0$ will always be zero and it would make no difference if we compact the string there.

The next point to be discussed is the reflection of a periodic wave. Suppose that the wave represented by $F(x - ct)$ is a sine wave and has been reflected; then the reflected wave $-F(x + ct)$ is also a sine wave of the same frequency, but travelling in the opposite direction. This requires F to have simple exponential behaviour, since the complex function relation: $F(x - ct) = e^{i\omega(x-ct)}$ and

49-2 The reflection of waves

49-2-1 Confined waves, with natural frequency

49-2-2 Modes in two dimensions

49-4 Coupled pendulums

49-5 Linear systems

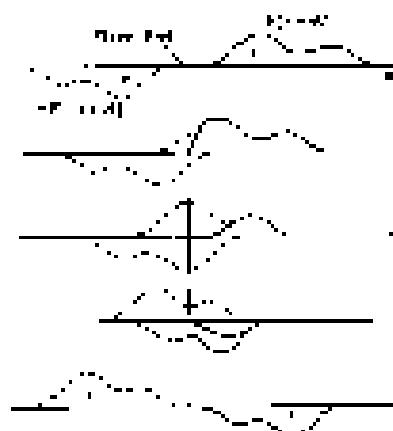


Fig. 49-1. Reflection of a wave and superposition of two travelling waves.

$\delta t = \pi - \alpha$ — *condition*. It can be seen that if these two values are in (49.2) and Δx is set equal to 0, then $\alpha = 0$ for all $\omega > \omega_0$, so it satisfies the *second condition*. Because of the properties of exponentials, this can be written in a simpler form:

$$x = e^{i\omega_0 t} (e^{-i\omega_0 t} + e^{i\omega_0 t}) = e^{i\omega_0 t} \sin(\omega_0 t/2) \quad (49.3)$$

There is something interesting and new here in the displacement that we look at any fixed x on the string's rest frequency. No matter where this point is, the frequency is the same! But does the same place, in particular whatever site $\Delta x = 0$, where there is no displacement \Rightarrow all? Furthermore, if at any time we take a snapshot of the string's string, the picture will be like waves. However, the displacement of this wave x will depend upon the time t . From inspection of Eq. (49.3) we can see that the length of one cycle of the sine wave is equal to the wavelength of either of the superimposed waves.

$$\lambda = 2\pi/\omega_0. \quad (49.4)$$

The points where there is no motion satisfy the condition $\sin(\omega_0 t) = 0$, which means that $\omega_0 t = 0, \pi, 2\pi, \dots, m\pi$. These points are called nodes. Between any two successive nodes, every point moves up and down sinusoidally, but the position of center stays fixed \pm nodes. This is the characteristic of standing waves on a string. One can find a pattern of nodes where are the present. Just at any point the object moves perfectly sinusoidally, and that all points move at the same frequency (though some will move more than others), when we have what is called a mode.

49-2 Standing waves, Mathematical Considerations

The next interesting problem is to consider what happens if the string is held at different, say sites. Assume $\alpha \neq 0$. We can begin with the idea of the reflection of waves coming from some kind of a boundary in one direction. As time goes on, we would expect the string to go back and forth, and as time goes still farther it will become a kind of little wobba, because it is combining with the reversed image wave which is coming from the other side. Finally the string's wavy will disappear and the string's energy will move in the other direction to repeat the process in the other end. This problem has an easy solution, but an interesting question is whether we can have continuous standing waves. Theory of diffraction is possible, but of course it is not necessarily periodic. Let us try to get a sinusoidally periodic wave on a string. If the string is tied at one end, we know it cannot look like our corner solution (49.3). If it is tied in the other end, it has to look like sum of the other end. So the only possibility for periodic waves is when the string is tied at both ends. When the string is tied at both ends, the only possibility is that $\sin(\omega_0 t) = 0$, however this is the only condition that will keep the ends fixed. Now to order the strings to be zero, the angle must be either $0, \pi/2, \pi$, or some other integer, k multiple of π . The condition

$$\delta t = \pi \quad (49.5)$$

will therefore give only one of the possible λ 's, depending on which integer is put in. For each of the λ 's there is a certain frequency ω which according to (49.3) is simple:

$$\omega = \Delta x \pi \omega_0 / \lambda. \quad (49.6)$$

We have found the following: that a string has a property that it can have sinusoidal motions, say with constant frequencies. This is the most important characteristic of confined waves. No matter how complicated the system is, it always turns out that there are some patterns of motion which have a perfect sinusoidal time dependence, but with frequencies that are a property of the particular system and the nature of boundaries. In the case of the string we have many different possible frequencies, such as, by definition, corresponding to a mode, because a mode is a pattern of motion which repeats itself sinusoidally. Figure 49-2 shows the first three modes for a string. For the first mode the wavelength is $\lambda = 2L$. This must be zero if one continues the wave out to $x = L$ to obtain one complete cycle of the sine wave. The angular frequency ω is then divided by the wavelength, in general, one in this case, exactly in 2π . The frequency is $\omega/2\pi$, which is in agreement with (49.2) with $v = 1$. Let us call the first mode frequency ω_1 . Now the next mode has two loops in the middle. For this mode the wavelength, then, is simply L . The corresponding value of ω is twice as great and the frequency is twice as large, $2\omega_1$. For the third mode, it is $3\omega_1$, and so on. So all the different frequencies of the string are multiples, 1, 2, 3, 4, and so on, of the lowest frequency ω_1 .

Returning now to the general method of the string, let us consider any possible motion can always be analyzed by adding the more than harmonic in operating at the same time. In fact, for general motion the actual number of modes must be excited at the same time. To get some idea of this, let us illustrate what happens when there are two modes oscillating at the same time. Suppose that we have the first mode oscillating as shown by the sequence of pictures of Fig. 49-3, which illustrates the deflection of the string for equally spaced time intervals, starting through one cycle of the lowest frequency.

Now, at the same time, we suppose that there is an oscillation of the second mode also. Figure 49-3 also shows a sequence of pictures of this mode, which at the start is 90° out of phase with the first mode. This means just or the start it has no displacement, but the two halves of the string have oppositely directed velocities. Now we recall a general principle relating to linear systems: if there are only two solutions, then their sum is also a solution. Therefore a third possible motion of the string would be a displacement which will be adding the two solutions shown in Fig. 49-3. The result, also shown in the figure, begins to suggest the idea of a bump running back and forth between the ends of the string, although with only two modes we cannot make a very good picture of its motion unless we extend. The result is, in fact, a special case of a great principle for waves generally.

Any motion at all can be analyzed by adding them it is the "principle of superposition" of all the different modes, combined with appropriate amplitudes and phases.

The importance of this principle derives from the fact that each mode is very simple—it is nothing but a sinusoidal function of time. It is true that now, the general motion of the string is not really very complicated, but there are other systems, for example the whipping of an airplane wing, in which the motion is a lot more complicated. Nevertheless, even with an airplane wing, we find cases in a certain particular way of twirling which has one frequency, and other ways of twirling may have other frequencies. If these modes can be found, then the complete motion can always be analyzed as a superposition of harmonic oscillations (except when the whipping is of such degree that the system can no longer be considered as linear).

49-3 Modes in one dimension

The best example to be considered is the interesting situation of modes in two dimensions. Up to this point we have talked only about one-dimensional situations—a stretched string or spring moving in a tube. Let's try to think of two dimensions, but an easier step will be due to two dimensions. Consider for definite a rectangular elastic membrane which is confined to a box; that is, it displaces only anywhere on the two angular edges, and let the dimensions of the rectangle

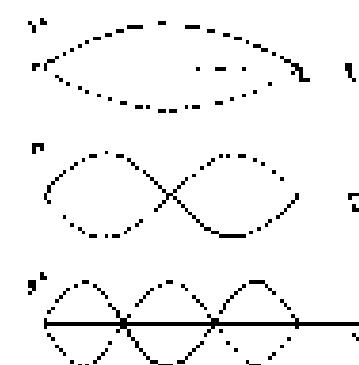


Fig. 49-2. The first three modes of a vibrating string

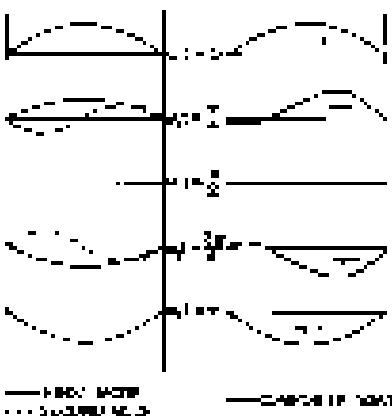


Fig. 49-3. Two modes combine to give a travelling wave.

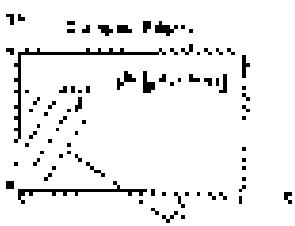


Fig. 19-4. Vibrating rectangular plate.

harmonic, as shown in Fig. 19-4. Now the question is, what are the characteristics of the possible modes? We consider as before the procedure easier for the x -mode. If we had no information at all, we would expect waves traveling along all paths that fit our mirror. For example, $(x^2)(y^2)(e^{i\omega t})$ would represent a wave traveling in some direction which depends on the relative values of x and y , and also has axes parallel to the x -axis, that is, the $\partial u/\partial y = 0$, $\partial v/\partial y = 0$. Using the ideas developed for the one-dimensional string, we can imagine another wave represented by the complex function $(-x^2)(y^2)(e^{i\omega t})$. The superposition of these waves will give us a displacement at $y = 0$ equal to the values of u and v (although these functions are defined for $y \neq 0$, let y where there is no displacement to $y = 0$, this can be ignored, since the displacement is only zero at $y = 0$). In this case we can look upon the second function as the reflected wave.

However, we want a nodal line at $y = b$ as well as at $y = 0$. How do we do this? The solution is similar to scattering, we did when studying reflection from crystals. These waves reflect off either at $y = 0$ or at $y = b$, only if the surface index of refraction is the same at $y = b$ (although it is not necessarily the same at $y = 0$), as shown in Fig. 19-5:

$$n_0 = 2n \sin \theta, \quad n = n_0 / k, \quad (19.7)$$

Now in the same way we can take the previous procedure by adding two more functions $-(x^2)(y^2)(e^{i\omega t})$ and $-(y^2)(y^2)(e^{i\omega t})$, each representing a reflection of one of the other two waves from the $y = b$ line. The condition for a nodal line at $y = b$ is similar to the one for $y = 0$. It is that $2k$ and $n_0 k$ be an integral multiple of π :

$$\pi n = 2k(b - a) \quad (19.8)$$

To conclude I recall that the waves forming a nodal line produce a standing-wave pattern, that is, a definite mode.

So we must satisfy the above two conditions if we want to have a mode. Let us find the wavelength. This can be done by eliminating the angle θ from (19.7) and (19.8) to obtain the wavelength in terms of a , b , n and ω . To easiest way to do this is to divide both sides of the relevant equations by $2k$ and $2n$, multiply them, and add the two equations together. The result is $\sin^2 \theta + \cos^2 \theta = 1 = (a/b)^2 - (n_0/n)^2$, which can be solved for λ :

$$\lambda = \frac{a^2}{\sqrt{1 + (n_0/n)^2}} + \frac{b^2}{\sqrt{1 + (n_0/n)^2}}. \quad (19.9)$$

In this way we have determined the wavelength in terms of the edges, and from the wavelength we immediately get the frequency ω . However, as we know, the frequency is equal to $2\pi f$ divided by the wavelength.

This result is interesting, and important enough that we should discuss it by a purely mathematical analysis instead of by an argument about the reflections. Let me repeat: we vibrated by a superposition of two waves chosen so that the four lines $x = 0$, $x = a$, $y = 0$, and $y = b$ are all nodes. In addition we shall require that all waves have the same frequency, and if f is the resulting motion with respect to a mode, f will be called the sum of light reflection; we know that $(y^2)(y^2)(e^{i\omega t})$ represents a wave traveling in the direction $\tan^{-1}(\omega/k) = \tan^{-1}(a/b)$. Equations (19.6) and (19.7) tell us $k = \omega/c$, and this provides

$$\lambda^2 = a^2 + b^2. \quad (19.10)$$

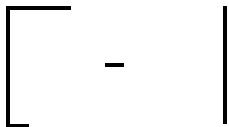
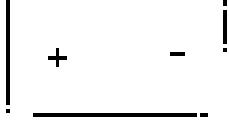
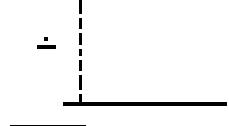
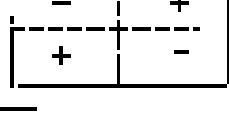
It is clear from this that $k_x = k$ and $k_y = k$.

Now our equation for the displacement, say u , of the rectangular membrane takes on the general form

$$u = [e^{i\omega t} V^{(1)} e^{i k_x x} e^{i k_y y} + e^{i\omega t} V^{(2)} e^{-i k_x x} e^{i k_y y} - e^{i\omega t} V^{(3)} e^{i k_x x} e^{-i k_y y} - e^{i\omega t} V^{(4)} e^{-i k_x x} e^{-i k_y y}], \quad (19.11)$$

Although this looks rather a mess, the sum of these things now is not very hard. D-4

Table 48-1

Mode shape	m	n	$\omega_{n,\text{expt}}$	$\omega_{n,\text{cal}}$
	1	1	1.9	1.82
	1	2	2.00	1.41
	1	1	1.9	.92
	2	1	1.74	-1.4
	2	2	3.00	2.24

The ω_n constants can be combined to give the functions $\psi_m(x)$ of the displacement form and $\psi_n(y)$:

$$\psi_n = (-1)^m \sin k_n x \sinh k_n y \quad (48-13)$$

In other words, it is a sinusoidal excitation, all right, with a position func. is also sinusoidal in both the x - and the y -direction. Our boundary conditions are of course satisfied at $x = 0$ and $y = 0$. You can easily verify as we did at $x = a$ and when $y = b$. Therefore we have to satisfy two other conditions: ψ_m must be an integer multiple of π , and ψ_n must be another integer k_n multiple of π . (Remember we have seen that $k_n = \lambda_n / a$; see 3 and $\lambda_n = 2\pi n$.) We immediately get equations (48-7) and (48-8) and from these the final result (48-9).

Now let us take as an example a rectangle whose width is twice the height. If we take $a = 2b$ and use Eqs. (48-7) and (48-8), we get the frequencies of all of the modes:

$$\omega_n^2 = \left(\frac{\pi n}{b}\right)^2 \frac{4\pi^2}{3} + m^2 \quad (48-14)$$

Table 48-1 gives some of the simple modes and can show other shapes in qualitative form.

The most important point to be emphasized about this part is the note "Ex.: the frequencies are not multiples of each other," or are they multiples of any number. The idea that the natural frequencies are harmonically related is not generally true. This is true for a system of one dimension, *i.e.* if it has no mass-dissipative systems which are more complicated than a string with uniform density and tension. A simple example of the latter is a hanging string; its natural frequencies are proportional to the length. It is also true if, in its mode oscillation, there are various masses and "springs," but the frequencies are not simple multiples of one another, nor are the mass-springs constant.

The modes of more complicated systems are *not* necessarily harmonic. For example, make the moves we have a long y above the x -axis, and by moving the x -axis and the y -axis, and so forth, we make an open-end pipe in a closed tube; types of different diameters and shapes; it is a terribly complicated oscillator, but

it is evident nevertheless. Now when one talks with gut vocal cords, they are made to produce some kind of tone. The tone is rather complicated and there are many sounds coming out, but the cavity of the mouth further modifies that tone because of the various resonant frequencies of the cavity. For instance, a singer can sing various vowels, e., i, oo or ee, and so forth, in the same pitch, but they sound different, because the various conditions are in resonance at this cavity to different degrees. The very great importance of the resonant frequencies of a cavity in modifying the voice sounds can be demonstrated by a simple experiment. Since the speed of sound goes as the reciprocal of the square root of the density, the speed of sound may be varied by using different gases. If one uses helium instead of air, since the density is lower, the speed of sound is much higher, and all the frequencies of a cavity will be raised. Consequently if one fills one's lungs with helium before speaking, the character of his voice will be drastically altered, even though the vocal cords may still be vibrating at the same frequency.

40-4 Coupled pendulums

Finally we should emphasize that not only do modes exist for consolidated oscillating systems, but also for very simple linear-like systems. A good example is the system of two coupled pendulums discussed in the preceding chapter. In that chapter it was shown that the motion could be analyzed as a superposition of two harmonic motions with different frequencies. So even this system can be analyzed in terms of harmonic motions or modes. The string has an infinite number of modes and the two-dimensional surface also has an infinite number of modes. In a sense it is analogous really. If we know how to count identities. But a simple "pendulum," which has only two degrees of freedom, and requires only two variables to describe it, has only two modes.

Let us make a mathematical analysis of the two modes for the case where the pendulums are of equal length. Let the displacement of mass be x , and the displacement of the other be y , as shown in Fig. 40.5. Without a spring, the force on the first mass is proportional to the displacement of that mass, times of gravity. There would be, if there were no spring, a certain natural frequency ω_0 for this one alone. The equation of motion would be

$$m \frac{d^2x}{dt^2} = -\omega_0^2 x \quad (40.13)$$

The other pendulum would swing in the same way if there were no spring. In addition to the force of restoration due to gravitation, there is an additional force pulling the first mass. That force depends upon the excess of displacement of the second mass, and is proportional to that difference, so it is a restoring force which depends on the symmetry, times $(x - y)$. The same force in reverse sense acts on the second mass. The equations of motion then have to be solved like this:

$$m \frac{d^2x}{dt^2} = -\omega_0^2 x - k(x - y), \quad m \frac{d^2y}{dt^2} = -\omega_0^2 y - k(y - x). \quad (40.14)$$

In order to find a motion in which both of the masses move at the same frequency, we must determine how much each mass moves. In other words, pendulum x and pendulum y will oscillate at the same frequency, but their amplitudes must have certain values, A and B , whose relation is fixed. Let us try this solution:

$$x = A e^{i\omega t}, \quad y = B e^{i\omega t}. \quad (40.15)$$

If these are substituted into Eqs. (40.14) and like terms are collected, the result is:

$$\begin{aligned} \left(\omega^2 - \omega_0^2 - \frac{k}{m}\right) A + \frac{k}{m} B &= 0, \\ \left(\omega^2 - \omega_0^2 - \frac{k}{m}\right) B - \frac{k}{m} A &= 0. \end{aligned} \quad (40.16)$$

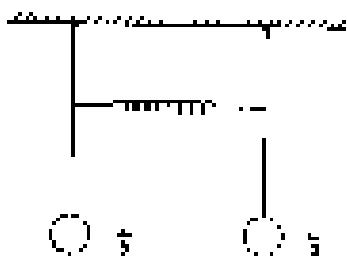


Fig. 40.5. Two coupled pendulums.

The equations we have have had the constant factor ω^2 removed and have been divided by m .

Now we see that we have two equations for what looks like two unknowns. For there can be no two unknowns because the whole idea of the motion is something that we cannot determine from these two alone. The above equations can relate one only the ratio $\omega_1^2/\omega_2^2 = k_1/k_2$, but they *never* both give the same result. The necessity for both of these equations to be consistent is a requirement that the frequency be something very special.

In this particular case this can be worked out rather easily. If the two equations are multiplied together, the result is

$$\left(\frac{\omega_1^2}{\omega_2^2} - \frac{k_1}{k_2}\right) \cdot k_2 = \left(\frac{k_1}{k_2}\right)^2 \cdot k_2 \quad (49.17)$$

The term k_2 can be removed from both sides unless k_2 is 0 or ∞ , which means there is no motion at all. If that's not true, then the two terms must be equal, giving a quadratic equation to solve. The result is that there are two possible frequencies:

$$\omega_1^2 - \omega_2^2 = \omega_2^2 - \omega_0^2 + \frac{2k_1}{m}. \quad (49.18)$$

Furthermore, if other values of frequency are substituted back into Eq. (49.16), we find that for the first frequency $\omega = \omega_1$, and the second frequency $\omega = \omega_2$. These are the "natural" shapes, as can be readily verified by experiment.

It is clear that in the first mode, where $\omega = \omega_1$, the spring is never stretched and both masses oscillate at the frequency ω_1 , as though the spring were absent. In the other situation, where $\omega = \omega_2$, the spring contributes a restoring force and raises the frequency. A more interesting case results if the pendulums have different lengths. The analysis is very similar to that given above, and is left as an exercise for the reader.

49-5 Linear systems

Now let us summarize the ideas discussed above, which are all expected to be probably the most general and wonderful principle of mathematical physics. If we have a linear system whose character is independent of the time, then the motion does not have to have any particular simplicity, and it fact may be exceedingly complex, but there are still special motions, usually a series of modes, motions, in which the whole position of motion varies exponentially with the time. For the vibrating systems that we are talking about now, the exponential is imaginary, and instead of saying "exponentially" we might prefer to say "sinusoidally" with time. However, on some higher level we may say that the motion will vary exponentially with the time in very special modes, with very special shapes. The next general motion of the system can always be represented as a superposition of motions involving each of the different frequencies.

This is worth stating again for the case of sinusoidal motion: a linear system need not be moving in a purely sinusoidal motion, i.e., at a definite single frequency. But no matter how it does move, this motion can be represented as a superposition of pure sinusoidal motions. The frequency of each of these motions is a characteristic of the system and the form or waveform of each motion is also a characteristic of the system. The general motion in any such system can be determined by giving the strength and the phase of each of these modes, and adding them all together. Another way of saying this is, letting linear vibration system be equivalent to a set of independent harmonic oscillators, just the natural frequency is corresponding to the modes.

We conclude the chapter by remarking on the connection of waves with quantum mechanics. In quantum mechanics the vibrating object, or the drop test particle, is again, the amplitude of a probability function that gives the probability of finding an electron, or system of electrons, in a given configuration. This amplitude function can vary in space and time, and satisfies, in fact, a linear equation

But in quantum mechanics there is a fundamental, in-between state, in which we will find some of the probability amplitude is equal, in the classical idea, to energy. Therefore we can consider the principle of least action to also be holding the word frequency and replacing it with energy. It becomes one of the like this a quantum mechanical system, for example an atom, need not have a definite energy, just as a simple mechanical system does not have a definite frequency; but consider how the system behaves, its behavior can always be represented as a superposition of states of definite energy. The energy of each state is a definite value of the sum, and so is the sum of amplitude which determine the probability of finding particles in different places. The general motion can be described by adding the amplitude of each of these different energy states. This is the origin of energy levels in quantum mechanics, since quantum mechanics is represented by waves in the wavefunction in which the electron does not have enough energy to ultimately escape from the proton, they are confined waves. For the quantized states of existing, there are definite frequencies, as the solution of the wave equation, as given in mechanics. The quantum-mechanical interpretation is that there are definite energies. The above quantum-mechanical system, because it is represented by waves, can have definite states of fixed energy; its amplitudes are the energy levels of definite states.

Musical tones

50.1 Musical tones

Pythagoras is said to have discovered the fact that two similar strings under the same tension are in tune only if their length ratio is such that it is a ratio of small integers. If the length ratio is 1 to 2, they will correspond to the octave interval. If the lengths are in the ratio of 2 to 3, they will correspond to the interval between C and G, which is called a 5th. These intervals are generally recognized as "pleasant" sounding chords.

Pythagoras was so interested by his discovery that he made it the basis of a school. Pythagoreans they were really—where real mystic beliefs in the great powers of numbers. It was believed that something similar would be found in about the planets, or "aphorae." We sometimes hear this expression "the music of the spheres." The idea was that there would be some kind of relationship between the orbits of the planets or between other things in nature. Pythagoreans think that this is in a kind of a way "inherited" from the Greeks, who is also different from our own scientific interest in quantitative relationships. Pythagoras' discovery was the first example, outside geometry, of any numerical relationship in nature. It must have been very surprising to suddenly discover that there was a law of nature that involved a simple geometrical regularity. Simple measurements of lengths give a prediction about something which has no apparent connection to geometry—the production of pleasant sounds. This discovery led to the extension that perhaps a general law for understanding nature could be found in mathematical regularities. The results of modern science justify that point of view.

Pythagoras would only have made his discovery by making an experimental observation. Yet this important aspect does not seem to have impressed him. It is hard, despite all the literature that has been written about it, to imagine why he did not decide simply have had a much earlier start. (It is always easy to look back at what someone else has done and to decide what he should have done.)

We might remark on a third aspect of this very interesting discovery: the discovery had to do with two issues that are significant in the ear. We may question whether we are any better off than Pythagoras in understanding why such certain sounds are pleasant to our ear. The general theory of aesthetics is probably an entirely different issue from a theory of Pythagoras' or the first discovery of the Greeks. There are, however, aspects: experiment, mathematical relationships and aesthetics. Physics has made great progress on only the last two parts. This chapter will deal with our present-day understanding of the discovery of Pythagoras.

Among the sounds that we hear, there is one kind that we call tone. It corresponds to a sort of inharmonic vibration of the continuum that is produced by the irregular vibration of some string in the neighbourhood. If we make a drumskin to indicate the pressure of the air on the ear-drum, such, however, the displacement of the drum, is a function of time, the graph which corresponds to a twice night such like that shown in Fig. 50-1(a). (Such a noise might correspond roughly to the sound of a cracked pot.) The sound of music has different characteristics. Music is characterized by the presence of more-or-less sustained regular or musical "tones." Musical instruments may produce notes as short as this tone, but for a relatively short time, as when a key is pressed on a piano, or it may be sustained almost indefinitely, as when a flute player holds a long note.

What are the special characters of a musical note from the point of view of the pressure in the ear? A musical note differs from a noise in that there is a periodicity in its graph. There is some uneven shape to the variation of the air pressure with

50-1 Musical tones

50-2 The Fourier series

50-3 Quality and resonance

50-4 The Fourier condition

50-5 The energy theorem

50-6 Nonlinear responses



Fig. 50-1. Pressure as a function of time for (a) a noise, and (b) a musical tone.

time, and the shape repeats itself over and over again. An example of a pressure-time function that would correspond to a musical note is shown in Fig. 30-1(b).

Musical notes usually consist of a musical note in terms of their characteristic tones: the loudness, the pitch, and the "quality." The "loudness" corresponds to the amplitude of the note or changes. The "pitch" corresponds to the period of time between repetition of the basic waveform. ("Low" notes have longer periods than "high" notes.) The "quality" of a tone has to do with the difference of "overtones" between two notes of the same loudness and pitch. An oboe, a violin, or a soprano will distinguishable even when they sound notes of the same pitch. The quality has to do with the structure of the note during pitch.

Let us consider, for a moment, the sound produced by a vibrating string. If we pluck the string, by pulling it to one side and releasing it, the subsequent motion will be determined by the motions of the waves we have generated. We know that these waves will travel in both directions, and will be reflected at the ends. They will reflect back and forth for a long time. We realize how complicated the wave is, however, if we repeat itself. The period of repetition is just the time required for the wave to travel the full length of the string. The result is just the same as general for any wave, once started. It reflects at each end and returns to its starting position, and so proceeding in the original direction. The time it becomes tor waves to stabilize is either dimension. Touching on the string will, therefore, return to its starting position after one period, and again one period later, etc. The sound wave produced... we also have the same repetition. We see why a plucking produces a musical note.

30-1 The Fourier series

We have discussed in the preceding chapter another way of looking at the motion of a vibrating system. We have seen that a simple harmonic motion, modes of oscillation, and that any particular kind of vibration that may be set up by the initial conditions can be thought of as a combination of multiple proportions—of several of the natural modes, oscillating together. For a string we found that the natural modes of oscillation had the frequencies $\omega_0, 2\omega_0, 3\omega_0, \dots$. The most general motion of a plucked string, therefore, is composed of the sum of a sinusoidal oscillation at the fundamental frequency ω_0 , and at the second harmonic frequency $2\omega_0$, and the third harmonic $3\omega_0$, etc. Now the fundamental mode repeats itself every period $T_0 = 2\pi/\omega_0$. The second harmonic mode repeats itself every $T_0/2 = \pi/\omega_0$, i.e., it repeats itself every $T_0/2$, also, i.e., it is periodic. Similarly, the third harmonic mode repeats itself after a time $T_0/3$, which is 1/3 of its periods. We see again why a plucked string, repeats its whole pattern with a periodicity of T_0 . It produces a periodic wave.

We have been talking about the motion of the spring. On the spring, which is a conductor of tension, is produced by the motion of the string, so its vibrations too must be composed of the same manner as though we were changing the ring about the center of mass of the ring. As on, the relative strengths of the amplitudes may be different in the various parts of the string, particularly if the string is "tautened" to one end and left free at the other. The efficiency of the coupling to the ring is different for different harmonics.

If we let $f(t)$ represent the air pressure as a function of time for a mono-harmonic case (as in Fig. 30-1(c)) then we expect the $f(t)$ can be written as the sum of two simple harmonic oscillations of time. Like cosine. As such, it has two inharmonic frequencies. If the period of the vibration is T_0 , the fundamental angular frequency will be $\omega = 2\pi/T_0$, and the harmonics will be $2\omega, 3\omega$, etc.

There is one slight complication. For each frequency we now expect that the various phases will not necessarily be the same for all frequencies. We should, therefore, use functions like $\cos(\omega_0 t + \phi)$. It is, however, simpler to use instead both the sine and cosine functions for each frequency. We recall that

$$\cos(\omega_0 t + \phi) = (\cos \omega_0 t) \cos \phi - (\sin \omega_0 t) \sin \phi \quad (30.1)$$

and since ω is a constant, any sinusoidal oscillation ω , the frequency ω can be written as the sum of a term $\omega_1 \cos \omega t$ and another term with an ω :

We conclude then, that any function $f(t)$ that is periodic with the period T can be written mathematically as

$$\begin{aligned} f(t) &= a_0 \\ &\quad + a_1 \cos \omega t + b_1 \sin \omega t \\ &\quad + a_2 \cos 2\omega t + b_2 \sin 2\omega t \\ &\quad + \dots = \dots \end{aligned} \quad (30.2)$$

where $\omega = 2\pi/T$ and the a_i 's and b_i 's are numerical constants which tell us how much of each component oscillation is present in the oscillation $f(t)$. We have called the "harmonics" terms a_i , so that the formula will be completely general, although it is usually used for a musical tone. It represents a start of the analysis called about ω , the "fundamental" term, of the sound pressure. With ω , our formula can take care of any case. The equality of Eq. (30.2) is represented schematically in Fig. 30-2. (The amplitudes, a_i , and b_i , of the harmonic fractions must be suitably chosen. They are drawn schematically and without any particular scale in the figure.) The series (30.2) is called the Fourier series for $f(t)$.

We have said that any periodic function can be made up in this way. We should even go back and say that any sound wave, or any function we ordinarily encounter in physics, can be made up of such ω 's. The mathematical equivalent functions which cannot be made up of simple harmonic functions—for instance, a function that has a "twelve-twist" so that it has two values for some values of t . We need not worry about such functions here.

30-3 Quality and consonance

Now we are able to answer the question: what is the "character" or "quality" of a musical tone? It is the relative amplitudes of the various harmonics. The ratios of the a_i 's are f_i 's. A tone with only one harmonic is a "pure" tone. A tone with many strong harmonics is a "rich" tone. A violin produces a different proportion of harmonics than does an oboe.

We can "manufacture" various musical tones if we connect several "resonators" to a loudspeaker. (An oscillator usually produces a nearly pure single harmonic function.) We could choose the frequencies of the oscillators to be ω , 2ω , 3ω , etc. Then by adjusting the volume control on each oscillator, we can add in, say, a_{10} , a_{15} , etc., of each harmonic, thereby producing tones of different quality. An oboe soprano does it in such this way. The "keys" select the frequency of the fundamental note, and the "keys" also switches that control the relative amplitudes of the harmonics. By throwing, these switches, the organ can be made to sound like a flute, or like a violin.

It is interesting that to produce such "artificial" tones we need only one oscillator for each frequency—we do not need separate oscillators for the different harmonic components. The ear is not very sensitive to the relative phases of the harmonics. It pays attention mainly to the ratio of the amplitudes of each frequency. Our "ear" is more accurate than is necessary to explain the subjective aspect of music. The response of a microphone or other physical instrument does depend on the phases, however, and one complete analysis may be needed in these cases.

The "quality" of a spoken word also determines the novel sounds that we recognize in speech. The shape of the mouth determines the frequencies of the natural modes of vibration of the air in the mouth. Some of these modes are set into vibration by the sound waves from the vocal cords. In this way, the amplitudes of some of the harmonics of the sound are increased with respect to others. When we change the shape of our mouth, harmonics of different frequencies are given preference. These effects account for the differences between an "o-o-o" sound and an "e-e-e" sound.



Fig. 30-2. Any periodic function $f(t)$ is equal to a sum of simple harmonic oscillations.

We know that a particular vowel sound, say "ee-eh", will sounds slightly the same vowel whether we say it in singing at a high or a low note. From the evidence we describe, we would expect that the voice frequency to increase linearly when we change our mouth so that "ee-eh" could have different changes in range the pitch of our voice. So the relation of the important harmonics to the fundamental tone is, that is, the "tonality" changes as we change pitch. Apparently the mechanism by which we recognize speech is not based on specific frequency relationships.

What should we say now about Pythagoras' discovery? We understand that two similar strings with length λ in the ratio $\sqrt{2} : 1$ will have frequencies in the ratio $\sqrt{2} : 1$. But why should they "sound pleasant" together? Perhaps we should take out clue from the frequencies of the instruments. The second harmonic of the lower string's song will have the same frequency as the fundamental of the longer song. (It is easy to show or to believe that a plucked string produces strongly the second lowest harmonic.)

Because we cannot make the "dissonant" rules. Not because compound waves they have harmonics with too many frequency. Notes sound dissonant if their upper harmonics have frequencies near to each other but "harmonious" agent that it is no rapids beats between the two. Why beats do not come possible, and why union of the upper harmonics were called consonant, is something that we do not know how to reason or describe. We can not say true this knowledge of what sounds good, who, say, is enough to say good. In other words, our understanding of it is not anything more general than the statement that when they sound in unison they sound good. It does not permit us to deduce anything more than the properties of consonances in music.

It is easy to check on the harmonic relationships we have discussed by some simple experiments with a piano. Let us take the piano C note near the middle of the keyboard by C, C', and C'', the note C'' just above by C, C', and C''. Then the fundamental will have effective frequencies as follows:

$$\begin{array}{ll} C = 2 & C = 3 \\ C' = 4 & C' = 6 \\ C'' = 6 & C'' = 12 \end{array}$$

These harmonic relationships can be demonstrated in the following way: Suppose we press C sharply—so that it does not sound, but we cause the damper to be lifted. If we then sound C, it will produce its own fundamental and some second harmonics. The second harmonic of C is the same as the C of C'' is a vibration. If we now release C, leaving C' pressed the damper will stop the vibration of the C strings, and we can hear (only) the note C'' as it dies away. In similar way, the third harmonic of C is the same as the C of C'. Or the note of C'' after getting released is the same as a vibration in the fundamental of C''.

A somewhat different result is obtained if we press C sharply and then sound C''. The third harmonic of C will correspond to the fourth harmonic of C, so only the fourth harmonic of C will be excited. We can hear (if we strain hard) the sound of C'', while, of course, nevertheless the C and "the piano". It is easy to think up more interesting combinations for this game.

We may remark by analogy that the major scale can be defined just by the condition that the three major chords (C-A-G, F-A-D, and G-B-E) each consist four consecutive notes in frequency ratio of 3:5:6. These ratios plus the fact that an octave (C-C', B-B') has the ratio 1:2 determine the whole scale for the "Major" mode, or for what is called "just intonation." A keyboard instrument like the piano are not usually tuned in this manner, but a flute "tuning" is done so that the frequencies of a soprano vocal range fall on "just" intonation points. For this tuning, which is called "tempered," the octave (still C-C') is divided into 12 equal intervals for which the frequency ratio is $(2)^{1/12}$. A fifth no longer has the frequency ratio 3/2, but 2^{1/5} ≈ 1.199 which is apparently close enough for most uses.

We have stated now for consonance differences at the considered time the meaning. Is this coincidence perhaps the reason that we make no statement? One reason for this is that we just cannot really expect it to be free of harmonics—the only way of consonance & dissonance as the relative frequencies are placed one after the other, the expected ratio. Such experiments are difficult because it is difficult to measure the pitch ratios, for reasons that we will see later. We cannot say for certain whether the ear is perceiving harmonics or being attracted when we decide that we like a sound.

5.4 The Fourier coefficients

For us to know in the Fourier synthesis theorem, a periodic signal $x(t)$ can be represented by a suitable combination of harmonics. We would like to show how we can find out what numbers of each harmonic are required. This, of course, does not apply to $f(t)$, using Eq. (50.2), if we are given all the coefficients a_n and b_n . The question now is, if we are given $f(t)$ how can we know what the coefficients of the various harmonic terms should be? (It is very suitable to take this recipe, but first we write down the recipe to we also get a cake!)

Fourier discovered that it was not really very difficult. The answer is certainly yes. We know already that it is $f(t)$ the average value $\bar{x}(t)$ of $x(t)$ over one full cycle $T = 2\pi/\omega = 1/f$. We can easily see that this is indeed so. The average value of a sinusoidal function over one period is zero. Thus two, three, or any whole number of periods, it is also zero. So the average value of all of the terms on the right hand side of Eq. (50.2) is zero, except for a_0 , unless—that we must remember $\omega = 2\pi/T$.

Now the average of a sum is the sum of the averages. So the average $\bar{f}(t)$ is just the average of a_0 . But a_0 is a constant, so its average is just the same as its value. Recalling the definition of an average, we have

$$a_0 = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) dt. \quad (50.3)$$

The other coefficients are much more difficult. To find them we can use a trick discovered by Fourier. Suppose we multiply both sides of Eq. (50.2) by some harmonic function—say by $\cos 2\omega t$. We then have

$$\begin{aligned} f(t) \cos 2\omega t &= a_0 \cdot \cos 2\omega t \\ &\quad + a_1 \cos \omega t \cdot \cos 2\omega t + b_1 \sin \omega t \cdot \cos 2\omega t \\ &\quad + a_2 \cos 2\omega t \cdot \cos 2\omega t + b_2 \sin 2\omega t \cdot \cos 2\omega t \\ &\quad \cdots \quad + \cdots \\ &\quad + a_n \cos n\omega t \cdot \cos 2\omega t + b_n \sin n\omega t \cdot \cos 2\omega t \\ &\quad \cdots \quad + \cdots \end{aligned} \quad (50.4)$$

Now let us average both sides. The average of $a_n \cos n\omega t$ over the time T is proportional to the average of $\cos^2 n\omega t$ over T whole periods. But that is just zero. The average of $b_n \sin n\omega t$ of the left of the terms is also zero. That is back to the a_n terms. We know, in general, that

$$\cos n(2\omega t) = 1/2(a_n + A_n) + 1/2a_n(A_n - B_n). \quad (50.5)$$

5.4.1 Amplitudes

$$a_n = 1/2(a_n + A_n). \quad (50.6)$$

We thus have two terms to deal with with $2\omega t$ periods in T and the others with ω . Their total average is zero. The average of the a_n terms is therefore zero.

By the same token, we would find a_n to be the total average, each of which also has to be zero. For the A_n term we would find a_n after multiplying $(-A_n)$. That multiplication is the same as $\cos 2\omega t$, so both of these have zero averages. It is clear

that all of the n terms will have a zero average except one. And that one is the a_0 term. For this one we have

$$\langle \cos(\omega t + \pi/2) \rangle = 0. \quad (30.7)$$

The cosine of zero is one, and its average, of course, is one. So we know the result that the average of $\langle \cos^2 \theta \rangle$ of the n terms of Ψ_0 (30.4) equals to unity.

Two terms are even easier. When we multiply by any cosine term, like $\cos \omega t$, we can show by the same method that all of the n terms have the same $\langle \cos^2 \theta \rangle$ value zero.

We see that Fourier's "trick" has acted here to zero. When we multiply by $\cos^2 \theta$ and average, all terms drop out except one, and we find that

$$\text{Average } \langle (\cos \omega t)^2 \rangle = a_0^2/2, \quad (30.8)$$

or

$$a_0 = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cdot \cos \omega t dt. \quad (30.9)$$

We shall leave it for the reader to show that the coefficients b_n can be obtained by multiplying Eq. (30.2) by $\sin \omega t$ and averaging over sides. The result is

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cdot \sin \omega t dt. \quad (30.10)$$

Now what is true for T we expect is true for any integer. So we can summarize our proof and result in the following more elegant mathematical form. If n is odd (an integer other than zero), and if $\omega = 2\pi/T$, then

$$\text{I. } \int_0^T \sin n \omega t dt = 0 \quad (30.11)$$

$$\text{II. } \int_0^T \cos n \omega t dt = \begin{cases} 0 & \text{if } n \neq m \\ 1/2 & \text{if } n = m \end{cases} \quad (30.12)$$

$$\text{III. } \int_0^T \sin m \omega t \cos n \omega t dt = \begin{cases} 0 & \text{if } n \neq m \\ 1/2 & \text{if } n = m \end{cases} \quad (30.13)$$

$$\text{IV. } f(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n \omega t + \sum_{n=1}^{\infty} b_n \sin n \omega t. \quad (30.14)$$

$$\text{V. } a_0 = \frac{1}{T} \int_{-T/2}^{T/2} f(t) dt. \quad (30.15)$$

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cdot \cos n \omega t dt. \quad (30.16)$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cdot \sin n \omega t dt. \quad (30.17)$$

In earlier chapters it was convenient to use the exponential notation for representing simple harmonic motion. Instead of this we used $\cos \omega t$, the real part of the exponential function. We have used cosine and sine functions in this chapter because it made the derivations parallel with elements. Our final result of Eq. (30.17) can, however, be written in a more compact form

$$f(t) = R e \sum_{n=-\infty}^{\infty} A_n e^{i n \omega t}, \quad (30.18)$$

where A_n is the complex number $a_n - i b_n$ (with $b_0 = 0$). If we wish to use the same notation throughout, we can write also

$$a_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-i n \omega t} dt \quad (n \geq 1) \quad (30.19)$$

We now know how to "analyze" a periodic wave into its harmonic components. The procedure is called Fourier analysis, and the resulting series is called Fourier components. We have said, however, that since we find all of the Fourier components and add them together we do indeed get back our $f(t)$. The corresponding area shown, for a wide class of functions, is that for ψ . Just one of interest to visualize that if we can do the integral ψ we will get back $f(t)$. There is one minor exception. If the function $f(t)$ is discontinuous, e.g., it jumps suddenly from one value to another, the Fourier sum will give a value at the midpoint to "average between the jumps" and never reaches the discontinuity. So if we have the strange discontinuity $f(t) = 0.0 \leq t < t_0$, and $f(t) = 1.0 \text{ for } t_0 \leq t \leq T$, the Fourier sum will give the right value everywhere except at t_0 , where it will have the value $\frac{1}{2} \text{ instead of } 1$. This rather unpleasant ambiguity in integral that is fraction should be brought to t_0 , but I neglect it. So perhaps we should make the "rule" for discontinuous or very discontinuous function $f(t)$ that only be a singularity for a real physical function that it be differentiable halfway between all the discontinuities. Then any such function within any finite number of such jumps, as well as other physically interesting functions, are given exactly by the Fourier sum.

As an exercise, we suggest that one reader determine the Fourier series for the function shown in Fig. 20-3. Since the function cannot be written in an explicit algebraic form, you will not be able to do the integrals from zero to T in the usual way. The integrals are easy, however, if we separate them into two parts. The integral from zero to $T/2$ over which $f(t) = 0$ and the integral from $T/2$ to T (over which $f(t) = -1$). The result should be

$$f(t) = \frac{4}{\pi} \left[\sin \omega t + \frac{1}{3} \sin 3\omega t + \frac{1}{5} \sin 5\omega t + \dots \right] \quad (20.19)$$

where $\omega = 2\pi/T$. We thus see that our square wave (with the particular phase chosen) has odd harmonics and their amplitudes are in inverse proportion to their frequencies.

Let us check that Eq. (20.19) does indeed give us back $f(t)$ for some value of t . Let us choose $t = T/4$, or $\omega t = \pi/2$. We find

$$f(t) = \frac{4}{\pi} \left(\sin \frac{\pi}{2} + \frac{1}{3} \sin \frac{3\pi}{2} + \frac{1}{5} \sin \frac{5\pi}{2} + \dots \right) \quad (20.20)$$

$$= \frac{4}{\pi} \left(1 - \frac{1}{3} - \frac{1}{5} - \frac{1}{7} + \dots \right) \quad (20.21)$$

The series¹¹ has the value $\pi/4$, and we find $\text{Max}(f) = 1$.

20-2 The energy theorem

The energy in a wave is proportional to the square of its amplitude. For a wave of simple shape, the energy in one period will be proportional to $\int_{t_0}^{t_0+T} A^2 dt$. We can also relate the energy to the Fourier coefficients. We write

$$\int_{t_0}^{t_0+T} f^2(t) dt = \int_{t_0}^{t_0+T} \left[a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t \right]^2 dt. \quad (20.22)$$

When we expand the square of the bracketed term we will get a few terms, such as a_0^2 , etc. But there are two. We have shown above, however, that Eqs. (20.11) and (20.12) the n th integral of x^n over one period is zero.

¹¹ The series can be evaluated in the following way. First we notice that $\int_0^{\pi/2} \sin x dx = 1$, and $\int_0^{\pi/2} x \sin x dx = \text{constant}$, so expand the integral in powers of $x/\pi/2$: $x^2 = 1 - x^2/2 + x^4/2^2 - x^6/3! + \dots$ We integrate the series term by term. Then we substitute to obtain $\int_0^{\pi/2} x \sin x dx = \pi/2 - \pi^2/8 + \dots$. Setting $x = 1$, we have the stated result, since $\int_0^{\pi/2} \sin x dx = 1$.

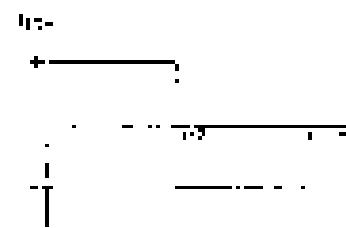


Fig. 20-3. Square-wave function
 $f(t) = 1 \text{ for } 0 < t < T/2$
 $f(t) = -1 \text{ for } T/2 < t < T$

We have left out the square term like $\omega_0^2 \sin^2 \omega_0 t$. The integral of any cosine squared or sine squared over one period is equal to $T/2$, so we get

$$\begin{aligned} \int_0^T f^2 dt &= T\omega_0^2 + \frac{T}{2}(\omega_0^2 - \omega_0^2 + \dots - \omega_0^2 - \omega_0^2 + \dots) \\ &= T\omega_0^2 - \frac{T}{2} \sum_{k=1}^{\infty} (\omega_k^2 + \omega_k^2) \end{aligned} \quad (30.23)$$

This equation is called the "energy theorem," and says that the total energy in a wave is just the sum of the energies in all of the Fourier components. For example, applying this theorem to the series (30.18), since $|f(n)|^2 = 1$ we get

$$\pi = \frac{\pi^2}{2} \cdot \left(\frac{4}{\pi}\right)^2 \left(1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \dots\right),$$

so we learn that the sum of the squares of the reciprocals of the odd integers is $\pi^2/8$. In a similar way, by calculating the Fourier series for the function and using the energy theorem, we can prove that $1 + 1/2 + 1/3 + \dots$ is $\pi^2/6$, a result we needed in Chapter 25.

30.6 Nonlinear responses

The y , in the theory of harmonics, has an important precursor which should be recalled since because of its practical importance that of nonlinear effects. In all the systems that we have been considering so far, we have supposed that everything was linear, that the response to forces, say the displacement or the accelerations, were always proportional to the forces. Or that the currents in the circuit were proportional to the voltages, and so on. We now wish to consider cases where there is no strict proportionality. We think of the manner of some device in which the response, which we will call x_{out} at the time t , is determined by the input x_{in} at the time t . For example, x_{in} might be the force and x_{out} might be the displacement. Or x_{in} might be the current and x_{out} , the voltage. If the device is linear, we would have

$$x_{\text{out}}(t) = K x_{\text{in}}(t), \quad (30.24)$$

where K is a constant independent of x_{in} and of t . Suppose, however, that the device is really, non-linearly, linear, so that we can write

$$x_{\text{out}}(t) = K[x_{\text{in}}(t)] + \epsilon x_{\text{in}}^2(t), \quad (30.25)$$

where ϵ is small enough compared with unity. Such linear and nonlinear segments are shown in the graphs of Fig. 30-4.

Nonlinear responses have several important practical consequences. We shall discuss some of them now. First we consider what happens if we apply a pure tone at the input. We let $x_{\text{in}} = \cos \omega t$. If we plot x_{out} as a function of time we get the solid curve shown in Fig. 30-5. The dashed curve gives, however, the response of a linear system. We see that the output is not just a cosine function. It is more peaked at the top and dithered at the bottom. We say that the output is distorted. The linear, however, has such a wave is no longer a pure wave, but it will have harmonics. We can find what the harmonics are. Using $x_{\text{in}} = \cos \omega t$ with Eq. (30.25), we have

$$x_{\text{out}} = K[\cos \omega t] + \epsilon \cos^2 \omega t. \quad (30.26)$$

From the equality $\cos^2 \theta = \frac{1}{2}(1 + \cos 2\theta)$, we have

$$x_{\text{out}} = K \left(\cos \omega t + \frac{1}{2} + \frac{1}{2} \cos 2\omega t \right). \quad (30.27)$$

The output has not only a component at the fundamental frequency that was present in the input, but also has some of its second harmonic. This may seem

appeared at the output a constant term $A(0/2)$, which corresponds to the shift of the average, $\omega_1 - \omega_2$, shown in Fig. 50.5. The process of calculating a shift of the average value is called **downmixing**.

Nonlinear responses will clearly not yet produce harmonics of the frequencies ω_1 and ω_2 . Although even linearity we assumed produced only second harmonics, nonlinearity will also produce terms which have terms like ω_1^2 and ω_2^2 . For example, " ω_1^2 " generates a harmonic higher than the second.

Another effect which results from a nonlinear response is modulation. If our linear function contains both (or more) poles at ω_1 , the output will have not only their harmonics, but still other frequency components. Let $x_{11} = A \cos(\omega_1 t + \theta_1)$ and $x_{12} = A \cos(\omega_2 t + \theta_2)$ be needed to be in the harmonic addition. In addition to the first term (which has K times the input) we shall have a component to be roughly given by

$$x_{11} = K(4 \cos(\omega_1 t + \theta_1) \cos \omega_2)^2 \quad (50.28)$$

$$= K(16 \cos^2 \omega_1 t + 8 \cos^2 \omega_2 t + 24 \cos(\omega_1 t) \cos(\omega_2 t)). \quad (50.29)$$

The first two terms in the parentheses of Eq. (50.29) are just those which gave the constant term and some harmonics terms we found above. The last term is new.

We can look at this new "cross term" in different ways in two ways. First, if the two frequencies are widely different (for example, if ω_2 is much greater than ω_1) we can consider that the cross term represents a cosine modulation of varying amplitude. That is, we can think of the factors in this way:

$$AB \cos(\omega_1 t) \cos(\omega_2 t) = C(t) \cos(\omega_1 t) \quad (50.30)$$

with

$$C(t) = AB \cos(\omega_2 t). \quad (50.31)$$

We say that the amplitude of $\cos(\omega_1 t)$ is modulated with the frequency ω_2 .

Alternatively, we can write the cross term in another way:

$$AB \cos(\omega_1 t) \cos(\omega_2 t) = \frac{AB}{2} [\cos(\omega_1 + \omega_2)t - \cos(\omega_1 - \omega_2)t] \quad (50.32)$$

We would now expect the new components have been produced, since the sum frequency ($\omega_1 + \omega_2$) and the difference frequency ($\omega_1 - \omega_2$).

We have two different, but equivalent ways of looking at the same result. In the special case that $\omega_1 \gg \omega_2$, we can relate these two different views by remarking that since $(\omega_1 + \omega_2)$ and $(\omega_1 - \omega_2)$ is closer to each other we would expect to observe beats between them. But these beats have just the effect of modulating the amplitude of the average frequency ω_1 by one-half the difference frequency ω_2 . We see, then, why the two descriptions are equivalent.

In summary, we have learned that nonlinear response produces some effects: rectification, generation, and mixing, and modulation, or the generation of components at sum and difference frequencies.

We should notice that all these effects (Eq. 50.29) are proportional not only to the nonlinear coefficient K but also to the product of their amplitudes—either A^2 , B^2 , or AB . We expect these effects to be much more significant for very large than for weak ones.

The effects we have been describing have many practical applications. First, with regard to sound, it is believed that the ear is nonlinear. This is believed to account for the fact that with local sounds we hear the same tone that we may perceive and also a certain difference frequency between the two sound waves coming out of the ears.

The components which are most important regarding equipment—among them, loudspeakers and valves—have some nonlinearity. They produce distortions in the sound—they generate harmonics, etc.—which were not present in the original sound. These new components are heard by the ear and are apparently objectionable. It is for this reason that "Hi-Fi" equipment is designed to be as linear as

possible. (Why the number of the war was not "objectionable" in case such war or how we even know that the neutrality is in the hydrocarbon oil, is true in the war is not clear.)

Non-precision and quite necessary, and one in fact, intentionally made traps in certain parts of radio broadcasting and receiving equipment. In an oscillator for the "carrier" signal (with frequency of some kilocycles per second) is combined with the "modulator" signal (with a frequency of some megacycles per second) in a nonlinear circuit called a modulator, to produce the modulated oscillation that is transmitted. In this case, the components of successive cycles are fed to a nonlinear circuit which combines the sum and difference frequencies of the modulated carrier to generate a modulated wave signal.

When we discussed the transmission of light we assumed that the induced oscillations of charges were proportional to the electric field of the light. That the assumption was correct. That is indeed a very good approximation. It is only within the last few years that light sources have been devised whose positive as intensity of light strong enough so that nonlinear effects can be observed. It is now possible to generate harmonics of light frequencies. When a strong red light passes through a piece of glass, a little bit of blue light is produced—summarized.

Waves

SI-1. Bulk waves

Although we have finished our quantitative analysis of waves, this added chapter on the subject is intended to give some approximation, qualitatively, for certain phenomena that one comes up with waves, which are too complicated to analyze in detail here. Since we have been dealing with waves for several chapters, there perhaps the subject might be called "one of the most complex phenomena associated with waves."

The first topic to be discussed concerns the effects that are produced by a source of waves, which is moving faster than the wave velocity, or the phase velocity. Let us first consider waves that have a definite velocity. The sound and light. If we have a source of sound which is moving faster than the speed of sound, then something like this happens: Suppose at a given moment a sound wave is generated from the source at point x_1 in Fig. SI-1; then, in the next instant, as the source moves to x_2 , the wave from x_1 expands by a radius $c_s t$, where t is the distance that the source moves, and, of course, another wave starts from x_2 . When the second source has moved a bit farther, to x_3 , and a wave is starting there, the wave from x_1 has now expanded to x_3 , and the one from x_2 has expanded to x_3 . Of course, the length does not increase, nor in steps, and therefore we have a series of wave circles with a common tangent line which goes through the center of the source. We see that the end of a source generating spherical waves, as it would if it were extending in σ , generates a wave front which forms a cone in three dimensions, or a pair of lines in two dimensions. The angle of the cone is very easy to figure out. In a given amount of time the source moves a distance, say $a c_s t$, perpendicular to σ , the velocity of the source. In the meantime the sound will have moved out a distance $c_s t$, proportional to $c_s t$, the speed of the wave. Therefore it is clear that the angle of opening has a sine equal to the ratio of the speed of the wave divided by the speed of the source, and this cone has a solution only if c_s is less than v , or the speed of the object is faster than the speed of the wave.

$$\sin \theta = \frac{c_s}{v}. \quad (SI.1)$$

Incidentally, although we implied that it is necessary to have a source of sound, it turns out, very interestingly, that once the object is moving faster than the speed of sound, it will move slower. That is, it is not necessary that it have a certain constant kinematic character. Any object moving through a medium faster than the speed v at which the medium carries waves will generate waves on each side, automatically, just from the motion itself. This is simple in the case of sound, but it also occurs in the case of light. At first one might think nothing can move faster than the speed of light. However, light in glass has a phase velocity less than the speed of light in a vacuum, and it is possible to share a charged particle of very high energy through a block of glass such that the particle velocity is close to the speed of light in a vacuum, while the speed of light in the glass may be only 3/4 the speed of light in a vacuum. A particle moving faster than the speed of light in the medium will produce a conical wave of light with its apex at the source, like the waves wake from a boat (which is from the same effect, as a matter of fact). By measuring the cone angle, we can determine the speed of the particle. This is most technically to determine the speed of particles in one of the methods of determining their energy in high-energy research. The direction of the light is all that needs to be measured.

SI-1. Bulk waves

SI-2. Shock waves

SI-3. Waves in solids

SI-4. Surface waves

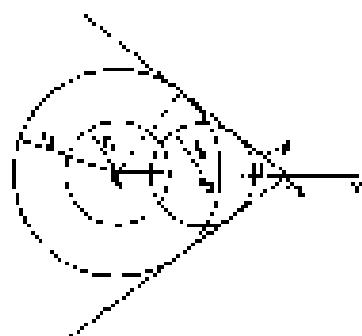


Fig. SI-1. The shock wave front lies on a cone with apex at the source and half-angle $\theta = \sin^{-1} v/c_s$.



Fig. 51-2. A shock wave emitted by an object moving faster than sound.

This light is sometimes called *Cerenkov radiation*, because it was first observed by Cerenkov. It is in effect light that is being emitted theoretically by a body and hence the name. The charged particle produces a change in refractive index, and since a suitable optical system the edge of the source can be made visible. The effect of the object moving faster than the speed of sound, however, makes production somewhat easier. The shock wave that occurs when the surface is actually cutted. It is stronger asymptotically, which is curved near the apex, and we have now discussed how the wave front, which comes up to the second ray of this chapter.

The corresponding experiment has in the last chapter been illustrated in Fig. 51-2, which is a photograph of an object moving through a gas. The object moves faster than the speed of sound. The shock wave, a wave produced because of a change in refractive index, and since a suitable optical system the edge of the source can be made visible. The effect of the object moving faster than the speed of sound, however, makes production somewhat easier. The shock wave that occurs when the surface is actually cutted. It is stronger asymptotically, which is curved near the apex, and we have now discussed how the wave front, which comes up to the second ray of this chapter.



Fig. 51-3. Wavefront propagation over time.

51-2. Shock waves

We expect that depends on the amplitude, and on the size of course the speed depends upon the amplitude in the following way. An object moving through the air has to move through all of the way, so the disturbance produced in this case is a wave front that propagates with the speed of light, which is much faster than the undisturbed region set up reached by the wave front, which is at its actual speed, say. But the air that is left behind, after the wave front passes, has been compressed irreversibly, and so since the wave front is increased. Now the speed of sound increases with the temperature, and the speed in the region behind the jump is finite, that is to say, in the air, in front. This means that every other disturbance, will be faster than the front, the front, the speed increasing with time, because the air is finite, thus makes the disturbance will receive little jumps or pressure pulses, so the pressure contour to aid visualization. We see that the first pulse pressure region in the air, to make the front, as time goes on, until eventually the compression front becomes a sharp front, if the thing is moving right, "sharpened" more and away; it is rather weak, it takes a long time, it may not be seen, that the sound is spreading out, it is not able to do this.

The sound we move is nothing else, it merely wave reflected to the atmospheric pressure, only 1 part in a million cent. But for pressure changes of the order of atmosphere, the wave velocity increases by about twenty per cent, and it makes the sharpness of a disturbance almost negligible. In nature nothing happens nothing is still, permanently and with respect to a "sharp" front has, as well, a very slight compression, it is not infinitely sharp. The distance over which the varying part of the

order of one mean free path, in which the theory of the wave equation begins to fail because we do not consider the structure of the gas.

Now, referring again to Fig. 51-2, we realize more fully why one can be understood. If we suppose that the pressures near the apex are higher than they are farther back, and so the angle α is greater. That is, v_{rel} is the result of the fact that the speed depends upon the strength of the wave. Therefore the wave does an adiabatic compression doesn't do it much faster, having speed about c , while until gets as far as the air is weakened to such an extent that spreading the air pressure bump is small compared with atmospheric pressure. The speed of the bump then approaches the speed of sound in the gas into which it is going. (In other words, it always turns out that the speed of the shock is higher than the speed of sound in the gas class, but is lower than the speed of sound in the gas behind.) That is, impinges from the front will arrive at the front but the front reduces the medium in which it is going faster than the normal speed of sound. So one cannot tell immediately that the shock is coming until it is too late. (The light from the source arrives first, but one cannot tell that the shock is coming until it arrives, because there is no sound signal coming at that time.)

This is a very interesting phenomenon, the piling up of waves and the point at which it occurs is that after a wave is passed, the speed of the resulting wave should be higher. And an example of this same phenomena is the following. Consider water flowing in a long channel with finite width and finite depth. If a piston, or a wall across the channel, is moved along the channel fast enough, water piles up like snow before a snowplow. Now suppose the situation is as shown in Fig. 51-4, with a sudden step in water along x coordinate in the channel. It can be demonstrated that long waves are cleaned out faster in deeper water than they are in shallow water. Therefore any new bump is irregular. A energy supplied by the piston runs forward and piles up at the front. Again, summary what we have is just water not a gas if you, thermally. However, as Fig. 51-4 shows there is a complex flow. Picture is a wave coming up a channel, the piston is at constant height and in the channel. At first it might look appearance like well-behaved wave, as one might expect. But then along the channel, it has become sharp and changes out. It means you get scattered. There is a tendency of moving at the surface, as the pieces of water fall down. In, it is essentially a very sharp rise with distance of the water above.

Actually water is much more complicated than wind. However, just will take a point, we will be interested in point at $x = 0$ in a shallow box, as a model. The point $x = 0$ is not really of any great importance for our purposes; it is not a generalization, it is only to illustrate that the laws of mechanics that we already know are capable of explaining the phenomena.

Imagining for a moment that the water does look something like Fig. 51-5, show the water in the higher height h_2 moving with a velocity v , and the front is moving with velocity v in uniform horizontal water which is of height h_1 . We would like to determine the speed at which the front moves. In this we draw a vertical plane initially at $x = 0$, moves a distance s to the right, while the front of the wave has moved $s + a$.

Now we apply the equations of conservation of matter and momentum. First the former: Per unit channel width, we see that the amount of air at earlier that has moved past $x = 0$ (area shrinks) is compensated by the air created again when it moves to $(x_0 + s)$. So, dividing by dt we get $s(h_2 - h_1)$. This does not yet give us enough, because while up we have h_2 and h_1 , we do not know either v or a ; we are trying to get both of them.

So the next step is to use conservation of momentum. We have now discussed the problems of water pressure, in any long hydrocyclone, but it is clear anyway that the pressure of water at a given depth is just enough to hold up the column of water above it. Therefore the volume of water is equal to the density of water times the depth below the surface, since the pressure increases linearly with depth b , the average pressure over the column is $\rho g b$, which is also the average force per unit width and per unit height pressing the base toward $x = 0$. Now multiply by another v to get the total force which is acting on the water,



Figure 51-4

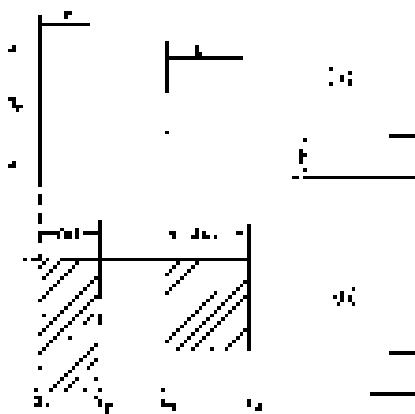


Fig. 51-5. Two cross sections of a wave in a channel, with z the depth of interface at later time t .

passing from the left. On the other hand, the \hat{v} produced by the water on the right also, exerting an opposite force on the right in \hat{v}_2 , relation, where \hat{v}_2 is, by the same kind of analysis, $\hat{v}_2 = \hat{v}_1 - \frac{1}{2} \Delta v$. Now we must balance the forces against the rate of change of the momentum. Thus we have to figure out how much momentum there is in motion (say in \hat{v}_1) \times $-$ from there (say \hat{v}_2). We see that the additional mass that has acquired the speed \hat{v}_2 is just $m \hat{v}_2 \Delta t$ — where m (per unit width), and multiplying this by \hat{v}_1 gives the additional momentum to be supplied to the impulse $F \Delta t$.

$$F \Delta t = m \hat{v}_2 \Delta t = (\hat{v}_1 - \frac{1}{2} \Delta v) m \Delta t.$$

If we eliminate Δv from this equation by substituting $\hat{v}_2 = \hat{v}_1 - \frac{1}{2} \Delta v$, we may divide, and simplify, we get finally that $m \hat{v}_1^2 / \hat{v}_2 = \Delta v / \Delta t$.

If the length difference is very small, so that \hat{v}_1 and \hat{v}_2 are nearly equal, this says that the velocity $= \sqrt{\hat{v}_1}$. As we will see later, that's only one criterion the wavelength of the wave is longer than the depth of the channel.

We would like to do the analogous thing for sound waves. Including the consideration of internal energy, and the conservation of energy, because this theory is reversible. In fact, it can carry the conservation of energy in the here problem, even though that energy is not conserved. If the length difference is small, it is found perfectly agreeable, but as soon as the length difference becomes very appreciable, there is a net loss of energy. This is illustrated as the falling water and the changing \hat{v}_1 in Fig. 8-4.

In shock waves there is a corresponding apparent loss of energy, from the point of view of adiabatic relations. The energy in the sound wave, behind the shock, goes into heating of the gas after shock passes, corresponding to turning of the wave at the bow. In focusing it out, three equations for the sound case turn out to be necessary for solution, and the temperature behind the shock is not the same as the temperature in front as we have seen.

If we try to make a bore that is upscale down ($\hat{v}_2 < \hat{v}_1$), then we have had the energy loss just second is negative. Since that is not available from anywhere, that bore cannot then maintain it; it is unusable. If we were to start a wave of that sort it would dampen out, because the speed dependence on height that resulted in sharpening in the case we discussed would now have the opposite effect.

8-3 Waves in solids

The next kind of waves to be discussed are the more complicated waves in solids. We have already discussed sound waves in gas and in liquid, and there is something in a solid wave in a solid. If a suddenly push is applied to a solid, it is compressed. Increases the compression, and a wave propagates outward if it is strong. However, there is another kind of wave that is possible in a solid, and which hence propagates in a fluid. If a solid is disturbed by pushing it sideways (either sideways displacement to pull itself back), that is by definition what distinguishes a solid from a liquid: if we distort a liquid (like milk), hold it a minute so that it comes down, and then let go, it will stay that way, but if we take a solid and push it, like shearing a piece of "Velva," and let it go, it does both and starts a shear wave, traveling to the wave very fast compression wave. In all cases, the shear wave speed is less than the speed of longitudinal waves. These shear waves are somewhat more longitudinal, so far as their polarizations are concerned to light waves. Sound has no polarization, i.e., it is pressure wave. Light has a characteristic polarization perpendicular to its direction of travel.

In a solid, the waves are of both kinds. First, there is a compressional wave analogous to longitudinal waves of one speed. If the solid is not yet broken, then a shear wave polarized in any direction will propagate at a characteristic speed. Of course, it will be one solid, i.e., the wave has a characteristic speed established by all scientists, the crystallographers, to say, and so on.

Another interesting question concerning sound waves is the following: What happens if the wavelength in a solid gets shorter, and shorter, and shorter? Thus shock wave gets? It is interesting that it cannot get any shorter than the space $\lambda = \frac{c}{f}$.

between the atoms. Because "there is supposed to be a wave in which one point goes up and another goes down," the energy "available wavelength" is clearly the atom spacing. In terms of the modes of oscillation, we say that there are longitudinal modes, and transverse modes, long waves, called shear waves, etc. As we consider various types of propagation, the spacing between the atoms, even the speed are no longer constant; there is a dispersion effect where the velocity is not independent of the wave number. But, if a wave, the higher mode of vibration always would be that in which every atom is doing the opposite of neighboring atoms.

Now from the point of view of seismic, vibration is like the two pendulums that we were talking about, for which there are two modes, one in which they both go together, and the other in which they are apart. It is possible to imagine the longitudinal wave in a very, in terms of a system of coupled harmonic oscillators, like a chain of mass-spring pendulums, with the highest mode such that they oscillate together, and lower modes with different characteristics of the timing.

The shortest wavelengths are so short that they are not usually available technically. However they are of great interest because, in the theory of thermodynamics, for solid, the heat properties of a solid, for example specific heats, can be analyzed in terms of the properties of the various sound waves. During all the waves of sound waves of ever shorter wavelength, one necessarily comes to the individual "intrinsic" frequencies that we think are the most ultimate.

A very interesting example of seismic waves in a solid, both longitudinal and transverse, is the waves that are in the solid earth. What makes the waves we do not know, but since the earth, from time to time, has some transverse waves rock slides, just to the other rock. That is like a little noise. The waves that are caused by rock slides are something very much longer in wavelength than usually considered seismic waves. Some of them are transverse, and they travel much in the earth. The earth is not homogeneous, however, and the properties of pressure, density, compressibility, and so on, change with depth, and therefore the speed varies with depth. Then the waves do not travel in straight lines—there is kind of like a refraction, and they go in curves. The longitudinal waves and the transverse waves have different speeds, so they are different at different depths in the earth. Therefore if we place a seismograph or some longitudinal waves, we see the longitudinal to the first moment of quake somewhere else, then we do not just get it immediately. We might get a jiggling, and a ground-shake, and then another jiggling, and an instant of seismic when the longitudinal. If it were close enough, we would first receive longitudinal waves from the disturbance, and then, a few moments later, transverse waves, because they travel more slowly. By measuring the time difference between the two, we can tell how far away the earthquake is, and we know enough about the speeds and properties of the interior regions involved.

An example of the seismic pattern of waves in the earth is shown in Fig. 21-6. The outbreak of waves is triggered off by different methods. If there were an earth quake at the place marked "source," the transverse waves and longitudinal waves would arrive at different times at the station by the solid lines. But, there would also be reflections at the boundary, reflecting waves back, and there would also be reflections at the boundary, reflecting waves back. In, in general, there is a source in the earth which does not carry transverse waves. If the source is supposed to consist, transverse waves still arrive, but the timing is not right. What happens is that the transverse wave comes in the same, and when it is reflected from some surface, it is reflected, between the boundaries, two new waves are generated, one transverse and one longitudinal. But inside the core of the earth, a transverse wave is not propagated (at least, there is no evidence for it), only the longitudinal wave, it comes out again in both forms and arrives in the station.

This incident behavior of these earthquake waves has not been determined. But, however, we can say that a propagating wave in the inner circle. This means that the outer shell of the earth is opaque in the sense that it is not propagating transverse waves. The only way to know what is inside the earth is by studying earthquakes and, by using a large number of observations of many seismographs at different

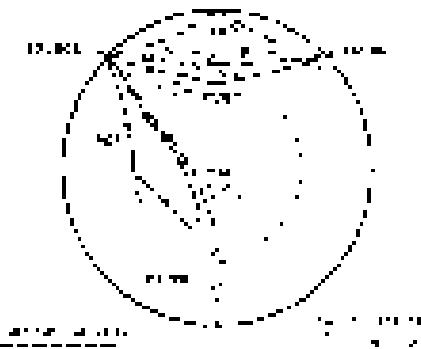


Fig. 21-6. Structure of the earth, showing paths of longitudinal and transverse seismic waves.

and one that has been worked out—the speed. We can do this all along. We know what the speeds of various kinds of waves are in any crystal. Knowing that, therefore, it is possible to figure out what the normal modes of the crystal are. Because we know the speed of propagation of sound waves—in other words, the elastic properties of both kinds of waves in every crystal. Suppose the waves were scattered like an elliptical and let us. It is just a matter of superimposing waves travelling out and in. In principle we determine the individual strings in a framework. We have figured out that if there is a line, hence, between two or more nodes, from the lowest, which is ellipsoidal, to higher modes with more structure.



Fig. 31-7. Power ratio spectrum as detected at seismographs in Chile, Peru, and Antioquia, Colombia. The coherence is a measure of the coupling between the oscillations. (From Van der Pol, Proc. Roy. Soc. (London), A, 200, 605 (1961)).

The Chilean earthquake of May 1960 made a good enough "model" for the signals were simple and not many others, and now when you have a good definition made just right to determine the frequency of the fundamental modes of the earth, to compare them with values that were calculated from the theory of sound with the known velocities, as assumed for the longitudinal waves. The result of this experiment is illustrated in Fig. 31-8, where a plot of the strength of the signal versus the frequency of its oscillation is Fourier analyzed. Note that in each particular frequency, there is much more being received than at other frequencies, the ones very definite numbers. These are the natural frequencies of the earth. Even so, there are two main frequencies at which the earth is oscillating. In other words, if we take a section of the earth, there would be many different modes, we would expect to obtain, in each section, irregular beatings which indicate a superposition of many frequencies. If we analyze this in terms of frequencies, we should be able to find the characteristic frequencies of the earth. The vertical axis in the figure is the superposed frequencies, and we find two characteristic agreement by agreement are to the fact that the theory of sound is right for the inside of the earth.

A very curious point is revealed in Fig. 31-8, which shows a very careful measurement, with below resolution of the lowest mode, on a large range of the earth. Note that it is very simple minimum, but it double set, 7.7 minutes and 37.1 minutes—slightly different. The reason for the two different frequencies was not known at the time that it was measured; either it may have been beyond the time limit. There are at least two possible explanations. One would be that there may be asymmetry in the earth's distribution, which would result in two different modes. Another possibility, which is even more interesting, is this: Imagine the waves going round the earth in two directions from the source. The speeds will not be equal because of effects of the rotation of the earth in the equations of motion, which does not mean that they are out of step during the analysis. Motion in a rotating system is modified by Coriolis forces, and they may cause the observed splitting.

Repeating the method by other cases quakes have set up already, which is obtained on the seismograph is not a curve of amplitude as a function of frequency but, disconcerting as it looks at first, is very irregular. Instead, to had the amount of all the different size waves for all different frequencies, we know that the total ω is roughly the sum by a sine wave of a given frequency and intensity, i.e., average \bar{A} , and in the average all other frequencies disappear. The results were thus, peaks of the integrals (sums) when the data were multiplied by one waves of different waves per minute and integrated.

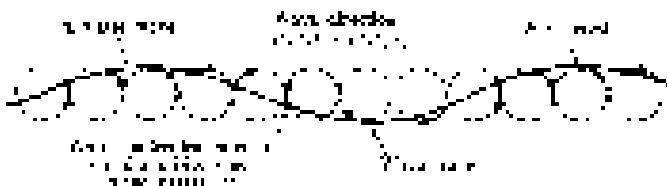


Fig. 31-8. Digital computer analysis of one of the Wadsworth records, showing spectral doublet.

51-4 Surface waves

Now, the next waves of interest, that is easily seen by everyone and which everybody sees or has ample of water in elementary courses, are **wave waves**. As we shall soon see, they are the worst possible example, because they are in no respects like sound and light. They have all the complications that waves can have. For instance w₁ large water waves in deep water. If one ocean is considered, it is very deep and a disturbance is made on the surface, waves are generated. All kinds of irregular motions occur, but the simplest type motion, with a very small disturbance, trips back like the common elastic wave waves coming in toward the shore. Now with such a wave, the water, of course, on the average, is standing still, but the wave moves. What is this motion, is it transverse or longitudinal? It may be either; it is not transverse, nor is it longitudinal. Although the water at a given place is alternately thrown up, it cannot simply harmonizing up and down, by the conservation of energy. That is, if it goes down, water is too water going to go up. Water is essentially incompressible. The speed of compression of waves, that is, sound in the water, is much, much higher, and we are not considering that now. Since water is incompressible on this scale, as a hill comes down the water must move away from the region. What actually happens is that particles of water, etc., the surface moves approximately in circles. When smooth wheels are turning, a person looking in a tire can look at a nearby object and see it going in a circle. So it is a mixture of longitudinal and transverse, breaking in the confusion. At greater depth in the water the motions are smaller circles until, reasonably far down, there is nothing left of the motion (Fig. 51-9).

Fig. 51-9. Deep-water waves, now formed from particles moving in circles. Note the systematic phase shift from wave to wave.



To find the velocity of such waves is an interesting problem: It must be some combination of the density of the water, the acceleration of gravity, which is the restoring force, that makes the waves, and possibly of the wavelength and of the depth. If we take the case where the depth goes to infinity, it will, in some sense depend on the depth. Whatever formula we are going to get for the velocity of the surface of the waves must combine the various factors to make the proper discussion, and if we try this in various ways, we find only one way to combine the density, ρ , and g in order to make a velocity, namely, $\sqrt{\rho g}$, which does not include the density at all. Actually, this formula for the phase velocity is not exactly right, but a complete analysis of the dynamics, which we will not pursue, shows that the formula we do have them, except for $\sqrt{\rho g}$:

$$v_{phase} = \sqrt{gk}/2\pi \text{ (for gravity waves).}$$

It is interesting that as long waves go faster than the short waves. This is about right, but not, however, there is some apparent contradiction, travelling by, this is not so, where the waves come to shore with slow sloshing at first and then more and more rapid sloshings, because the short waves travel longer. The waves get shorter and shorter as the time goes on, because the velocities grow in proportion to the wavelength.

One may object, "That is not right, we must look at the group velocity in order to figure it out." Of course this is true. The formula for the phase velocity does not tell us what is going to arrive first: what tells us is the group velocity. So we have to work out the group velocity, and it is left as a problem to show that, in fact, of the phase velocity, assuming that the velocity goes as the square root of the wavelength, which is all that is needed. The group velocity also goes as the square root of the wavelength. How can the group velocity go half as fast as the phase? If one looks at the bunch of waves that are made by a boat travelling

along. Moving a particular wave, he finds the wave front in the group and gravity gets weaker and distorted in the front, and magnetically the waves have a weak one in the back which is now lower than the rest of the magnet. In short, the waves are moving through. In a magnetized plasma, the group is only moving at half the speed that the waves are moving.

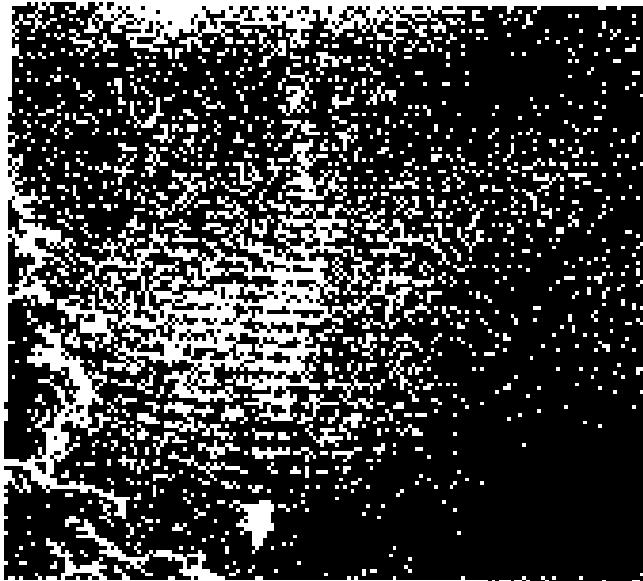


Fig. 51-10. The wake of a boat.

Because the group velocities and phase velocities are not equal, there are waves that are produced by an object moving through air no longer almost a wave, but it is really more interesting. We can see this in Fig. 51-10, which shows the waves produced by a object moving through the water. Now, too, it is quite different than what we expect from for sound, in which the velocity is independent of wave length, where we would have wanted to calculate the wave amplitude by summing them, we know waves in the water with fronts moving parallel to the motion of the boat, and then we have little waves on the sides of the boat. This is the pattern of waves coming, with amplitude, for example, by knowing easily that for the phase velocity is proportional to the square root of the wavelength. For small k , that is just the pattern of waves is stationary relative to the (constant) velocity; but, for other patterns would get lost from the boat.

The water waves that we have been considering so far are long waves in which the large of restoration is due to gravitation. But when waves get very short in the water, the main restoring force is capillary attraction, i.e., the energy of the surface, the surface tension. For capillary tension waves, it turns out that the phase velocity is

$$v_{\text{phase}} = \sqrt{\frac{2\pi}{\rho g}} \cdot \frac{1}{k} \quad (\text{capillaries})$$

where ρ is the surface tension and g the density. If k is the wave, opposite the phase velocity is higher, the shorter the wavelength, when the restoring force very small. When we have both gravity and capillary action, as we always do, we get the cancellation of the two together,

$$v_{\text{phase}} = \sqrt{\frac{2\pi}{\rho g} + \frac{1}{\rho k^2}}$$

where $k = 2\pi/l$ is the wave number. So the velocity of the waves which are really going along is ... The phase velocity as a function of the wave length is shown in Fig. 51-11; for very long waves it is fast, for very long waves it is fast and there is a minimum wavelength at which the waves stop going. The group velocity is to be calculated from the formula: it goes to λ the phase velocity too, up to the phase velocity for a given wave. To the left of the minimum the group velocity is 1, plus when the phase velocity goes to the right, the group velocity is less than or

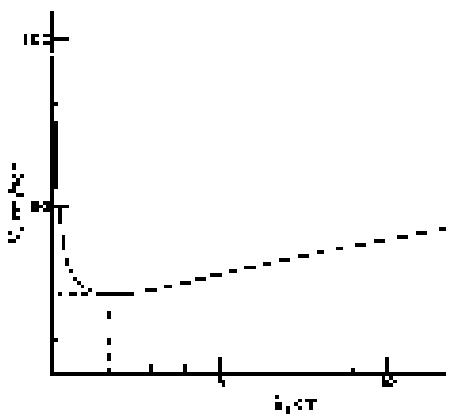


Fig. 51-11. Phase velocity vs. wave length in water.

phase velocity. There are a number of interesting phenomena associated with these facts. In the first place, since the group velocity is increasing as rapidly as the wavelength goes down, it can make a difference that will be a thousand fold in the dispersion going at the minimum, equal with the corresponding wavelength, and then it follows, going at a higher speed, will be a short wave and a very long wave. It is easy to see the long waves, but it is easy to see the difference in a water tank.

So we see that the ripples which result in shallow simple waves are quite interesting and complicated; they do not have a sharp wave front. At all, we are rather like simple waves like sound and light. The ripples are like little ripples which run out away. A sharp discontinuity in the wave does not produce a sharp wave because of the dispersion. This comes the very fine waves. Naturally, if an object moves through the water at a certain speed, a rather complicated pattern results, because of the different waves it is going at different speeds. One can demonstrate this very simply in water and see that the waves are not the true equilibrum waves. There are shallow waves, of a certain kind, which go forward by touching the bottom. One sees that when the depth h is large, the speed is lower. The wave moves in an angle; the rate of maximum slope it tends and tends to follow that line. In this way one can show various things, and we can only say waves are more complicated in water than in air.

The speed of long waves in water with no translation motion is slower when the depth is less, lesser in deep water. That is, waves come toward a beach when the depth is great, the waves go slower. But when the water is deeper, the waves go faster, so we get the effects of break waves. This runs, since the wave is not so simple, the sheets are much more scattered, and the wave overcomes itself, in the familiar way shown in Fig. 51-12. This is what happens when waves encounter the shore, and the next chapter's lecture will reveal to such circumstances. Nature has yet been able to figure out what shape the wave should take just before it breaks. It is easy enough when the waves are small, but when the sea gets large and winds, then it is a difficult problem.



Fig. 51-12. A Water Wave.

An interesting feature about equilibrum waves can be seen in the disturbances made by an object moving through the water. From the point of view of the object itself, the water is flowing past, and the waves which ultimately catch up with it are always the waves which have had the right speed to stay still with the object in the water. Similarly, consider an object in a stream, with the current flowing by, the pattern of waves is stationary, and at least the right wavelength to go at the same speed as the water goes up the. But if the group velocity is less than the phase velocity, then the C-tube has group velocity going backwards in the stream, because the group velocity is not quite enough to keep up with the stream. If the group velocity is less than the velocity of the plane, the pattern of waves will appear in front of the object. If one walks directly across just in a stream, one can see that there are little ripples in front and long "valleys" in the back.

Another interesting feature of this solution can be observed in pouring liquids from a funnel held through air or a bottle. For instance a large number of waves can be seen crossing both ways in the outgoing stream. They are wave scattering along the surface at the edges, and running out much like the waves about an object in a stream. There are effects from both sides which produce the crossed patterns.

We have investigated some of the interesting properties of waves, and the very many complications of dependence of phase velocity on wavelength, the speed of two waves on each other and so forth, that produce the really complex, and therefore interesting, phenomena of nature.

Symmetry in Physical Laws

52-1 Symmetry operations

The subject of this chapter is what we may call symmetry in physical laws. We have already discussed certain features of symmetry in physical laws in connection with wave analysis (Chapter 1), the theory of relativity (Chapters 16), and rotation (Chapter 26).

Why should we be concerned with symmetry? In the first place, symmetry is fascinating to the human mind, and we find objects or patterns that are in some way symmetrical to be appealing. It is no accident that most of our children attain levels of symmetry in the mirror we find in the world around us. Perhaps the most symmetrical things imaginable are spheres, and we are full of spheres—sun, planets, water droplets in clouds. The crystals found in rocks exhibit many different kinds of symmetry, the study of which tells us some important things about the structure of solids. Even the simplest and very early works done concerning symmetry, although the symmetry of a flower or of a bee is not as perfect as we sometimes think, is full of insights.

But our main concern here is not with the beauty of the objects of nature and their symmetries. Rather, we wish to examine some of the even more remarkable symmetries of the universe—the symmetries that exist in the basic laws themselves which govern the operation of the physical world.

First, what is symmetry? How can a physical law be "symmetrical"? The problem of defining symmetry is an interesting one and we have already noted that West gave a good definition, the substance of which is that a thing is symmetrical if there is something we can do to it so that after we have done it, it looks the same as it did before. For example, a symmetrical equation of motion will look the same if we reflect in time; it will look the same as it did before. The question we wish to consider here is what we can do to physical phenomena, or to a physical situation *at the same instant*, and yet leave them the same. A list of the known operations which various physical phenomena remain invariant is shown in Table 52-1.

52-2 Symmetry in space and time

The first thing we might try would, for example, be to translate the phenomenon in space. If we do an experiment in a certain region, and then build another apparatus at another place in space for doing the same sort every time, whatever went on in one apparatus, in a certain order in time, will occur in the same way if we have arranged the same conditions, with all care according to the restrictions that we mentioned before. But all of these features of the environment which make it act the same way have also been "fixed" here—we could always move ... define how much we should include in these circumstances, and we shall not go into those details again.

In the same way, we can do better today than yesterday or later, or we can do worse. (That is as far as we know today)—all of these things are as far as we know today.) That means that, if we're able to do an experiment and start it at a certain time, say on Tuesday at 10:00 a.m., and then build the same apparatus and start it at, say, three days later, in the same location, the two experiments will go through the same motions in nearly the same way as a function of time, no matter what the starting time, provided again, of course, that the relevant features of the environment, and so on, don't change significantly. That's symmetry because,

52-1 Symmetry operations

52-2 Symmetry in space and time

52-3 Symmetry and conservation laws

52-4 Mirror reflections

52-5 Uniform and axial rotation

52-6 Which hand is right?

52-7 Parity is not conserved!

52-8 Antimatter

52-9 Broken symmetries

Table 52-1

Symmetry Operations

Translators in space

Translators in time

Rotation through fixed angle

Uniform velocity (in space
(the distance transformation)

Creation of time

Reflection of space

Interchange of identical atoms
or identical particles

Quantum mechanical phase

Mass antimatter (charge conjugation)

of course, cost of one bought. General Motors check those numbers up, the same thing would happen to us if he bought it now!

We have to watch out for geographical differences too. On the one, of course, rotations in the characteristics of the earth as far. So, for example, if we rotate the magnetic field in a certain region and move the apparatus to some other region, it may not work in precisely the same way because the magnetic field is different, but we say that is because the magnetic field is associated with the earth. We can imagine that "we turn the whole system upside down, it would make no difference in the operation of the apparatus."

Another thing that we discussed in considerable detail was rotation in space: if we turn an object round an angle it works just as well, provided we turn everything else that is relevant along with it. In fact we discussed the problem of symmetry under rotation in space in considerable detail in Chapter 11, and we invented a mathematical system called tensor analysis to handle it as neatly as possible.

On a more advanced level we had another symmetry—the symmetry under uniform velocity in a straight line. That is, in regard to the remarkable effect that if we leave a piece of apparatus working a certain way and then take the same apparatus and put it in a car, and move the whole car, plus all the relevant circumstances, at a uniform velocity in a straight line, then to keep the phenomena inside intact we must make no difference; at the basis of this is again the same. We even know how to express this more technically, and that is that the mathematical equations of the physical laws must be unchanged under a Lorentz transformation. As a matter of fact, it was a result of the relativity problem that unconnected physical entities can apply in symmetry in physical laws.

Now the above mentioned symmetries have all been *x*-symmetries, i.e., in time and space being more or less the same, but there are other symmetries of a different kind. For example, there is a symmetry which describes the fact that we can replace one atom by another of the same kind. To put it differently, there are atoms of the same kind. It is possible to hold a sample of some particular sort of ion-charges *a* *plus*, it makes no difference—the atoms are *identical*. Wherever one atom of oxygen of a certain type will go, another atom of oxygen of that type will go. One may say, "That is ridiculous, that is the definition of equal types." That may be merely the definition, but then we still do not know whether there are many "atoms of the same type"; the fact is that there are many, many atoms of the same type. Thus it does mean something to say that "no difference if we replace one atom by another of the same type." The so-called elementary particles of which the atoms are made are also identical, and also in the above sense—all electrons are *the same*; all protons are *the same*; all positive pions are *the same*; and so on.

After such a long list of things that can be done without changing the phenomenon, one might think we could do practically anything; so let us give some examples of the contrary, just to see the difference. Suppose that we ask: "What the physical laws symmetric under a change of size?" Suppose we build a certain piece of apparatus, and then build another apparatus five times bigger; is every part will it work exactly the same way? The answer is, in this case, no! The wavelength of light emitted, for example by the atoms inside, one hole of sodium atoms and the wavelength of light emitted by a gas of sodium atoms the same in volume is not five times longer, but is, in fact, exactly the same as the older. So the ratio of the wavelength to the size of the emitter will change.

An other example: we see in the newspaper, every once in a while, pictures of a given correspondence with real materials, e.g., sometimes such a set by some noted letter etc. keeps going mathematical figures. They much more elaborate and wonderful than any real correspond. If we imagine that this wooden cathedral were suddenly built on the scale of a m^3 as usual, we see where the building is—it would not last—the whole thing would collapse because of the enormous strength involved and just not strong enough. "Yes," one might say, "but we also know that water exerts an influence from the outside—this must be changed in proportion!" We are talking about the ability of the object to withstand gravitation. So what we should do is first to take the model cathedral of real materials and

the nuclear, and then we know it is stable. Then we should take the finger out of the hole and have a bigger error. But keep it at exact zero, here [sic] the gravitation is increased still more!

Today, of course, we might extend the last one, phenomena depending on the weak on the grounds that matter is always in contact, and certainly if we break an egg, the fact that we can't there are only five atoms in it. It would be really something you could never compare down arbitrarily. The size of an individual atom is not at all arbitrary—it is quite definite.

The fact that the laws of physics are not unchanged under a change of scale was discovered by Galileo. He realized that the strength of materials is constant in exactly the right proportion to their sizes, and he illustrated this property that we were just discussing, about the strength of materials, by showing, for instance, the bone of one dog, in the right proportion for holding up his weight, and the imaginary bone of a "super dog" that would be, say, ten or a hundred times bigger—that bone was a big solid thing with give it. To cut you a long story, we do not know whether he ever argued the argument similar to the one given that the laws of nature must have a definite scale, but he was so impressed with this discovery that he considered it to be so important in the character of the laws of motion, because he published from time in the same volume, called "In Two New Sciences."

Another example in which the laws are not symmetric etc., that we know quite well, is that a system in rotation at a uniform angular velocity does not give the same apparent laws as one that is not rotating. If we make an experiment and take our apparatus, a soccer ball and have the ground spinning it, carry a particle at a constant angular velocity, the apparatus will not work the same way however, as we know, things inside the apparatus will be drawn to the outside, and so on, by the centripetal or centrifugal forces, etc. In fact, we can tell that there is rotation, by using a so-called Foucault pendulum, without looking outside.

Next we mention a very interesting symmetry which is obviously false, i.e. reversibility in time. The obvious laws apparently except by symmetry in time, because, as we know, all obvious phenomena are irreversible on a large scale: "The moving finger writes, and having writ, moves on." So far as we can tell, this irreversibility is due to the very large number of particles involved, and if we could see the individual molecules, we would not be able to see which in the machinery was working forward or backwards. To make it more precise we make a *moving* picture in which we know what all the atoms are doing, in which we can see them jiggling. Now we have another appearance like it, but which can be made, at the final condition of the other one, with all the velocities perfectly reversed. If not then you through the same medium, but exactly in reverse. Building a molecular movie if we take a moving picture, at sufficient detail, of all the inner motion of a piece of material and show it on a screen and put it backwards the physician will be able to say, "That is against the laws of physics, that's doing something wrong." If we do the same in the other, of course, he can tell well by perfectly clear. If we see the egg shattering on the sidewalk and the shell cracking open, and so on, then we say "simply no." That is impossible. Because if we put the moving picture backwards the egg will collect together and the shell will go back together, and that is obviously ridiculous! But if we look at the molecular atoms themselves, the laws look completely reversible. This is, "comes, a man," under directly to have made, but apparently it is true that the fundamental physical laws, on a microscopic and fundamental level, are completely reversible—true!

52-3 Symmetries and conservation laws

The symmetries of the physical laws are very interesting at this level, but only just as in the end, do we ever need to be using and applying when we come to quantum mechanics. There is reason when we cannot make clear at the level of the present discussion, in that the most physicists still find somewhat surprising, a most profound and basic thing is that in quantum mechanics, for every symmetry of symmetry there is a corresponding conservation law; there is a definite

connection between the two of conservation and the symmetries of physical laws. We can only conclude that at present, without any attempt at explanation.

In fact, for one, plus the two laws are symmetrical but translation is static when we add the p insights of quantum mechanics, there can be no such connection or symmetry.

But the laws of symmetries under translation in time means, in quantum mechanics, that energy is conserved.

Translational invariance through a fixed angle is space corresponds to the conservation of angular momentum. These connections are very interesting and beautiful things, among the most beautiful and profound things in physics.

Incidentally, there are a number of symmetries which appear in quantum mechanics which have no classical analog, which I use in another of describing it, classical physics. One of these is as follows. If ψ is the amplitude for some process or other, we know that the absolute square of ψ is the probability that the process will occur. Now if someone else were to make his calculations, not with this ψ , but with ψ' which differs merely by a change in phase: that is to be some constant, and multiply ψ' times the old ψ , the absolute square of ψ' , which is the probability of the event, is then equal to the absolute square of ψ .

$$|\psi'|^2 = |\psi|^2; \quad |\psi'|^2 = |\psi|^2. \quad (52.1)$$

Therefore, the physical law is unchanged if the phase of the wave function is shifted by an arbitrary constant. That is another symmetry. Physical laws must be of such a nature that a shift in the quantum-mechanics phase makes no difference. As we have mentioned in quantum mechanics, there is a renormalization law for this symmetry. The conservation law which is common to all the quantum-mechanical phase seems to be the conservation of electrical charge. This is together a very interesting business.

52-4 Mirror Reflections

Now the next question which is going to concern us for most of the rest of this chapter, is the question of symmetry under reflection in space. The question is this: Are the physical laws symmetrical under reflection? We may put it this way. Suppose we hold a piece of equipment, let us say a clock, with lots of wheels and cogs and so forth, in a box. It works, and it goes ticking round and inside. We look at the clock in the mirror. Now a clock in the mirror is not the question. But let us actually build another clock which is exactly the same as the first clock looks in the mirror. Now if there is a gear with a right-hand thread in one, we also have with a left-hand thread in the corresponding place of the other, where one is marked "2" on the face we mark a "5" on the face of the other. Each cogged spring is twisted one way in one clock and the other way in the mirror-image clock: when we are all finished, we have two clocks, both physical, which have in them other the relation of a mirror and its mirror image, although they are both actual, material objects we emphasize. Now the question is: If the two clocks are started in the same condition, do they go in corresponding right motion, will the two clocks tick side by side, forever after, as exact mirror images? (This is a physical question, zero-dimensional question). Our intuition about the laws of physics would suggest that they would.

We would suggest that, first of all, in the case of these clock reflections in space, is one of the symmetries of physical laws, that if we change everything from "right" to "left" and leave it otherwise. To some, we cannot tell the difference. Let us, then, suppose it is a comment that this is true. If it is true, then it would be impossible to distinguish "right" and "left" by any physical phenomenon, just as it is, for example, impossible to define a "universal" positive velocity by a physical phenomenon. So it should be impossible, by any physical phenomenon, to define absolutely what we mean by "right" as opposed to "left," because the physical laws should be symmetrical.

Of course, the world does not have to be symmetrical. For example, using what we may call "geography," such "right" can be defined. For instance, we stand

to New Orleans and look at Chicago, and Florida is to our right (where our feet are on the ground). Surprised? Right" and "left" by geography. Of course, the same situation in our system does not have to be. We can say that we are talking about it; it is a creation of mine, the human imagination—in other words, which is, is where the person has to have a space like the earth with "left-handed air" on it and a person like ourselves standing looking at a city like Chicago from a side like New Orleans, but with everything the other way around, so Florida is on the other side. I think it seems sort of impossible, yet again, the person now has to have everything changed left for right.

Another point is that our definition of "light" should not depend on history. An easy way to distinguish right from left is to go to a bad neighborhood pick up a bunch of sand. The odds are it has a right-hand twist—not necessarily, but it is much more likely to have a right-hand twist than a left-hand one. This is a question of history or convenience, or the fact things happen to be, and is again not a question of fundamental laws. As we can well appreciate, everyone could have started out making left-handed screws!

So we must try to find some polarization in which "right" and "left" is intended fundamentally. The most possibility we have is the fact that polarized light rotates the plane of polarization as it goes through, say, sugar water. As we saw in Chapter 29, it rotates, let us say, to the right, a certain amount of them. This is sort of defining "left-hand" because we may dissolve some sugar in the water and then the polarization goes to the right. But sugar has no living things, and if we try to make it, say, artificially, then we discover that after you dissolve the plane of polarization. But if we then take that same sugar which is made artificially and which does not rotate the plane of polarization, and put it back and at it take out some of the sugar and then blow it in the heat lamp, we find that we still have sugar left (so much less sugar as we had before), and this time it does rotate the plane of polarization, but the other way. It seems very convincing, but is really explained.



Fig. 32-1. L-L Alanine (left) and D-alanine (right).

Take another example. One of the substances which is common in all living creatures and that is fine-surfaced is the lipoprotein. Proteins consist of chains of amino acids. Figure 32-1 shows a model of an amino acid like, amino acid L-alanine. This form of acid is called L-alanine, and the molecular arrangement would look like that in Fig. 32-1, if it came out of a protein of a real living thing. On the other hand, if we try to make alanine from carbon dioxide, chlorine, and some kind we can make it, it is not a complicated molecule; we discover that we are making right, mirror image of the molecule and the one shown in Fig. 32-1 (left). The first molecule, the one that comes from the living thing, is called *D-alanine*. The *D*-alanine, which is the same chemically, in that it has the same kinds of atoms and the same connections of the atoms, is a "right-hand" molecule, compared with the "left-hand" L-alanine and it is called *D-alanine*. The interesting thing is that when we make alanine at home in a laboratory, more simple process, we get an equal mixture of both kinds. However, the only thing that life uses is L-alanine. (This is not exactly true. Here and there in living creatures there is a special use for D-alanine, but it is very rare. All proteins use L-alanine exclusively.) Now if we make both kinds, and we find that this is the same animal which likes to "eat" our right hands, it cannot eat D-alanine, so it only uses the L-alanine, and is well accustomed to the "right" after the last time. In short, that makes

well for the *s*, only the "wrong" kind is well. β -D-glucose sugar tastes sweet but not the same as right-handed sugar.

So it looks as though the whole game of "left" versus "right" distinction between " left " and " right " in chemistry permits a discrimination, because the two molecules are chemically different—but no, it does not! So far as physical measurements can be made, such as cf energy, the rates of chemical reactions, and so on, the two kinds work exactly the same way if we make everything else in a mirror image too. One molecule is well suited to one sugar, and the other will confuse it at the "if" it scatters the same amount, through the same volume of fluid. Thus, so far as physics is concerned, these two amine acids are equally spectroscopic. & that is the unfortunate thing today, the fundamental of the Schrödinger equation however that the two molecules should behave in exactly corresponding ways, so that one is to the right as the other is to the left. Nevertheless, in life, it is all over now!

It is presumed that the lesson here is the following. If we suppose, for example, that life is somehow living matter in a very high tension at which all the proteins in some creatures have left-handed amino acids, and at the same time are imposed—every substance in the living creature is imposed—it is non-symmetrical. So when the living creature comes by its living chemicals to the food from the tree, he would turn his head or swallow "this" into his stomach. But the snake king does not like Cinderella and the prince, except that it is a "left hand" (that we are told up). So far as we know, in principle, we could build a frog, for example, in which every molecule is reversed, everything is to the "left-hand" in the image of a real frog, we have a left-hand frog. This left-hand frog would go on all right for a while, but he would find according to us, because it has "hands over there, his enzymes are not built to digest it. The frog has the wrong "hand" of amino acids unless we give him a left-hand frog. So as far as we know, the chemicals and life processes would continue in the correct manner if everything were reversed.

If there is actually a physical and chemical phenomenon, then we can understand that the proteins are all made in the same way, or you know the idea that all the non-living forms living creatures, by accident, got started and a few more, somewhere, once, one organic molecule was kapooshed in a certain way, and from this *accident* came the "right" happens in nature in our particular perspective; a particular "bias", however was one-sided, and ever since then the "bias" has propagated itself. Once having begun that bias, that is, in now of course, it will always continue—all the enzymes digest the right things, manufacture the right things; when the carbon dioxide and the water vapor and so on, go in the plant leaves, the enzymes that take the oxygen in them happen because the enzymes are kapooshed. Many new kind of virus or living things were to originate at a later time, we would survive only if we could "read" the kind of living matter already present. This is the most of the problem.

There is no conservation of the number of right-handed molecules. Once started, we could keep increasing the number of right-handed molecules. So the presumption is, that the replacement in the course of time shows a loss of symmetry in physical law, but it *fails*, on the contrary, the universal nature and the commonness of electric charge, will exist on our earth in the sense described above.

52.5 Polar and axial vectors

Now we go together. We observe next in discussing the case of light in other places where we have "right" and "left," the situation is a matter of fact when we carried about vector analysis with the right-hand rule; we have to use it in order to get the angular momentum, torque, magnetic field, and so on, all come out right. The case of a vector involving a magnetic field, for example, is $\mathbf{E} \times \mathbf{B}$. In a given situation, in which we knew \mathbf{E} , \mathbf{B} and \mathbf{B} , isn't that sufficient enough to define right-handedness? As a matter of fact, if we go back and look at where the vectors start from, we know that the "right-hand rule" was really a convention; it wasn't true. The original quantities, like the angular momenta and the angular velocities, and charges of this kind, were not really vectors in all. They are al-

motion associated with a surface plane, and it is just because there are three dimensions to space that we can associate the quantity with a direction perpendicular to that plane. Of the two possible directions, we choose the "right-hand" direction.

So if the laws of physics are symmetric, we should find that if some denoted were to break into 2, the physics abides the same. Right" for "left" in every book in which "right-hand rules" are given, and instead we were to use all "left-hand" rules," nothing there would make no difference whatsoever in the physics area.

Let us give an illustration. There are two kinds of vectors. There are "translatory" vectors. For example, see diagram: If in our system of there is a piece of wood moving elsewhere, then its motion component in there will be the image piece Q. The image since being one, and if we draw a vector from the "piece" to the "something else," the vector is the mirror image of the other (Fig. 52-2). The vector arrow changes its head, just as the whole wood turns inside out; such a vector we call a polar vector.

The other kind of vector which has to do with rotations, is of a different nature. For example, suppose that in three dimensions something is rotating as shown in Fig. 52-3. Then unless it is frozen in a mirror, it will be rotating as indicated, namely as the mirror image of the original motion. Now you cannot expect to represent the direction of rotation by the same rule it is a "vector" which, on reflection, does not change around as the polar vector does, but is reversed relative to the polar vector and to the geometry of the space, such a vector is called an axial vector.

Now if the law of reflection symmetry is right in physics, then it must be true that the equations must be unchanged that it is clearly the case of each polar vector and each corresponding axial vector, which would be what corresponds to reflection, nothing will happen. For instance, when we write a formula which says that the angular momentum is $\mathbf{L} = \mathbf{r} \times \mathbf{p}$, that equation is all right, because, if we change to a left-handed coordinate system, we change the sign of \mathbf{L} , and if and we do not change the cross product sign is changed, since we must change from a right-hand rule to a left-hand rule. As another example, we know that the force on a charge moving in a magnetic field is $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$, but, if we change from a right- or a left-handed system, since Faraday's law is to be preserved the sign of \mathbf{F} is also changed, required by the correspondence that we reflected by a sign change in \mathbf{B} . When this is so, \mathbf{B} must be an axial vector. In other words, if we make such a reflection, \mathbf{B} must go to $-\mathbf{B}$. So if we change our coordinates from right to left, we must also change the place of magnets from north to south.

Let us assume that there is no asymmetry. Suppose that we have two magnets as in Fig. 52-4. One is a magnet with the ends pointing toward a certain way, and were placed in a given direction. The other magnet looks like the reflection of the first magnet in a mirror. The coil will wind the other way, everything that happens inside the coil is exactly reversed, and the current goes as shown. Now, from the law for the propagation of magnetic fields, which we do not know yet exactly, but which we most likely learned at high school, it turns out that the magnetic field is as shown in the figure. In one case, the pole is a south magnetic pole, while in the other magnet the current is going the other way, and the magnetic field is reversed—it is a north magnetic pole. So we see that when the problem right about instead magnet from north to south.

Never mind changing $\mathbf{magnetic}$ field; these too are mere conventions. Let us talk about electric fields. Suppose now, that we have an electron moving through one field, going into the page. Then, if we use the formula for the force $\mathbf{F} = q\mathbf{E}$ (remember, the charge is minus), we find that the electron will deviate in the indicated direction according to the physical law. But the phenomenon is that we know a coil with a current going in a specified direction, an electron moves in a certain way. This is the physics, never mind how we label every thing.

Now let us do the same experiment with a mirror: we send an electron through a coil carrying a current and now the force is reversed, if we calculate it here the same rule, and it is very good because the corresponding mirror is the mirror image!



Fig. 52-2. A step in space and its mirror image.

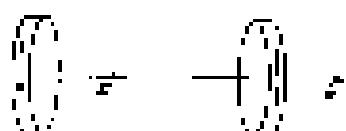


Fig. 52-3. A rotating wheel and its mirror image. Note that the angular velocity "vector" is not reversed in direction.

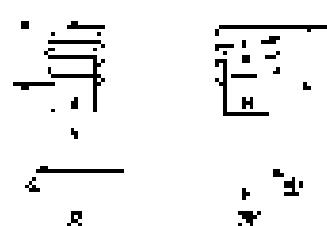


Fig. 52-4. A magnet and its mirror image.

52-6 Which hand is right?

So the fact of the matter is that in studying any phenomenon there are always two right-hand rules, or an even number of them, and the net result is that the phenomena always looks symmetrical. In your receiver, we can see left-right from left if we turn it around with north pointing south. However, it may seem that we can feel the more pull of a magnet. The north pole of a compass needle, for example, is one that prefers to be north. This is of course not a physical property that has to do with geography of the earth; that's just the talk-up about it which makes us Chicago, so it does not count. If we have seen compass needles we may have noticed that the north-seeking pole is a sort of "visible" pole. But that is just due to the man who painted the magnet. These are all local, comparative effects.

However, if a magnet were to have no property, that if we looked at it closely enough we would see same bars pointing in its north pole but not on its south pole, if that were the general rule, or if there were no unique way to distinguish the north from the south pole of a magnet, then we would tell which of the two poles was actually the , and that would be the end of the law of reflective symmetry.

To illustrate this whole problem of irreversibility, imagine that we are telling two Indians, or someone very far away, by telephone. We cannot allow us and him any visual contact to inspect; so, I suppose if we can't send light, we could send some right-hand or rather polarized right and left, "This is right and light just watch the way it is going," but we cannot give him anything, we can only tell him. He is far away, or in some strange location, and he cannot see anything we can say. For instance, we cannot say, "Look at this magnet; now see? Two thumb-screws are arranged. What we mean by right is . . ." We are only allowed to describe him.

Now we want to tell him all about it. Of course, first we can define my thumb-screw, and say, "This, tick, tick, too, too, tick, tick, tick, here . . ." so that gradually he can understand a couple of words, and so on. After a while we may become very familiar with this fellow and he says, "What do you guys think?" He can't visualize correctly, and say, "Well, we are six feet tall." He says, "Wait a minute, what is she fast?" Is it possible to tell him what six feet is? Certainly! We say, "You know about the number of hydrogen atoms we are: 7,000,000,000 hydrogen atoms high?" That is possible because hydrogen atoms are not scattered hither and thither, and therefore, we can define an absolute length. And so we define the size of the body, and tell him what the general shape is—the two prongs were the arms sticking out on the ends, and so on, and he follows us along, but we finally describe no motion on the outside, presumably without ever under any particle-to-particle collision. He is even making a model of a screwdriver—says, "Say, you are definitely very familiar with it now, now what is on the inside?" So we start to describe the various parts on the inside, and we come to the heart and we carefully describe the shape of it, and say, "Now put the heart on the left side!" He says, "Which is the left side?" Now our problem is to describe to him which side the heart goes on without his ever seeing anything. Just we see, and we hear him ever sending any sample to him of what we mean by "right" the standard right-handed action. Can we do it?

52-7 Finally & not answered:

It turns out that the laws of gravitation, the laws of electricity and magnetism, nuclear forces, all or by the principles of reflection, are all these laws, or anything derived from them, cannot be used. But we know about the many particles that are found . . . among them is a phenomenon called beta decay, or weak decay. One of the examples of weak decay, in connection with a part of the material in about 1937, posed a strange puzzle. There was a certain charged particle which disintegrated into three neutrinos as shown schematically in Fig. 52-5. This particle was called, for a while, a π -meson. Now in Fig. 52-5 we also see another particle which disintegrates into two neutrinos, the third component, from the π -meson.

spectrum of energy. The particle was called a π -meson. So on the one hand we have a particle called π , which disintegrates into three π 's, and on the other it disintegrates into two π 's. Now it was soon discovered that the π 's and the π 's are almost equal in mass, in fact, rather than experimental error, they are equal. Next, the length of time it took for them to disintegrate into three π 's and two π 's was found to be almost exactly the same; they live the same length of time. Next, whenever they were made, they were since at the same proportion, say, 14 percent to 86 percent of.

Any one of his higher mind realizes immediately that they must be the same particle, that we merely produce an object which has two different ways of disintegrating—two very different possibilities. This object, too, can disintegrate in two different ways has, therefore, the same lifetime and the same disintegration ratio. Because this is simply the ratio of the odds with which it disintegrates into these two kinds.

However, it was possible to prove (and we cannot here explain it in detail) from the principle of reflection symmetry in quantum mechanics that it was impossible to have two such forms from the same particle—the same particle could not disintegrate in both of these ways. The observations are corresponding to the principle of reflection symmetry is something which has no classical analog, and so this kind of quantum-mechanical conservation law is called the conservation of parity. So, it was a result of the conservation of parity or, more precisely, from the symmetry of the quantum-mechanical equations of the weak decay under reflection, that the same particle could not go one way, so it must be some kind of antiparticle. It does, however, and we do. But, the particle was studied, the more remarkable it is to find that, and the suspicion gradually grew that possibly the deep law of reflection symmetry of nature may be false.

As a result of this apparent trouble, the physicists Fermi and Tamm suggested that other experiments be done in metal crystals to try to see whether the law was correct in other cases. The first such experiment was carried out by Miss Wu from Columbia, and was done in Cobalt. Using a very strong magnet at a very low temperature, it turns out that a certain isotope of cobalt which has electrons by colliding, an electron, is magnetic, and if the temperature is low enough and the thermal oscillations do not jolt the atomic magnets about too much, they line up in the magnetic field. As the cobalt atoms will all line up in this strong field, they then disintegrate with polarization, and it was discovered that when the atoms were lined up in a field where it would polarize upward most of the electrons were emitted in a downward direction.

It was, in my "lisp" to the world, such a result that did not yield anything of significance, but it was apparently a problem and interesting. Many of us could, then, see that as a most dramatic discovery: Were we not cobalt atoms in some arbitrary other magnetic field, something like electrons going down than up. Therefore, we went to put in a current-carrying experiment in a "magnet" which the electric charge would be used up at the opposite direction, they would split their electrons up, not down! No action is ungrammatical. For example, our power failed. The south pole of a magnet is of such a kind that the electrons in a direction of migration will begin to turn in, but, unfortunately, in a clockwise way, the north pole from the south pole.

After this, a lot of other experiments were done: the disintegration of the π in a spin experiment and also new times nowadays the Λ , etc., excited and π , disintegration of Z 's, and many other disintegrations. In fact, in almost all cases where it could be expected, all have been found not to obey reflection symmetry. Fundamentally, the law of reflection symmetry at this level in physics, is broken.

In short, we can tell the world, when to put the heart, we say, "Listen, build yourself a magnet, and put the wire in, and p. Turn it on, and then take some cobalt and lower the temperature. Arrange the experiment so the current goes from the feet to the head. Then the direction in which the current goes through the coils is the direction that goes up on what we call the right and comes up on the left." So it is possible to do, a right and left, now, in doing an experiment of this kind.



Fig. 52-5. A schematic diagram of the disintegration of a n into a p + $e^- + \bar{\nu}$.

There are lots of other features that may have been predicted. For example, it turned out that the spin, the angular momentum of the electron alone before disappearance, is always odd, and after disappearance, is 4 units. The electron carries an angular momentum... and that is known to me now. It is easy to see from this that the electron must carry its spin angular momentum aligned along its direction of motion, the neutrino likewise. So, it looks as though the electron is spinning to the left, and the neutrino also clockwise. In fact, it was calculated right here at Caltech by Bethe and Weizsäcker, that the electron spins really to the left. (There were some other predictions that gave the opposite answer, but they were wrong.)

The next problem, of course, was to find the law of the behavior of parity conservation. What is the rule that tells us how saving the baryon number is going to be? The rule is that it occurs only in every very slow reaction, called weak decays, and when I say "slow," the rule is that the particle which is my syn. has the electron, neutrino, and so on, etc. travel with a spin (length) to the left. That is a important rule. It connects a particle's spin or velocity and an axial vector angular momentum, and says that the angular momentum is uniquely to be opposite to the velocity vector along it.

Now that is the rule, but today we do not really understand the why and the physics of it. Why is this the right rule, who is the "impartial" reason for it, and how is it connected to something else? At the moment we have been unable to be the "why" that the thing is unsymmetrical. But we have not been able to recover enough to understand what it means with regard to the other rules. However, the subject is still活着, maybe, and still unbroken, so it seems appropriate that we discuss some of the questions a moment with it.

52-4 Antiparticles

The last thing is the third one of the symmetries which is to be immediately go back over the list of known or assumed symmetries and see whether any of the others are lost. Now we did not mention the question of our list, which most interestingly is quibbled, and that is the relation between matter and antimatter. Dirac postulated first in addition to electrons there must be another particle, called the positron discovered by Compton by Anderson, that is necessarily related to the electron. All the properties of these two particles obey certain rules of correspondence; the energies are equal, the masses are equal, the charges are equal; but, more important than anything, the two of them, when they come together, can annihilate each other and liberate their entire mass in the form of energy, say $E=mc^2$. The positron is called a conjugate to the electron, and these are the characteristics of a particle and its antiparticle. It was clear from Dirac's argument that all the rest of the particles of the world should also have corresponding antiparticles. For instance, for the proton there should be an antiproton, which is now synthesized by α . They would have a negative total charge, and the same mass as a proton, and so on. The most important feature, however, is that a proton and an antiproton containing bags, but not opposite each other. The reason we emphasize this is that people do not understand it when we say, "There is a proton and a neutron antiproton, because they say, "A neutron is neutral, so how can it have the opposite charge?" The rule of the "rule" is not just that it has the opposite of bags, it has a certain set of properties. One thinks he is talking about opposites. The antiproton is distinguished from the neutron this way: if we bring two neutrons together, they just stay as two neutrons, but if we bring a neutron and an antineutron together, they annihilate each other with a great explosion of energy being liberated, with various pions, etc., bags, and whatever.

Now if we have antineutrons, antiprotons, and antideuterons, we can make a hydrogen, in principle. They have mutual repulsion, but it is possible to accelerate. For instance, a hydrogen atom has a proton at the center with an electron going around it. Now imagine that somewhere we can make an antiproton with a positron going around, would it go around? Well, first of all, the antiproton is electrically negative and the antineutron is electrically positive, so they attract each other in a corresponding manner. One would do all the same everything as the

same. It is one of the principles of the symmetry of physics, the equal right to show, that if a clock, say, were made of matter on one hand, and then we make the same clock out of antimatter, it would run in this way. (Of course, if we put the clocks together, they would annihilate each other, but that is another story.)

An immediate question then arises. We can build, say, in matter two clocks, one which is "left-handed" and one which is "right-handed." For example, we could build a clock which is set up in a simple way, has two cathodes and magnets and electron detectors, which detect the presence of 0-energy electrons and count them, feel time has occurred, the second hand moves next. To the right of each, respectively, there must now will not run on the same time. So evidently we can make two clocks such that the left-handed clock does not agree with the right-handed one. Let us make, out of matter, a clock which has with the standard or "right-hand" clock. You can make, also, out of matter, a clock which you call a "left-hand" clock. We have just discovered that, in general, these two will not run the same way, perhaps last famous physical discovery, it was thought, that they would. Now it was also suggested that, rather than antimatter were equivalent, that is, if we made a standard clock, right-handed, based on β -ray. Then it would run the same as the right-handed matter clock, and if we made the same clock in the left, it would run the same. In other words, at the beginning it was believed that all four of these clocks were the same; but, of course, we know that the right-hand and left-hand matter are not interchangeable. Presumably, therefore, the right-handed antimatter and the left-handed antimatter are not the same.

So the question is, which goes with which, if either? In other words, does the right-handed matter have the same way as the right-handed antimatter? Or does the right-handed matter believe the same as the left-handed antimatter? Perhaps, to get into it using particle decay instead of time analogy, indicate that this is the difference from matter to the "right" versus the same way as gamma to the "left."

Therefore, at long last, it is really true that right and left symmetry is still violated! If we make a left-handed clock, but since it out of the other kind of matter, antimatter, instead of matter, it would run in the same way. So what has happened is that instead of having two independent types of matter-like symmetries, now, if these race go large, here comes a new rule, which says that matter to the right is symmetrical with antimatter to the left.

So our theory is made of antimatter and we give him his own symmetries, this "right" handed model like us, it will, of course, run in the other way around. What would happen if we, after much conversation back and forth, we each leave taught the other to make every thing and we meet halfway in empty space? You have the right side, either on our computers, and so forth, and the two of us come rushing out in opposite hands. Well, if he puts out his left hand, watch it.

21.9 Broken symmetries

The real question is, what sort of rules out of laws which are *not* symmetrical? The most interesting about all of this, I think, is a wide range of important strong consequences—nuclear forces, electrical phenomena, and much more—but give them over a tremendous range of physics, all the laws for these seem to be symmetrized. On the other hand, the last term goes to say, "No, the laws are not symmetrical!" How is it the laws can be Lorentz-symmetrized, yet perfectly non-symmetrized? What shall we make of this? First, do we have any other examples? I mean, in, well, a foot, does a few other examples. For instance, the nuclear part of the force between proton and proton, between neutron and neutron, and between neutron and proton is all exactly the same. There is a symmetry for that. So, for a long time, that we can interchange neutron and proton. But it evidently is not a general symmetry. For the electrical repulsion between protons at a distance does not exist for neutrons. So, is not generally true that we can always replace a proton with a neutron, but only to a good approximation. Why not? Because the nuclei of protons are much stronger than the

electrical forces. So there are "valence" systems also. But we do have examples in other things.

We have, in our minds, a tendency to accept symmetry as some kind of perfection. In fact it is ~~like~~ ~~the~~ old idea of the Greeks that man were perfect, and it was rather terrible to believe that the planetary orbits were not circles, but only nearly circles. The difference between being a circle and being nearly a circle is not a small difference, it is a fundamental difference for us the mind is concerned. There is a sense of perfection and symmetry in a circle that is not there the moment the circle is slightly off—~~not~~ is ~~not~~ end of it—it is no longer symmetrical. Then the question is why it is only nearly a circle—it's a more critical question. The actual nature of the planets, in general, would be elliptical, but during the ages, because of tidal forces, and so on, they have been made more asymmetrical. Now the question is whether we have a similar problem here. The problem from the point of view of the Greeks is if they were perfect circles there would be nothing to explain, that is clearly simple. But since they are only nearly circles, there is a job to explain, and the job turned out to be a big cosmological problem, and now our problem is to explain why they are nearly symmetrical by looking at tidal forces and so on.

So our problem is to explain where symmetry comes from. Why is nature so nearly symmetrical? We have ~~no~~ ~~any~~ idea why. The only thing we might suggest is something like this. There is a pale in Japan, a pale in Neolithic, which is ~~suspected~~ ~~and~~ by the Japanese. In most religions, goes up all Jason; it was built in a time when there was great influence from Chinese art. This is very elaborate, with lots of pillars and beautiful carvings and lots of figures and what not. Heads and personages all over the pillars, and so on. But when one looks closely he sees that in the elaborate and complex design along one of the pillars, one of the ~~say~~ ~~design~~ elements is sexual reproduction; otherwise the thing is completely without material. One asks why this is, the story is that it was sexual appeal: down we that the gods will set the jealousy of the perfection of man. So they probably put someone in there, or for the gods would not be jealous and get angry with human beings.

We might like to turn the idea around and look at the true explanation of the near-symmetry of nature is that God made the laws only nearly symmetrical so that we should not be jealous of His perfection!

The Feynman LECTURES ON PHYSICS

MAINLY ELECTROMAGNETISM AND MATTER

RICHARD P. FEYNMAN

*Ronald Coase Fellow Physics Department
California Institute of Technology*

ROBERT B. LEIGHTON

*Professor of Physics
California Institute of Technology*

MATTHEW SANDS

*Professor
Stanford University*

OXFORD PUBLIC LIBRARY
751 GOLDEN GATE AVENUE
SAN FRANCISCO CALIFORNIA 94102



ADDISON WESLEY PUBLISHING COMPANY, INC.
PEABODY, MASSACHUSETTS • NEW YORK • LONDON

January 12, 1962

CALIFORNIA INSTITUTE OF TECHNOLOGY

Address to the Chair of Committee of Assessors

ALL THE INFORMATION WHICH WAS RECEIVED FROM THE COMMITTEE
WAS SUBMITTED WITH THE SAME CARE AND ATTENTION AS THOSE
PRESENTED IN THE PUBLICATION.

Letter of George C. Clark, Chairman, A.C.M.T.C.

5000 University, Pasadena, Calif.

Introduction & Preface



These are the lectures I planned that I gave last year and the year before in the fall and spring semesters at Caltech. The lectures are, of course, not complete—they have been edited, some parts necessary and some not, so the lectures form only part of the complete course. The whole group of 196 students is listed—a big because most were weak in fact these lectures did not fully update some large portion of the students in certain sections, either too good or else too strong. In addition, there was a laboratory every seven weeks.

The space problem we had to grapple with was how to give a minimum number of lectures yet enthusiasm and interest among students coming out of 116 high schools and 116 backgrounds. This has included students less interesting and exciting physics—such as theory of relativity, quantum mechanics, and other modern areas. By the end of two years or one previous course, study would be very advanced, but not necessarily in the same, say, order, since it is not necessary. They were interested in relativity, quantum mechanics, and so forth, and also in general theory of everything. The question was whether or not we could make it possible to do so the more advanced and excited students by maintaining enthusiasm.

The culture from my point of view is to take every course, but not very seriously. I thought to address myself to the most intelligent in the class and to make sure, if possible, that what I used to help them was available to everybody, especially my bright students in the course—by providing in appendices an approach of the sort and success in various directions outside the main line of attack. For this reason, though, I tried very hard to "simplify" the situation so as always as possible, to point out in every case where the student has an interest and also the body of material, and then, when they wanted more—things were to be modified. I also felt like for such students it is important to indicate what it is they can do—say in a difficult problem—so that it could be distinguished from what has been done before, so that it is being presented as something new. When new ideas came in, I would try either to deduce them if they were deductive, or to explore them in a new field which had likely been in the past. If so, then what had already learned and what was not supposed to be previously—but was just called “”.

In several of the lectures I assumed that the student knew something about the course of high school—with that as background, implies, say, geometry, logic, and so on. I also, of course, left some basic material to make the lectures

for a definite order, ... I assume that I would not be allowed to introduce something until I've really understood it in detail. This was a great deal of trouble at things, to do, to teach, without complete consequences. These more complete consequences would come later on, the dependent lecture more advanced. Examples are the two responses of molecular and of the laser levels, which are at the moment ... in a very qualitative way and the laser development completely.

In the same time that I was talking to the most active student, I also wanted to understand the others, for whom the extra knowledge that applications are already cropping up and who cannot be expected to have most of the material in the book at all. The basic student I wanted there to be at least a general idea of problems of material which he could get. Even if he didn't understand everything, I hoped he wouldn't get nervous. I didn't expect him to understand everything, all the central and most direct features. In other words, of course, a certain intelligence on his part is necessary, and the central feature, a fundamental concept, and where the more elaborate calculations begin, which he may understand only at best.

In giving these lectures there was one serious difficulty. In the way the course was going, I was getting very feedback from the audience to the beginning, indicating how well the lectures were going over. This is almost a very serious difficulty, and I don't know how avoid the interaction with the audience. The whole thing was essentially an experiment. And it is not always a question about the same way that I hope I don't have to do it again. I think though, that things worked somewhat to the degree I experienced satisfactorily in the first year.

In the second year I was much less satisfied. In the first year of the course, dealing with electricity and magnetism, I couldn't think of any really unique or different way of doing it, of any way that would be substantially more exciting than the usual way of proceeding. So I didn't think I'd deliver very much in the lectures on electricity and magnetism. In the second year I had originally intended, as in the first, to discuss quantum theory, by giving some more lectures on the properties of matter, so as mainly to take up things like fundamental waves, solutions of the Schrödinger equation, interacting systems, orthogonal functions, calculating the first stage of what are usually called "the mathematical methods of physics." In retrospect, I think now, if I were doing it again, I would probably do this straightforward. I mean, I was not planning, in the second year, giving these lectures again, never suggesting reading assignments, or exercises or the like, or any direction in the quantum mechanics which you see, said in Volume II.

It is perfectly clear now, students were still majoring physics, even to limit their third year to quantum mechanics. Of course, based on the argument we made that many of the students in our course study physics as a background for their primary interests in other fields. And the usual way of dealing with quantum mechanics makes them pay a heavy price, probably more than necessary, to students because they have to take a long time to learn the details of real applications. Such skills in more complex applications such as solid state engineering and electronics - the full machinery of the Schrödinger equation appears to be rarely used. So I tried to describe the basics of quantum mechanics in a way which wouldn't require the user to know the mathematics of perturbation theory, scattering, etc., for a progress. I might say, in an interesting thing to try to do - to present quantum mechanics in the reverse direction, for several lessons which may be apparent in the literature elsewhere. However, I think that the experiments in the quantum mechanics part were not completely successful, in any part, because I really did not have enough time at the end of it should, for instance, have had them for some lectures in order to deal more explicitly with matter in energy bands and the overall dependence of amplitudes. And, I had never presented the subject this way before, so the bulk of feedback was predominantly negative. I now believe the quantum mechanics should be given at a lower level. Maybe, I now believe by non-symmetry. Then I think it's right.

The reason there are no lectures on how to solve problems is because there were no assignments. At 2, only I did problems, so I can't say what I did in the first year or how to solve problems. They are not important here. As a general way, I think, of solving

gives no search for only cataloged items by subject or reading system, but why was nothing easily available? The 10th and 11th editions of *Calculus* due to Mathematics Faculty, and from my own.

The question of course is how well the experiment has succeeded. My own point of view, which however does not seem to be shared by most of the people who worked out the system is, I assume, I didn't do it and we will be in trouble. When I look at the way the majority of the students handle the systems *Calculus* *Computers*, I think that the system is a failure. Of course, if I assume that, as I did, there were no errors, our students were more surprised single-handedly able to understand in all of the lectures, and who were quite active, I working with the project, and working upon their own projects in all subjects and interested was. I like people have now. I believe a first rate computing program they are excellent. I think I even can be asked. A man "in possession" of certain possibilities of mathematics except in his happy "specimens" where it cannot cope due to "difficulties".

Still, I didn't want to leave my student in silence, almost as perhaps I did. I think one way was probably better than none more useful, by putting more time into developing a series of problems which would give the only use of the books in the lecture. Problems give a great opportunity to fill out the material in the lecture and make it more meaningful, more complete and more selfful in the mind the ideas that have been exposed.

I think, however, that there isn't any solution to this problem. A plausible answer is to teach the best teaching you, reading only when, but less a real individual relationship between a teacher and a good teacher-student, where the student reads & the teacher talks about the things, and talk about the things. It's impossible to have such a simple situation, so much be simple reading problem that was assigned. But in our modern times we have so many students to teach that we have to hire the same substitute for the teacher. Perhaps we "hires" as a teacher or as substitute. Perhaps it is a small place where there are "hired" a teacher and student. They may get some improvement since there there is better. Perhaps they will have run through them through, or going on to another one of the best teacher.

John C. P. Tassoudji

JPNW 2967

Foreword

Over seven years Richard P. Feynman focused his interests on the subjects comprising the physical world, and it is his skillful writing and the depth of his ideas, now, in his given two years of his activity and his energy to his Lectures on Physics for beginning students. For them he has distilled the essence of his knowledge, and his efforts—so far as they can help to give a picture of the physicist's attitude. In his lectures he has brought the brilliance and clarity of his thought, the originality and vitality of his approach, and the courageous enthusiasm of his delivery. It was a joy to behold.

The first year's lectures formed the basis for the first volume of this set of books. We have inserted in this one second volume of texts, some kind of a part of the second year's lectures—which were given as the sophomore course during the 1949-1950 academic year. The rest of the second year's lectures will make up Volume II.

Of the second year of lectures, the first two-thirds were devoted to a fairly complete treatment of the physics of electricity and magnetism. The presentation was intended to serve a dual purpose. We hoped, first, to give the students a straightforward view of the great chapters of physics—from the early grouping of brackets through the great synthesis of Maxwell, on to the complete electron theory of the solid, properties and finally with the still unbroken connection of the electromechanical cell energy. Are we hopeful, second, by introducing at the outset the basic notion of vector fields, to give a subtle education in the mathematical language of field theories?—a knowledge, I repeat, of the mathematical methods needed subjects from 2000 parts of physics were sometimes analyzed together with their classical counterparts. We maintained, in this case, however, the generality of the mathematics, while concentrating upon the few, the more difficult. And we emphasized, again, by the kinds of exercises and exercises, to give you both the concepts.

Following the electromagnetism there are two chapters on an auxiliary topic: "Time." In the first chapter of each pair, the elementary and practical views are presented. The second chapter on each topic attempts to give an analysis of the whole complex subject of quantum mechanics. The subject can lead to, these two chapters, and well be pursued without serious loss, even though they are not at all necessary preparation for Volume II.

The last quarter, approximately, of the second year was dedicated to an introduction to quantum mechanics. This material was left out in the third volume.

In the second year, Feynman's lectures, we wished to do more than provide a transcription of what was said. We hoped to make the written version as close as possible to the ideas on which the original lectures were based. For this reason figures due credit be given by making only minor adjustments of the wording in the original. Likewise, the others of the lectures, I may say, the reporting and rearrangement of the material was equal. So, lectures we felt we should now better prepared to improve the clarity or balance of the presentation. Throughout the process we consulted from the original books and corrected our material whenever.

The translation of the 1949-1950 spoken words into a coherent text is a difficult task; it is, however, the primary aim to be accomplished by the

other academic buildings which come with the introduction of a new course—preparing for written exams, and giving students, designing exercises and examinations, and grading them, and so on. Many hours—and heads—will now have to be spent on this task, one which, I believe, from 1946, no longer is of vital interest or a readily-reachable *product* of the original *Program*. In other words we have failed to shorten this task. Our successors and ours' will therefore inherit the legacy, we regret.

As explained in detail in the *Procedure to Admit*, these lectures were but one aspect of a programme which was supervised by the Physics Course Committee (R. P. Leighton, Chairman, L. V. Bates, and M. Sones); at the Communication of Technology, and suggested, informed by the Technical Education. In addition the teaching people helped with any aspect of another of the responsibilities of Oxford Institute for the Armed Services: T. K. Crowley, M. V. Clegg, J. R. Clegg, J. B. Flavin, T. W. H. Harvey, M. H. Jones, W. J. Keras, A. W. Langford, R. P. Leighton, J. Matthews, M. S. Mayes, F. J. Quinn, W. Whalley, C. H. White, and E. Z. Williams. Others who assisted indirectly through their work on the courses: J. Bush, G. D. Caudine, M. J. Chesser, R. Dolan, H. H. Hall, and A. M. T. Dr. Professor Gerry Mullings, though he did not always do his task with intelligence, did nevertheless become the butt of much

The story of physics you find here would, however, not have been complete for the armed services had it not included the following: Richard P. Feynman,

Review Series

March 1944

Contents

CHAPTER 1. ELECTROSTATICS

- 1.1 Electrical forces 1.1
- 1.2 Electric field and potential 1.2
- 1.3 Characteristics of electric fields 1.4
- 1.4 The law of conservation of charge 1.5
- 1.5 What we do field? 1.5
- 1.6 Electrostatic phenomena in science and technology 1.10

CHAPTER 2. DIFFERENTIAL CALCULUS OF VECTOR FIELDS

- 2.1 Understanding physics 2.1
- 2.2 Scalar and vector fields 2.1
- 2.3 Derivatives of field's the gradient 2.4
- 2.4 The operator ∇ 2.4
- 2.5 Operations with ∇ 2.4
- 2.6 The differential equation of heat flow 2.8
- 2.7 Second derivatives of vector fields 2.9
- 2.8 Divergence 2.11

CHAPTER 3. VECTOR INTEGRAL CALCULUS

- 3.1 Vector integrals: the line integral (Eq. 3.1)
- 3.2 The flux of a vector field 3.2
- 3.3 The flux from a surface Gauss theorem 3.4
- 3.4 Flux continuation; the diffusion equation 3.6
- 3.5 The circulation of a vector field 3.8
- 3.6 The circulation current; Ampere's theorem 3.9
- 3.7 Coulomb and divergence-free fields 3.10
- 3.8 Summary 3.1.

CHAPTER 4. ELECTRODYNAMICS

- 4.1 Gauss 4.1
- 4.2 Continuity law (dissipation) 4.2
- 4.3 Electric potential 4.3
- 4.4 $E = -\nabla V$ 4.5
- 4.5 The flux of E 4.5
- 4.6 Gauss law; divergence of E 4.9
- 4.7 Field of a spherical charge 4.10
- 4.8 Field lines, eq. potential surfaces 4.11

CHAPTER 5. EQUILIBRIUM OF CHARGE

- 5.1 Equilibrium of charge example 5.1
- 5.2 Equilibrium in an electrostatic field 5.1
- 5.3 Equilibrium with conditions 5.2
- 5.4 Stability of charge 5.2
- 5.5 The field of a line charge 5.3
- 5.6 A sheet of charge; resistance 5.4
- 5.7 A shell of charge; a spherical shell 5.4
- 5.8 Is the field of a point charge constant? 5.7
- 5.9 Instability of a conductor 5.7
- 5.10 The field in a cavity of a conductor 5.8

CHAPTER 6. THE ELECTRIC FIELD BY PARTICLES

- 6.1 Equations of the electrostatic potential 6.1
- 6.2 The electric dipole 6.2
- 6.3 Boundary value equations 6.4
- 6.4 The dipole potential in a problem 6.4
- 6.5 The dipole approximation for an arbitrary distribution 6.6
- 6.6 The fields of charged condensers 6.7
- 6.7 The method of images 6.7
- 6.8 A point charge near a conducting plane 6.9
- 6.9 A point charge near a conducting sphere 6.10
- 6.10 Conductors: parallel plates 6.11
- 6.11 High-voltage breakdown 6.13
- 6.12 The dielectric breakdown 6.14

CHAPTER 7. THE DIELECTRIC FIELD & MAXWELL'S EQUATIONS (CONTINUED)

- 7.1 Methods for finding the dielectricity 7.1
- 7.2 Two-dimensional fields; functions of the complex variable 7.2
- 7.3 Phonon oscillations 7.5
- 7.4 Colloidal particles in an electrolyte 7.6
- 7.5 The electrostatic field of a grid 7.10

CHAPTER 8. ELECTROSTATIC ENERGY

- 8.1 The electrostatic energy of charges. A uniform sphere 8.1
- 8.2 The energy of a condenser. Based on energy conservation 8.2
- 8.3 The electrostatic energy of a dipole 8.4
- 8.4 Dielectric energy in space 8.5
- 8.5 Energy in the electrostatic field 8.9
- 8.6 The energy of a point charge 8.12

CHAPTER 9. ELECTROLYSIS IN THE ATMOSPHERE

- 9.1 The electric potential gradient of the atmosphere 9.1
- 9.2 Electric currents in the atmosphere 9.2
- 9.3 Origin of the atmospheric currents 9.4
- 9.4 Tornadoes 9.5
- 9.5 The mechanism of charge separation 9.7
- 9.6 Lightning 9.10

CHAPTER 10. DIPOLESES

- 10.1 The dielectric constant 10.1
- 10.2 The polarization vector P 10.2
- 10.3 Polarization charges 10.3
- 10.4 The electrostatic equations with dielectrics 10.6
- 10.5 Fields and forces with dielectrics 10.7

CHAPTER 11. ISOTROPIC DIPOLES

- 11-1 Molecular dipoles 11-1
- 11-2 Dielectric polarization 11-1
- 11-3 Polarizability; orientation polarization 11-2
- 11-4 Electric field in vicinity of a dipole 11-3
- 11-5 The electric potential resulting from a dipole: Maxwell's equation 11-5
- 11-6 Self-electrostatic energy 11-7
- 11-7 Electrostatics: Biot-Savart law 11-8

CHAPTER 12. ELECTROSTATICS AND FIELD

- 12-1 The same charges have the same solutions 12-1
- 12-2 The flux of field at point source due to infinite plane boundary 12-2
- 12-3 The electrostatic moment 12-3
- 12-4 The diffusion of neutrons: a uniform spherical source in a homogeneous medium 12-6
- 12-5 Levitating fluid over the Moon just a sphere 12-7
- 12-6 Chromatography: sorting of particles 12-10
- 12-7 The "underappreciation" of science 12-12

CHAPTER 13. MAGNETOSTATICS

- 13-1 The magnetic field 13-1
- 13-2 Electric current; the conservation of charge 13-1
- 13-3 The magnetic force on a current 13-2
- 13-4 The magnetic field of steady currents 13-4
- 13-5 The magnetic field of a straight wire and of a solenoid: atomic current 13-5
- 13-6 The relativity of magnetism and electric fields 13-6
- 13-7 The transformation of currents and charges 13-11
- 13-8 Susceptibility; the light bulb rule 13-11

CHAPTER 14. THE MAXWELL FIELD IN VARIOUS SITUATIONS

- 14-1 The scalar potential 14-1
- 14-2 The vector potential of known currents 14-2
- 14-3 A solenoid 14-4
- 14-4 A long solenoid 14-5
- 14-5 The field of a open loop: the magnetic dipole 14-7
- 14-6 The vector potential of a circuit 14-8
- 14-7 The law of Biot and Savart 14-9

CHAPTER 15. THE VARIATIONAL PRINCIPLE

- 15-1 The force on a current loop: center of a dipole 15-1
- 15-2 Mechanical and electrical energy 15-3
- 15-3 The energy of steady currents 15-6
- 15-4 Sources 15-9
- 15-5 The wave-polarized and quantum mechanics 15-8
- 15-6 What is the best state? Rutherford dynamics 15-17

CHAPTER 16. ISOTROPIC DIPOLES

- 16-1 Molecular polarizability 16-1
- 16-2 Transformation and reduction 16-4
- 16-3 Purcell's induced current 16-5
- 16-4 Technical technology 16-8

CHAPTER 17. THE LAW OF INVERSION

- 17-1 The physics of inversion 17-1
- 17-2 Description terms: "inversion" 17-2
- 17-3 Parity violation in the nuclear electric field; the Ising model 17-3
- 17-4 A paradox 17-5
- 17-5 Alternating-current generators 17-5
- 17-6 Natural selection 17-9
- 17-7 Solid-state 17-11
- 17-8 Infrasound and magnetic energy 17-12

CHAPTER 18. THE MAXWELL EQUATIONS

- 18-1 Maxwell's equations 18-1
- 18-2 How to view networks 18-2
- 18-3 A long circular pipe 18-3
- 18-4 A travelling field 18-5
- 18-5 The speed of light 18-8
- 18-6 Solving Maxwell's equations; the possibility and the necessity of light 18-9

CHAPTER 19. THE PHYSICS OF LIGHT FORCES

- A special lecture—*Light: radiation 19-1*
- A note added after the lecture 19-11

CHAPTER 20. SOLUTIONS OF MAXWELL'S EQUATIONS IN FREE SPACE

- 20-1 Waves in free space; power density 20-1
- 20-2 One-dimensional waves 20-3
- 20-3 Scientific imagination 20-9
- 20-4 Electrical waves 20-11

CHAPTER 21. SOLUTIONS OF MAXWELL'S EQUATIONS WITH CONDUCTORS AND CHARGES

- 21-1 Layered electromagnetic waves 21-1
- 21-2 Spherical waves from a point source 21-2
- 21-3 The general solution of Maxwell's equations 21-4
- 21-4 The incident and reflected wave 21-5
- 21-5 The potentials of a moving charge: the general solution of Liénard and Mathieu 21-9
- 21-6 The potentials for a charge moving over conductors; the Lorentz formula 21-12

CHAPTER 22. AC CIRCUITS

- 22-1 Impedance 22-1
- 22-2 Generators 22-2
- 22-3 Networks of ideal elements: Kirchhoff's rules 22-2
- 22-4 Equivalent circuits 22-10
- 22-5 Energy 22-11
- 22-6 A linear network 22-12
- 22-7 Filters 22-14
- 22-8 Overdamped motion 22-16

CHAPTER 23. COMMUNICATIONS

- 23-1 Rayleigh scattering 23-1
- 23-2 Resonance at high frequencies 23-2
- 23-3 Amplitude control 23-6
- 23-4 (X-ray, radio) 23-9
- 23-5 Communication-resistant circuits 23-10

CHAPTER 24. WAVEGUIDES

- 24-1 The transmission line 24-1
- 24-2 The radiation from a dipole 24-4
- 24-3 The wave function 24-6
- 24-4 The speed of the guided waves 24-7
- 24-5 Other modes 24-7
- 24-6 Wave mode coupling 24-8
- 24-7 Waveguide modes 24-10
- 24-8 Another way of looking at the guided waves 24-10

CHAPTER 25. ELECTRODYNAMICS IN RELATIVISTIC MOTION

- 25-1 Electromotion 25-1
- 25-2 The scalar product 25-1
- 25-3 The four-dimensional gradient 25-6
- 25-4 Electromagnetic current and charge density 25-10
- 25-5 The components of a moving charge 25-10
- 25-6 The invariance of the equations of electrodynamics 25-10

CHAPTER 26. ENERGY AND POWER IN ELECTRIC FIELDS

- 26-1 The free-potential of a moving source 26-1
- 26-2 The field of a point charge moving with velocity 26-2
- 26-3 Relativistic conservation of the fluxes 26-2
- 26-4 The equations of motion in relativity 26-11

CHAPTER 27. FIELD ENERGY AND FIELD DENSITY

- 27-1 Field energy 27-1
- 27-2 Energy conservation and electromagnetism 27-2
- 27-3 Energy density and energy flow in the electromagnetic field 27-2
- 27-4 The multipole of the field 27-3
- 27-5 Losses of energy flow 27-6
- 27-6 Dielectric loss 27-6

CHAPTER 28. THE ENERGY OF A FIELD AND MASS

- 28-1 The field energy of a point charge 28-1
- 28-2 The field momentum of a moving source 28-2
- 28-3 Electromagnetic mass 28-3
- 28-4 The field of a point charge moving with velocity 28-4
- 28-5 Anomalous rotatory inertia theory 28-6
- 28-6 The nuclear force field 28-12

CHAPTER 29. THE MOTION OF CHARGE IN ELECTRIC AND MAGNETIC FIELDS

- 29-1 Motion in a uniform electric or magnetic field 29-1
- 29-2 Mechanical analogies 29-1
- 29-3 An electromagnetic lens 29-1
- 29-4 A magnetic lens 29-2
- 29-5 The electron microscope 29-3
- 29-6 Accelerators 29-4 & 29-5
- 29-7 Alternating-gradient focusing 29-5
- 29-8 Motion in crossed electric and magnetic fields 29-6

CHAPTER 30. THE INTERNAL STRUCTURE OF CRYSTALS

- 30-1 The internal geometry of crystals 30-1
- 30-2 Chemical bonds in crystals 30-1
- 30-3 The growth of crystals 30-1
- 30-4 Crystal lattices 30-4
- 30-5 Symmetries in two dimensions 30-4
- 30-6 Symmetries in three dimensions 30-7
- 30-7 The anisotropic metals 30-8
- 30-8 Dislocations and crystal growth 30-9
- 30-9 The Hagen-Poelzelt model 30-10

CHAPTER 31. THERMOS

- 31-1 The vector of probability 31-1
- 31-2 Transforming the tensor components 31-1
- 31-3 The surface dipole 31-2
- 31-4 Ellipses versus the Jones of length 31-6
- 31-5 The one-particle 31-7
- 31-6 The wave of death 31-7
- 31-7 Velocity of higher order 31-11
- 31-8 The importance of electromagnetically induced anisotropy 31-12

CHAPTER 32. RADIATION BY INDEX OF REFRACTION

- 32-1 Polarization of matter 32-1
- 32-2 Maxwells eq. values in a dielectric 32-3
- 32-3 Waves in dielectrics 32-4
- 32-4 The complex index of refraction 32-6
- 32-5 The index of refraction 32-8
- 32-6 Waves in liquids 32-10
- 32-7 Low-frequency and high-frequency approximations for skin depth and dielectric frequency 32-11

CHAPTER 33. REFLECTION AND TRANSMISSION

- 33-1 Reflection and refraction of light 33-1
- 33-2 Waves in dielectrics 33-2
- 33-3 Total reflection conditions 33-4
- 33-4 Normal reflection and transmission waves 33-7
- 33-5 Diffusion from metals 33-11
- 33-6 Total internal reflection 33-13

CHAPTER 34. THE MAGNETISM OF MATTER

- 34-1 Demagnetization and paramagnetism 34-1
- 34-2 Magnetic moments and angular momentum 34-1
- 34-3 The precession of rotating charges 34-3
- 34-4 Paramagnetism 34-5
- 34-5 Curie's theorem 34-6
- 34-6 Chemical physics gives neither diamagnetic film nor paramagnetism 34-6
- 34-7 Angular momentum in diamagnetism 34-7
- 34-8 The magnetic entropy of atoms 34-11

CHAPTER 35. POLARIZATION AND MAGNETIC RESONANCE

- 35-1 Unpolarized magnetic down 35-1
- 35-2 The Stern-Gerlach experiment 35-3
- 35-3 The Earth's dipole 35-4
- 35-4 The paramagnetism of bulk materials 35-6
- 35-5 Curiously anisotropic demagnetism 35-10
- 35-6 Nuclear magnetic resonance 35-10

Chapter 36. Electromagnetism

- 36-1 Magnetization and χ 36-1
- 36-2 Earth's field 36-2
- 36-3 The magnetization curve 36-3
- 36-4 Curie's law 36-4
- 36-5 Electromagnets 36-5
- 36-6 Symmetric magnetization 36-6

Chapter 37. Dielectric Materials

- 37-1 Understanding ferro-magnetism 37-1
- 37-2 Thermodynamic properties 37-2
- 37-3 The hysteresis curve 37-3
- 37-4 Ferromagnetic hysteresis 37-4
- 37-5 Complementary magnetic materials 37-5

Chapter 38. Electricity

- 38-1 Ohm's law 38-1
- 38-2 Uniform resistors 38-2
- 38-3 The junction rule, chain rule 38-3
- 38-4 The heat term 38-9
- 38-5 Breaking 38-10

Chapter 39. Static Electricity

- 39-1 The nature of static 39-1
- 39-2 The forces of repulsion 39-2
- 39-3 The minimum attractive force 39-3
- 39-4 Conductive behavior 39-28
- 39-5 Calculating the electric constant 39-10

Chapter 40. The Flow of DC Water

- 40-1 Bernoulli's eq. 40-1
- 40-2 The equation of motion 40-2
- 40-3 Steady flow—Bernoulli's theorem 40-4
- 40-4 Circulation 40-5
- 40-5 Vortex flow 40-10

Chapter 41. The Flow of Water

- 41-1 Viscosity 41-1
- 41-2 Viscous flow 41-4
- 41-3 The Froude number 41-5
- 41-4 Flow past a circular cylinder 41-7
- 41-5 The limit of smoothness 41-9
- 41-6 Couette flow 41-10

Answers

Electromagnetism

1-1 Electrical Forces

Consider a force like gravitation which varies predominately inversely as the square of the distance, but which is about a billion-billion-billion-billion times stronger. And with greater difference. There are two kinds of "matter," which we can call positive and negative. Like kinds repel and unlike kinds attract with a static charge there is only attraction. What would happen?

A bunch of positives would repel with an enormous force and spread out in all directions. A bunch of negatives would do the same. But an evenly mixed bunch of positives and negatives would do something completely different. The opposite pieces would be pulled together by the enormous electrostatics. The net result would be that the terrible forces would balance the repulsion almost perfectly, by forming tight, fine mixtures of the positive and the negative, and the mean two separate bunches of such mixtures there would be practically no attraction or repulsion at all.

There is such a force: the electrical force. And it's nature is a mixture of positive protons and negative electrons which are attracting and repelling with this great force. So perfect is this balance, however, that when you stand next to someone else you don't feel any force at all. If there were even a little bit of imbalance you would know it. If you were standing at arm's length from someone and each of you had one proton more electrons than protons, the repelling force would be impossible. How great? Enough to lift the Empire State Building? Not To It. Never! Except? Not. The repulsion would be enough to lift a "weight" equal to that of the entire earth!

With such enormous forces so perfectly balanced in this intimate mixture, it's not hard to understand that another, trying to keep its positive and negative charges in this balance, has here a great stiffness and strength. The Empire State Building, for example, sways only eight feet in the wind because the electrical force holds every electron and proton onto its legs with a vice-like grip. On the other hand, if we look at a cluster of a scale small enough that we see only a few atoms, any small piece will not usually have an equal number of positive and negative charges, and so there will be strong repulsive electrostatic forces. Even when there are equal numbers of both charges in two neighboring small pieces, there may still be large net electrical forces because the forces between individual charges vary inversely in the square of the distance. A net force can arise if a negative charge of one piece is closer to the positive than to the negative charge of the other piece. The attractive forces can then be larger than the repulsive ones and there can be a net attraction between two small pieces with no excess charges. The forces that hold the atoms together, and the chemical bonds that hold molecules together, are really electrical forces, acting in regions where the balance of charge is not perfect, or where the distances are very small.

You know, of course, that around E & B there will be positive protons in the nucleus and with electrons outside. You may ask: "If this electrical force is so terrific, why don't the positive and electrons just get on top of each other? If they want to be like an isolated molecule, why isn't it still there isolated?" The answer lies in the uncertainty principle. If we try to confine one electron in a region that is very close to the positive, then according to the uncertainty principle they must have some mean kinetic momentum which is larger than the \hbar we use to confine them. It is this relation, required by the laws of quantum mechanics, that keeps the electrical attraction from bringing the charges way closer together.

1-1 Electrical forces

1-2 Electric and magnetic fields

1-3 Characteristics of vector fields

1-4 The laws of electromagnetism

1-5 What are the fields?

1-6 Electromagnetism in science and technology

Review Chapter 12, Vol. I, Characteristics of Force

There is another question: "What holds the nucleus together?" To understand, there are several problems, all of which are positive. Why don't they push themselves apart? It turns out that in nuclei there are in addition to attractive "nuclear" forces, called nuclear forces, others are present from the electric forces, and which are able to hold the protons together in spite of the electrostatic repulsion. The nuclear forces, however, have a short range—their force falls off much more rapidly than light. And this has an important consequence. If a nucleus has too many protons in it, it gets too big, and it will not stay together. The example is hydrogen, with 92 protons. The nuclear forces act mainly between each proton (or neutron) and its nearest neighbor, while the electrical forces act over large distances, giving a repulsion between each proton, and all of the others to the nucleus. The mass increases in a nucleus, the stronger is the electrical repulsion, and so in the case of a helium, the balance is so strong that the nucleus is always ready to fly apart from the repulsive electrical forces. If such a nucleus is just "disrupted" lightly one can be done by shooting it, a slow neutron; it breaks into two pieces, and with positive charges, and these pieces fly apart by electrical repulsion. The energy which is liberated is the energy of the nuclear bomb. This energy is usually called "nuclear" energy, but it is really "electrostatic" energy released when electrical forces have overcome the attractive nuclear forces.

We may ask, finally, what holds negatively charged dust together since it has no nuclear forces. If an electron is all made of one kind of substance, each part should repel the other parts. "No, wait, doesn't it do that?" But does the electron have "parts"? Perhaps we should say that the electron is just a point and that electrical forces only act between different point charges, so the electron itself does not yet again itself. Perhaps, all we can say is that the attraction of opposite signs in the electron together has produced many difficulties in the attempt to obtain a complete theory of electromagnetism. The question has never been answered. We will, unfortunately, not be discussing this subject too much in later chapters.

As we have seen, we should expect that it is a combination of electrical forces and gravitation-mechanical effects that will determine the detailed structure of matter's in bulk, and, therefore, their properties. Since atoms are held, some sort of. Some are electrical "conductors"—because their electrons are free to move about, others are "insulators"—because their electrons are held tightly in individual atoms. We will consider later how some of these properties come about, but first to a very incomplete subject, or we will begin by looking at the electrical forces only in simple situations. We begin by treating only the case of electrically insulating materials, which is really a part of the same subject.

We know well that the electrical forces, like gravitational forces, decrease inversely as the square of the distance between charges. This relationship is called Coulomb's law. But it is not precisely true when charges are moving—the electrical forces depend also on the motion of the charges in a complicated way. One part of the force between moving charges we call the magnetic force. It is really one aspect of an electric effect. That is why we call the subject "electromagnetism".

There is an important general principle that makes it possible to treat electromagnetism in a relatively simple way. We find, from experiment, that the force that acts on a particular charge—on other tiny minor other charges there are no force, they are moving—depends only on the position of the particular charge, on the velocity of the charge, and on the amount of charge. We can write the force F on a charging moving with a velocity v

$$F = q(E + v \times B). \quad (1.1)$$

We call E the electric field and B the magnetic field at the location of the charge. The important thing is that the electrical forces from all the other charges in the universe can be summarized by giving just these two forces. Their values will depend on where the charge is, and may change with time. Furthermore, if we replace a test charge with another charge, the force on the new charge will be just proportional to the amount of charge; so among all the rest of the charges in the

Table 1.1 Some basic Greek letters and commonly used capitals

α	alpha
β	beta
γ	gamma
δ	delta
ϵ	epsilon
ζ	zeta
η	eta
θ	theta
ι	iota
κ	kappa
λ	lambda
μ	mu
ν	nu
ρ	rho
σ	sigma
τ	tau
ϕ	phi
ψ	psi
ω	omega
Ω	Omega

would do in change their positions or motions. (In real situations, of course, each charge produces forces on all other charges in the neighborhood and may cause these other charges to move, and so in some cases the field can change if we replace our particular charge by another.)

We know from Vol. I how to find the motion of a particle if we know the force on it. Equation (1.1) can be combined with the equation of motion to give

$$\frac{d}{dt} \left[\frac{1}{r} - \frac{q_0 q}{\pi^2 c^2 r^2} \right] = F = q(E + v \times B). \quad (1.2)$$

If E and B are given, we can find the motion. Now we need to know how the E 's and B 's are produced.

One of the most important simplifying principles about the way the fields are produced is this: Suppose a number of charges moving in some manner would produce a field E_1 , and another set of charges would produce E_2 . If both sets of charges are in place at the same time (keeping the same locations and motions they had when considered separately), then the field produced is just the sum

$$E = E_1 + E_2. \quad (1.3)$$

This law is called the principle of superposition of fields. It holds also for magnetic fields.

This principle means that if we know the law for the electric and magnetic fields generated by a single charge moving in an arbitrary way, then all the laws of electrodynamics are complete. If we want to know the force on charge A we need only calculate the E 's and B 's produced by each of the charges B , C , D , etc., and then add the E 's and B 's from all the charges to find the fields, and from them the forces acting on charge A . It is bad only to remember that the field produced by a single charge was simple; this would be the easiest way to describe the laws of electrodynamics. We have already given a description of this law (of chapter 25, Vol. I) and it is, unfortunately, rather complicated.

It turns out that the form in which the laws of electrodynamics are simplest are not when you rightly expect. It is not simple to give a formula for the force that one charge produces on another. It is true that when charges are stationary the Coulomb force law is simple, but when charges are moving, from the relativity, the complications by delays in time and by the effects of acceleration, amount of time. As a result, we do not wish to present electrodynamics only through the force laws between charges; we find it more convenient, in another point of view—a point of view in which the laws of electrodynamics appear to be the most easily manageable.

1-2 Electric and magnetic fields

First we must extend somewhat our ideas of the electric and magnetic vectors, E and B . We have defined them in terms of the forces that are felt by a charge. We will now associate electric and magnetic fields with points even when there is no charge present. We are saying, in effect, that since there is a force "acting on" the charge, there is still "something" there when the charge is removed. If a charge located at the point (x, y, z) at the time t feels the force F given by Eq. (1.1) we associate this vector F with the point (x, y, z) . We may think of $F(x, y, z, t)$ and $B(x, y, z, t)$ as giving the vector that would be experienced at time t by a charge located at (x, y, z) , with the condition that placing the charge there did not disturb the positions or field values of all the other charges responsible for the fields.

Following this idea, we associate with every point (x, y, z) in space two vectors E and B , which may be changing with time. The electric and magnetic fields are, then, viewed as vector functions of x , y , z , and t . Since a vector is specified by its components, each of the fields $E(x, y, z, t)$ and $B(x, y, z, t)$ represent three mathematical functions of x , y , z , and t .

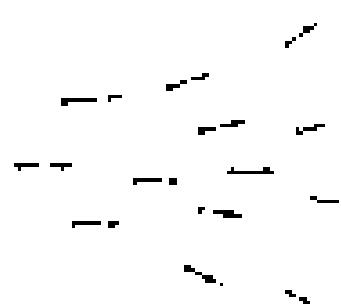


Fig. 1-1. A vector field may be represented by drawing a set of arrows, whose magnitudes and directions indicate the value of the vector field at the particular point where the arrows are drawn.

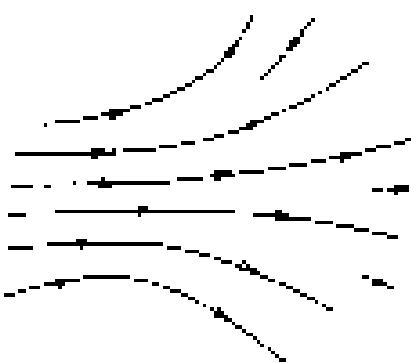


Fig. 1-2. A vector field can be represented by drawing lines which are tangent to the direction of the field vector at each point, and by drawing the density of these proportional to the magnitude of the field vector.

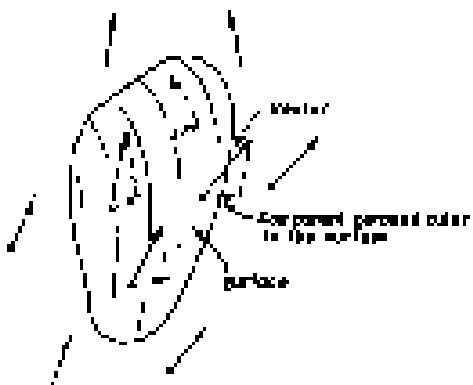


Fig. 1-3. The flux of a vector field through a surface is defined as the average value of the normal component of the vector times the area of the surface.

It is precisely because \mathbf{F} (or \mathbf{B}) can be specified at every point in space that it is called a "field." A "field" is any physical quantity which takes on different values at different points in space. Temperature, for example, is a field—in this case a scalar field, which we write as $T(x, y, z)$. The temperature could also vary in time, and we would say the temperature field is time-dependent, and write $T(x, y, z, t)$. Another example is the "velocity field" of a flowing liquid. We write $\mathbf{v}(x, y, z)$ for the velocity of the liquid at each point in space at the time t ; this is a vector field.

Returning to the electromagnetism field—although they are produced by charges according to complicated formulas, they have the following important characteristic. The relationship between the values of the fields at various points in the volume x is usually quite complex. With only a few such relationships in the form of differential equations we can describe the fields completely. It is a feature of such equations that the laws of electrodynamics are much simpler when they have been written in terms of the behavior of fields.

The most correct is also the most abstract: we simply consider the fields as mathematical functions of position and time. We can also attempt to get a certain picture of the field by drawing vectors at many points in space, each of which gives the field's strength and direction at that point. Such a representation is shown in Fig. 1-1. We can go further, however, and draw lines which are everywhere tangent to the vectors—which, to speak literally, follow the arrows and keep track of the direction of the field. When we do this we lose track of the length of the vectors, but we can keep track of the strength of the field, by drawing the lines far apart where the field is weak and close together where it is strong. We adopt the convention that the number of lines per unit area at right angles to the lines is proportional to the field strength. This is, of course, only an approximation, and it will require, in general, that we have spacetime start up in order to keep the number, up to the strength of the field. The field of Fig. 1-1 is represented by field lines in Fig. 1-2.

1-1. Characteristics of vector fields

There are two mathematically important properties of a vector field which we will use in our description of the laws of causality from the field point of view. Suppose we imagine a closed number of some kind and ask whether we are losing "something" from the inside; that is, does the field have a "valley" or "valley?" For instance, for a velocity field we might ask whether the velocity is always outward on the surface, or more generally whether most fluid flows out (or) into "valleys" coming in. We call the direction of fluid going out through the surface the "flux of velocity" through the surface. The flow strength an element of a surface is just equal to the component of the velocity perpendicular to the surface times the area of the surface. If an arbitrary closed surface, the net outward flux of the velocity is the average normal component of the velocity times the area of the surface.

$$\text{Flux} = \text{Average norm. comp.} \times (\text{Surface area}). \quad (1.1)$$

In the case of an electric field, we can mathematically define something analogous to an "outflow," and we again ask "is there loss?" In fact, however, it is not the flow of any substance, because the electric field is not the velocity of anything. It turns out, however, that the mathematical quantity which is the average normal component of the field will be a useful quantity. We speak, then, of the electric flux—also defined by Eq. (1.1). Finally, it is also useful to speak of the flux not only through a completely closed surface, but through any bounded surface. As before, the flux through such a surface is defined as the average normal component of a vector times the area of the surface. These ideas are illustrated in Fig. 1-3.

There is a second property of a vector field that has to do with a law, called "law" a surface. Suppose again that we think of a velocity field that describes the flow of a liquid. We might ask this interesting question: Is the liquid circulating? I.e.

By "lost we mean: Is there a net "rotational" motion around some loop? Suppose that we inconspicuously freeze the liquid everywhere except inside of a tube which is of uniform bore, and which goes in a loop that closes back on itself as in Fig. 1-4. Outside of the tube the liquid stops moving, but inside the tube it may keep on moving because of the momentum to the trapped liquid: one is, of course, more momentum heading one way around the tube than the other. We define a quantity called the circulation as the angular speed of the liquid in the tube times circumference. We can again extend our loops and define the "circulation" for any vector field (even when there isn't anything moving). For any vector field the circulation around any "closed" closed curve is defined as the average tangential component of the vector (in a consistent sense) multiplied by the circumference of the loop (Fig. 1-5).

$$\text{Circulation} = \text{average tangential component} (\text{reference arc only}). \quad (1.5)$$

You will see that this definition does indeed give a number which is proportional to the circulation velocity in the quickly frozen tube described above.

With just these two ideas—flux and circulation—we can describe all the laws of electricity and magnetism at once. You may not understand the significance of the last right away, but they will give you some idea of the way the physics of electromagnetism will be ultimately described.

1-4 The laws of electromagnetism

The first law of electromagnetism describes the flux of the electric field:

$$\text{The flux of } E \text{ through any closed surface} = \frac{q_0}{\epsilon_0} \text{ net charge inside.} \quad (1.6)$$

where ϵ_0 is a constant constant. (This constant ϵ_0 is usually read as "epsilon-zero" or "spacel-nought".) If there are no charges inside the surface, even though there are charges nearby outside the surface, the average normal component of E is zero, as there is no net flux through the surface. To show the power of this one statement, we consider that Eq. (1.6) is the same as Coulomb's law, provided only that we also add the rule that the field from a single charge is spherically symmetric. For a point charge, we draw a sphere around the charge. Then the average normal component is just the value of the magnitude of E at any point. Since the field must be altered radially and have the same strength for all points on the sphere. Our rule now says that the field at the surface of the sphere, times the area of the sphere—that is, the outgoing flux—is proportional to the charge inside. If we were to make the radius of the sphere bigger, the area would increase as the square of the radius. The average normal component of the electric field times that area must be equal to the same charge inside, and so the field must decrease as the square of the distance—we get an "inverse square" field.

If we have an arbitrary curve in space and measure the circulation of the electric field around the curve, we will find that it is zero, in general, even though there is no free charge inside. Rather, for electricity there is a second law that shows for any surface S (not closed) which cuts the curve C ,

$$\text{Circulation of } E \text{ around } C = \frac{\partial}{\partial t} (\text{flux of } E \text{ through } S). \quad (1.7)$$

We can complete the laws of the electromagnetic field by writing two corresponding equations for the magnetic field B .

$$\text{Flux of } B \text{ through any closed surface} = 0. \quad (1.8)$$

For a surface S bounded by the curve C ,

$$\begin{aligned} \text{Circulation of } B \text{ around } C &= \frac{\partial}{\partial t} (\text{flux of } B \text{ through } S) \\ &\quad + \frac{\text{flux of electric current through } S}{\epsilon_0}. \end{aligned} \quad (1.9)$$

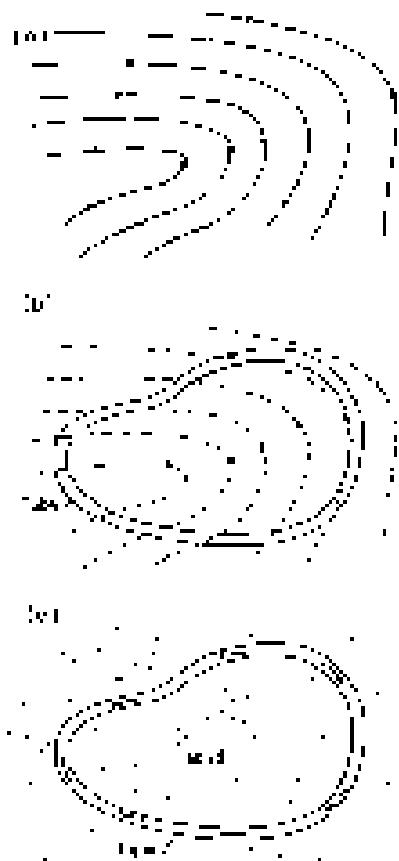


Fig. 1-4. (a) The velocity field in a liquid. Imagine a tube of uniform cross section that follows an arbitrary closed curve as in (b). If the liquid were suddenly frozen everywhere except inside the tube, the liquid in the tube would circulate as shown in (c).

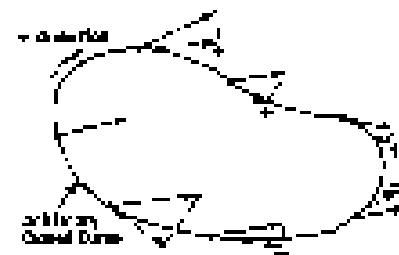


Fig. 1-5. The circulation of a vector field is the average tangential component of the vector (in a consistent sense) times the circumference of the loop.

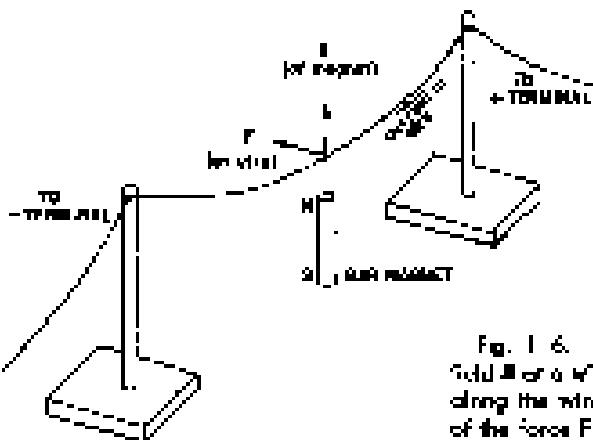


Fig. 1-6. A bar magnet gives a field B to a wire. When there is a current along the wire, the wire moves because of the force $F = qv \times B$.

The constant c^4 that appears in Eq. (1.1) is the square of the velocity of light. It appears because magnetism is in reality a relativistic effect of electricity. The constant c , i.e., the factor in v makes the units of electric current come out in a reasonable way.

Equations (1.6) through (1.8), together with Eq. (1.1), are all the laws of electrodynamics*. As you remember, the laws of Newton were very simple to write down, but they had a lot of complicated consequences and it took us a long time to learn about them all. These laws are not nearly as simple to write down, which means that the consequences are going to be more elaborate and it will take us quite a lot of time to figure them out.

We can illustrate some of the laws of electrodynamics by a series of small experiments which show qualitatively the interrelationships of electric and magnetic fields. We have expressed the first term of Eq. (1.1) when combing your hair, so we won't show that one. The second part of Eq. (1.1) can be demonstrated by passing a current through a wire which hangs above a bar magnet as shown in Fig. 1-6. The wire will move when a current is turned on because of the force $F = qv \times B$. When the current stops, the charges inside the wire are moving, so they have a velocity v and the magnetic field from the magnet exerts a force on them, which results in pulling the wire sideways.

When the wire is pushed to the left, we would expect that the magnet must feel a push to the right. (Otherwise we could put the whole thing on a wagon and have a propulsive system that didn't conserve momentum.) Although the force is too small to make movement of the bar magnet visible, a more sensitively supported magnet, like a compass needle, will show the movement.

How does the wire pull on the magnet? The current in the wire produces a magnetic field of its own that exerts forces on the magnet. According to the last

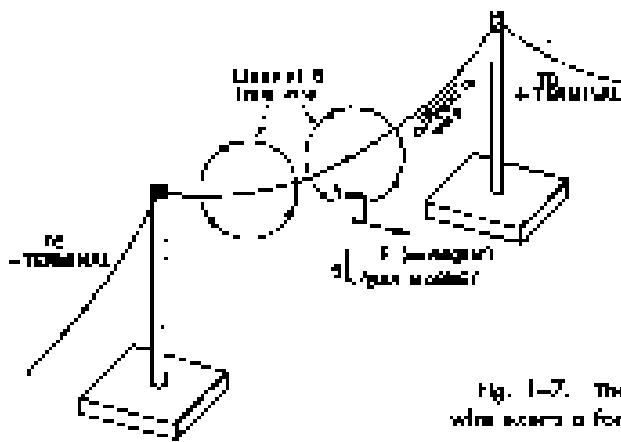


Fig. 1-7. The magnetic field of the wire exerts a force on the magnet.

* We need only add a remark about some conventions for the sign of the circulation.
L-6

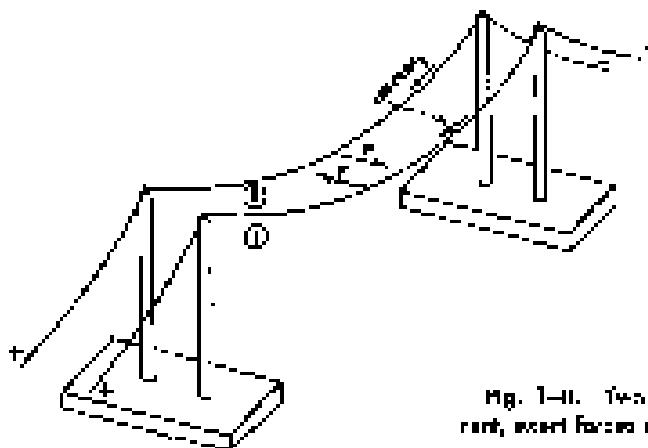


Fig. 1-8. Two wires, carrying current, exert forces on each other.

where in Eq. (1.10), a current much larger than one of I . In this case, the lines of B are loops around the wire, as shown in Fig. 1-7. This B -field is responsible for the force on the magnet.

Equation (1.9) tells us that for a fixed current through the wire, the magnitude of B is the same for any curve that surrounds the wire. For curves—say circles—that are farther away from the wire, the circumference is larger, so the tangential component of B may decrease. You can see this in Fig. 1-7; in fact, expect B to decrease inversely with the distance from a long straight wire.

Now, we have said that a constant current in the wire produces a magnetic field, and that when there is a magnetic field present, there is a force on a wire carrying a current. But we should also expect that if we make a magnet with a current in one wire, it should exert a force on another wire which also carries a current. This can be shown by using two hanging wires, as shown in Fig. 1-8. When the magnets are in the same direction, the two wires repel, because the currents are opposite, they repel.

Is direct electrical current as well as magnetism, make magnetic fields? But wait, what is a magnet, anyway? If magnetic fields are produced by moving charges, is it not possible that the magnetic field from a piece of iron is really the result of currents? It appears to be so. We can replace the bar magnet of our experiment with a coil of wire, as shown in Fig. 1-9. When a current is passed through the coil—as well as through the straight wire above it—we observe a motion of the wire exactly as before, when we had a magnet instead of a coil. In other words, the current in the coil creates a magnet. It happens, then, that a piece of iron acts as though it contains a permanent circulating current. We can, in fact, understand magnets in terms of permanent currents in the atoms of the iron. The force on the magnet in Fig. 1-7 is due to the second term in Eq. (1.1).

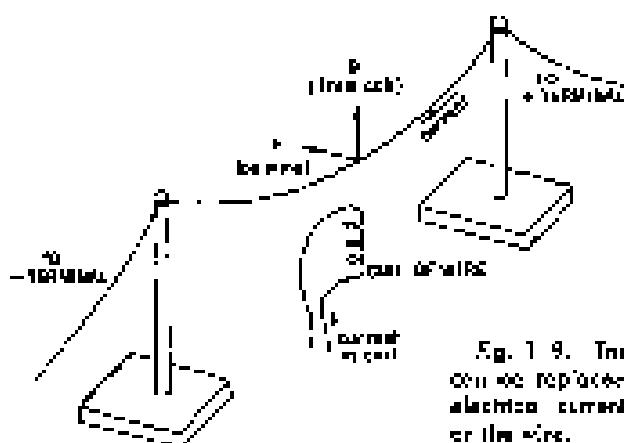


Fig. 1-9. The bar magnet of Fig. 1-6 can be replaced by a coil carrying an electric current. A similar force acts on the wire.

Where do the currents come from? One possibility would be from the motion of the electrons in atomic orbits. Actually, that is not the case for iron, although it is for some materials. In addition to rotating around its atoms, an electron also spins about on its own axis—anything like the spin of the earth—and it is the current along this spin that gives the magnetic field to iron. (We say “rotating like the spin of the earth” because the quantum mechanics tell us that the classical ideas do not really describe things very well.) In most substances, some electrons will rotate with some spin the other, so the magnetism cancels out, but in iron—for a mysterious reason which we will discuss later—many of the electrons are spinning with their axes lined up, and that is the source of the magnetism.

Since the fields of magnets are from currents, we do not have to add any extra term to Eqs. (1.8) or (1.9) to take care of magnetism. We just take all currents, including the circulating currents of the spinning electrons, and then the law is right. You should also notice that Eq. (1.8) says that there are no magnetic “charges” analogous to the electrical charges appearing on the right side of Eq. (1.6). None has been shown.

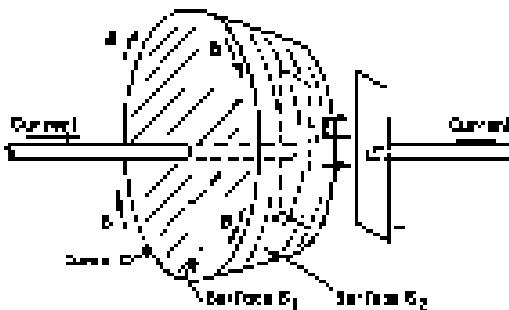


Fig. 1-16. The circulation of \mathbf{B} around the curve C is given either by the current passing through the surface S_1 , or by the rate of change of the flux of \mathbf{E} through the surface S_2 .

The first term on the right-hand side of Eq. (1.5) was discovered theoretically by Maxwell and is of great importance. It says that changing electric fields produce magnetic effects. In fact, without this term the equation would not make sense, because without it there would be no currents in circuits that are not complete loops. But such currents do exist, as we can see in the following example. Imagine a capacitor made of two flat plates. It is being charged by a current that flows between one plate and away from the other, as shown in Fig. 1-18. We draw a curve C around one of the wires and fit it in with a surface which covers the wire as shown by the surface S_1 in the figure. According to Eq. (1.9), the circulation of \mathbf{B} around C is given by the current in the wire times c^2 . But what if we fit in the curve with a different surface S_2 , which is shaped like a bowl and passes between the plates of the capacitor, staying always away from the wire? There is certainly no current through this surface. So, surely, just changing the location of an imaginary surface is not going to change a real magnetic field! The circulation of \mathbf{B} must be what it was before. The last term on the right-hand side of Eq. (1.5) does, indeed, combine with the second term to give the same result for the two surfaces S_1 and S_2 . For S_2 the circulation of \mathbf{B} is given in terms of the rate of change of the flux of \mathbf{E} between the plates of the capacitor. And it works out that the changing E is related to the current in just the way required for Eq. (1.9) to be correct. Maxwell saw that it was needed, and so was the desire to write the complete equation.

With the setup shown in Fig. 1-6 we can remember another of the laws of electromagnetism. We disconnected the ends of the bridging wire from the battery and connect them to a galvanometer which tells us when there is a current through the wire. When we push the wire sideways through the magnetic field of the magnet, we observe a current. Such an effect is again just another consequence of Eq. (1.1). The electrons in the wire feel the force $F = q\mathbf{v} \times \mathbf{B}$. The electrons have a sideways velocity because they move with the wire. This v with a vertical B from the magnet results in a force on the electrons directed along the wire, which makes the electrons moving toward the galvanometer.

Suppose, however, that we leave the car alone and move the magnet. We guess from relativity that it should make no difference, and indeed, we observe a similar current in the galvanometer. This means the moving coil produces forces on charged ones? According to Eq. (1.2) there must be an electric field. A moving magnet must make an electric field. How does it happen is told quantitatively by Eq. (1.7). This equation describes many phenomena of great practical interest such as those that occur in electric generation and transmission.

The most remarkable consequence of our equations is that the combination of Eq. (1.7) and Eq. (1.9) embodies the propagation of the waveform of electromagnetic effects over large distances. The reason is roughly something like this: suppose that somewhere we have a magnetic field which is oscillating because, say, a current is turned on suddenly in a wire. Then by Eq. (1.7) there must be a circulation of an electric field. As the electric field builds up to produce its circulation, then according to Eq. (1.9) a magnetic circulation will be generated. But the building up of this magnetic field will produce a new circulation of the electric field, and so on. In this way fields track their way through space without the trace of charges or currents except at their source. That is the way we see each other! It is all in the equations of the electromagnetic field.

1-5 What are the fields?

We now make a few remarks on the way of looking at this subject. You may be saying, "All this business of lines and distributions is pretty abstract. There are electric fields at every point in space. What does one these 'lines' or 'fields' actually represent?" Why can't you explain it, for instance, by whatever it is that goes between the charges? Well, it depends on your prejudices. Many physicists used to say that direct action with nothing in between was incomprehensible. (How could they find another incomprehensible why it had already been given?) They would say: "Look, the only choices we have is the above account of two pieces of matter interacting. It's impossible that there can be a force with nothing in between!" You what really happens when we apply the "cancellation" of one piece of matter right against another? We discover that it is not one piece right against the other; they are slightly askew, and there are electrical forces acting on a tiny scale. Thus we find that we are going to begin to call other interaction in terms of the picture for electrical forces. It is certainly not sensible to try to insist that an electric force has to look like the old. Similarly, a regular push is pull when we call out that the molecular pushes and pulls are going to be interpreted as electrical forces! The only sensible question is what is the most convenient way to look at electrical fields. Some people prefer to represent them as the interaction of a charge of charge, and to renounce compensated law. Others like the field lines. They draw field lines all the time, and look at things as if they were isolated. The field lines, however, are only a crude way of doing it in a field, and it is very difficult to give the correct quantification laws exactly in terms of field lines. Also, the laws of the field lines do not contain the deepest principle of electrodynamics, which is the superposition principle. Even though we know how the field lines look for one set of charges and what the field lines look like for another set of charges, we don't get any idea about what the field lines produce will look like when both sets are present together. From the mathematical stand point, on the other hand, superposition is easy—you simply add the two vectors. The field lines have some advantages in giving a visual picture, but they also have some disadvantages. The direct interaction view I think is having advantages when thinking of electrical charges at rest, but has great disadvantages when dealing with charges in motion.

The best way is to use the electric field law. That is also not superfluous, but necessary. The attempts to try to represent the electric field as the motion of some kind of particles, or in terms of forces, or of stresses in some kind of material have made to more effort of physicists than it would have taken simply to get the right answers about electrodynamics. It is interesting that the various experiments for the behavior of light in crystals were worked out by McCullough in 1970. (See

people said to him: "Yes but there is no real material whose mechanical properties could possibly satisfy those equations, and since light is an oscillating field, most vibrations are oscillations. We cannot believe this abstract electrical business." If people had been more open-minded, they might have believed in the right equations for the behavior of light & let science have to say the

In the case of the magnetic field we can make the following point: Suppose you finally succeeded in making up a picture of the magnetic field in terms of some kind of lines or of gear wheels running through space. Then you try to explain what happens to two charges moving in space, both of the same speed and parallel to each other. But as they are moving, they will behave like two cars and will have a magnetic field associated with them like the air pressure in the wind, as Fig. 1-1). An observer who was riding along with the two charges, however, would see both charges as stationary and would say that there is no magnetic field. The "gear wheels" or "lines" disappear when you ride along with the object. All we have done is to invent a new problem. How can the gear wheels disappear? The people who have light lines are in a similar difficulty. Not only is it not possible to say where the field lines move or do not move with charges—they may disappear completely in certain coordinate frames.

All that we are saying here is that magnetism is really a relativistic effect. In the case of the two charges we just considered, travelling parallel to each other, we would expect to have to make relativistic corrections to their motion, with terms of order v^2/c^2 . These corrections would correspond to the magnetic force. But what about the force between the two wires in the experiment (Fig. 1-2)? There the magnetic force is the whole force. It didn't look like a "relativistic correction." Also, if we estimate the velocities of the electrons in the wire (you can do this yourself), we find that their average speed along the wire is about 0.01 centimeter per second. So v/c^2 is about 10^{-12} . Surely a negligible "correction." But not! Although the magnetic force is, in this case, 10^{-11} of the "normal" electrical force between the moving electrons, remember now that "normal" electrical forces have disappeared because of the almost perfect balancing out because the wires have the same number of positive and negative. The balance is much more precise than one part in 10^{12} , and the small relativistic force which we call the magnetic force is the only one left. It becomes the dominant force.

It is the near-perfect cancellation of electric effects which allowed relativity theory (*i.e.* in magnetism) to be trusted and the correct equations—the order v^2/c^2 to be discovered, even though scientists didn't know that's what was happening. And that is why, when relativity was discovered, the electromagnetic law didn't need to be changed. Only—unlike mechanics—it was correct to a precision of v^2/c^2 .

1.6 Electromagnetism in science and technology

Let us end this chapter by pointing out that among the many phenomena studied by the Greeks there were two very strange ones: that if you rubbed a piece of amber you could lift up little pieces of paper, and that there was a strange rock from the island of Magnesia which attracted iron. It is interesting to think that these were the only phenomena known to the Greeks in which the effects of electricity or magnetism were apparent. The reason that these were the only phenomena that appeared is due primarily to the fortunate position of the balancing of charges that we mentioned earlier. Study by successive steps after the Greeks intersected one new phenomenon after another that were really quite typical of these amber and/or lodestone effects. Now we realize that the phenomena of chemical interaction and, ultimately, of life itself are to be understood in terms of electromagnetism.

At the same time that an understanding of the subject of electromagnetism was being developed, historical necessities that dictated the explanation of the people that came before were appearing. It became possible to signal by taking up one long distance, and to talk to another person miles away without any connection between, and to run huge power systems—a gear motor wheel, connected by

firearm over hundreds of miles to another engine that runs in resonance in the center wheel—every thousands of branching elements—an incandescent engine in ten thousand places running the machinery of industry and homes—burning because we have knowledge of the laws of Electromagnetism.

Today we are applying even more subtle effects. The electrical forces, enormous as they are, can also be very tiny, and we can control them, and use them in very many ways. So delicate are our instruments that we can tell what a wire is doing by the way it affects the electrons in a thin metal rod hundreds of miles away. All we need to do is to put the rod as an antenna for a television receiver!

From a long view of the history of mankind—seen from, say, ten thousand years from now—there can be little doubt that the most significant event of the 20th century will be judged as Maxwell's discovery of the laws of charge dynamics. The American Civil War will pale into provincial significance in comparison with this important scientific event of the same decade.

Differentiated Function of Vector Fields

2-1 Understanding physics

The physicist needs a facility in looking at problems from outside points of view. The exact analysis of real physical problems is nearly quite complicated, and any particular physical situation may be too complicated to analyze directly by solving the differential equation. But one can still get a very good idea of the behavior of a system if one has some feel for the character of the solution in different circumstances. Tools such as the field lines, capacitors, resistance, and inductance are, for such purposes, very useful. So we will spend much of our time analyzing them. In this way we will get a feeling for what should happen in different electromagnetic situations. On the other hand, some of the heuristic models, such as field lines, is really adequate and accurate for all situations. There is only one possible way of presenting the laws, and that is by means of differential equations. They have the advantage of being fundamental and, as far as we know, precise. If you have learned the differential equations, you can always go back to them. There is nothing to unknown.

It will help you twice now to understand what should happen in different circumstances. You will have to solve the equations. Each time you solve the equations, you will learn something about the character of the solutions. To keep these solutions in mind, it will be useful also to carry their meaning in terms of field lines and of other concepts. This is the way you will really "understand" the equations. That is the difference between mathematics and physics. Mathematicians, or people who have only mathematical minds, are often led away from "studying" physics because they lose sight of the physics. They say, "Look, these hideous equations—the Maxwell equations—are all there is in electromagnetism; i. e. is admitted by the physics and there is nothing else is not contained in the equations. The equations are complicated, but after all they are only mathematical equations and if I understand them mathematically, that's all, I will understand the physics inside out." Only it doesn't work that way. Mathematicians who only play with that pile of new and there have been many of them usually make little contribution to physics, even in fact little in themselves. They fail because the real physical situations in the real world are so complicated that it is necessary to have a much broader understanding of the equations.

What it means really to understand an equation—that is, in more than a strictly mathematical sense—was described by Dirac. He said: "I understand what an equation means if I have a way of figuring out the characteristics of its solution without actually solving it." So if we have a way of knowing what should happen in given circumstances without actually solving the equations, then we "understand" the equations as applied to these circumstances. A physical understanding is a completely nonmathematical, intuitive, and incisive thing, but absolutely necessary for a physicist.

Ordinarily, a course like this is given by developing gradually the physical ideas—by starting with simple situations and going on to more and more complicated situations. This requires that you continuously forget things you previously learned. Things that are true in certain situations, but which are not true in general. For example, the "law" that the electrical force depends on the square of the distance is not always true. We prefer the opposite approach. We prefer to prove first the complete laws, and then to show how and apply them to simple situations, developing the physical ideas as we go along. And that is what we are going to do.

2-1 Understanding physics

2-2 Scalar and vector fields—I and II

2-3 Derivative of fields—the gradient

2-4 The operator ∇

2-5 Operations with ∇

2-6 The differential equations of heat flow

2-7 Second derivatives of vector fields

2-8 Fields

Review: Chapter 1, Vol. I, Errors

Our approach is completely opposite to the historical approach, in which one coverage the subject in terms of the experiments by which the information was obtained. But the subject of physics has been developed over the past 200 years by some very important people, and as we have only a limited time to deduce our knowledge, we cannot possibly cover everything they did. Unfortunately one of the things that we shall have a tendency to lose in these lectures is the historical, experimental development. It is hoped that in the subsequent course of this book, can be corrected. You can also fill in what we must leave out by reading the Physics department's Bulletin, which has excellent historical articles on electricity and on other parts of physics. You will also find historical information in many textbooks on electricity and magnetism.

3-3 Scalar and vector fields—Part A

We begin now with the abstract mathematical view of the theory of electricity and magnetism. The ultimate aim is to deduce the meaning of the laws given in Chapter 1. In order to do this we must first explain a few new notation that we want to use. So let us begin electromagnetism for the moment and discuss the mathematics of vector fields. This is very useful both because, not only for electromagnetism, but for all kinds of physical circumstances. Just as ordinary differential and integral calculus is so important to all branches of physics, so also is the differential calculus of vectors. We turn to that subject.

Below follow some few facts from the algebra of vectors. It is assumed that you already know them.

$$A \times B = \text{scalar} \cdot (A_1 B_2 - A_2 B_1) = A_1 B_2 - A_2 B_1 \quad (2.1)$$

$$A \times B = \text{vector} \quad (2.2)$$

$$(A \times B)_1 = A_1 B_2 - A_2 B_1$$

$$(A \times B)_2 = A_1 B_3 - A_3 B_1$$

$$(A \times B)_3 = A_2 B_3 - A_3 B_2$$

$$A \times J = 0 \quad (2.3)$$

$$J \cdot (A \times B) = 0 \quad (2.4)$$

$$A \cdot (\nabla \times C) = (A \times B) \cdot C \quad (2.5)$$

$$A \times (B \times C) = B(A \cdot C) - C(A \cdot B) \quad (2.6)$$

Small systems are harder.

Also we will want to use the two following equations from the calculus:

$$\Delta f(x, y, z) = \frac{\partial^2 f}{\partial x^2} \Delta x + \frac{\partial^2 f}{\partial y^2} \Delta y + \frac{\partial^2 f}{\partial z^2}, \quad (2.7)$$

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}. \quad (2.8)$$

The first equation (2.7) is, of course, true only in the limit that Δx , Δy , and Δz are small enough.

The simplest possible physical field is a scalar field. By a field, you remember, we mean a quantity which depends upon position in space. By a scalar field we merely mean a field which is characterized at each point by a single number—a scalar. Of course the number may change its value, but we need not worry about that for the moment. We will talk about what the field looks like at a given instant. As an example consider a solid cube of material which has been heated at some places and cooled at others, so that the temperature of the cube varies from point to point in a complicated way. Then the temperature will be a function of x , y , and z , the position in space measured in a regular grid convenient system. Temperature is a scalar field.

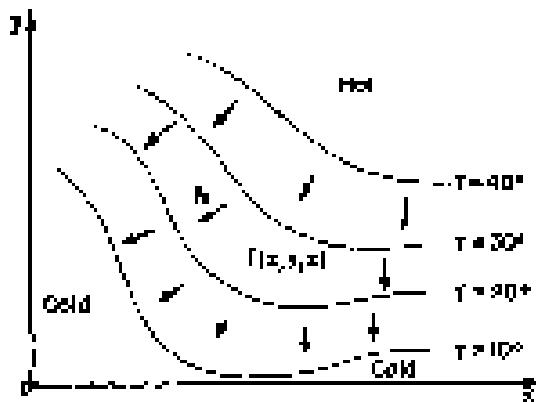


Fig. 2-1. Temperature T is an example of a scalar field. With each point (x, y, z) in space there is associated a number $T(x, y, z)$. All points on the surface marked $T = 20^\circ$ lie on a curve or "isotherm" at the same temperature. The arrows are samples of the heat flow vector \mathbf{h} .

One way of thinking about scalar fields is to imagine "contours" which are imaginary surfaces drawn through all points for which the field has the same value, just as contour lines on a map connect points with the same height. For a temperature field the contours are T and "isothermal surfaces" are "isotherms". Figure 2-1 illustrates a temperature field and shows the dependence of T on x and y when $z = 0$. Several isotherms are drawn.

There are also vector fields. The idea is very simple. A vector is given for each point in space. The vector varies from point to point. As an example, consider a rotating body. The velocity of the material of the body at any point is a vector which is a function of position (Fig. 2-2). As a second example, consider the flow of heat in a block of material. If the temperature in the block is constant in one place and free at another, there will be a flow of heat from the latter place to the former. The heat will be flowing in different directions in different parts of the block. This heat flow is a directional quantity which we call \mathbf{h} . Its magnitude is a measure of how much heat is flowing. Examples of the heat flow vector are also shown in Fig. 2-1.

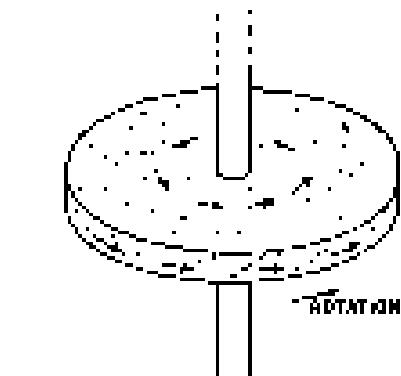
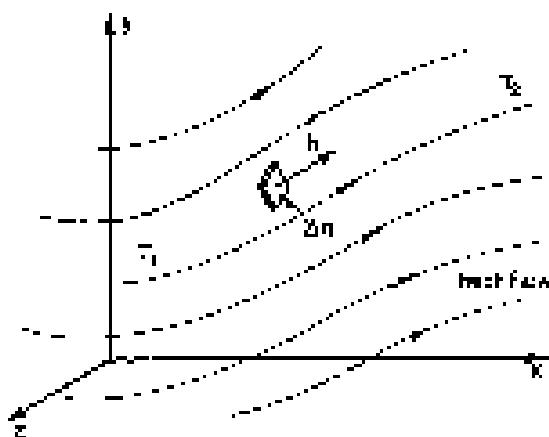


Fig. 2-2. The velocity of the atoms in a rotating object is an example of a vector field.

Let's take a more precise definition of \mathbf{h} : the magnitude of the vector \mathbf{h} at a point is the amount of thermal energy that passes, per unit time and per unit area, through an infinitesimal surface element, at right angles to the direction of flow. The vector points in the direction of flow (see Fig. 2-3). In symbols: If dA is the elemental area that passes per unit time through the surface element, then

$$h = \frac{dQ}{dA} \text{ W/m}^2 \quad (2-9)$$

where q_1 is a unit vector in the direction of flow.

The vector \mathbf{h} can be defined in another way — in terms of its components. We ask how much heat flows through a small surface at an angle with respect to the flow. In Fig. 2-4 we show a small surface dA outlined with dashed lines, which is perpendicular to the flow. The unit vector n is normal to the surface dA . The

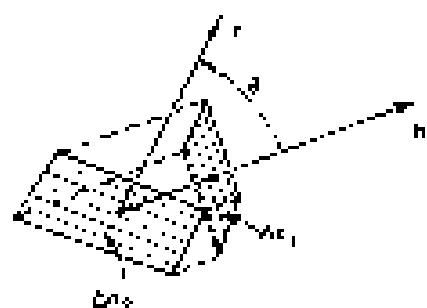


Fig. 2-4. The heat flow through dA is the rate of flow through dA .

angle between \hat{m}_1 and \hat{n} is the same as the angle between the heat flux (normal to \hat{m}_1) and \hat{n} . Now \hat{m}_1 is the heat flow per unit area through \hat{m}_1 . The flow through \hat{m}_1 is the sum of through \hat{m}_1 ; only the areas are different. In fact, $\hat{m}_1 = \hat{m}_2$ since the heat loss through \hat{m}_1 is

$$\frac{\partial T}{\partial n} = \frac{\partial T}{\partial \hat{m}_1} \cos \theta = \hat{m}_1 \cdot \hat{n} \quad (2.10)$$

We increase this equation: the heat flow (per unit time and per unit area) through any surface element whose unit normal is \hat{n} is given by $\hat{m} \cdot \hat{n}$. Equally, we could say: the component of the heat flow perpendicular to the surface element \hat{m}_1 is $\hat{m}_1 \cdot \hat{n}$. We can, if we wish, consider that these statements about \hat{m} are valid by applying the same ideas to other vector fields.

2.5 Derivatives of fields - the gradient

When fields vary in time, we can give the variation by giving their derivatives with respect to t . We want to describe the variations with respect to a similar way, because we are interested in the relationship between ΔT , the temperature at one place and the temperature at a nearby place. Then shall we take the derivative of the temperature with respect to position? Do we differentiate the temperature with respect to x_1 , or with respect to x_2 , or x_3 ?

Theful physics' law does not expand upon the orientation of the coordinate system. It says simply, therefore, *at position x there is a vector field whose components are scalar functions of the coordinates*. What is the derivative of a scalar field, say $\partial f / \partial x_1$? Is it scalar, or a vector, or what? It is neither a scalar nor a vector, so you can easily appreciate, because if we took a different x -axis, $\partial f / \partial x$ would certainly be different. But notice: We have three possible derivatives, $\partial f / \partial x_1$, $\partial f / \partial x_2$, and $\partial f / \partial x_3$. These three are three kinds of derivatives and we know that it takes three numbers to form a vector, perhaps these three derivatives are the components of a vector.

$$\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3} \right) \equiv \text{a vector} \quad (2.11)$$

Of course, it is not generally true that three numbers form a vector. It is true only if, when we rotate the coordinate system, the components of the vector transform like the components of the coordinate system. So it is necessary to analyse how these derivatives change by a rotation of the coordinate system. We shall show that (2.11) is indeed a vector. The derivatives do transform in the correct way when the coordinate system is rotated.

We can approach in two ways. One way is to ask a question whose answer is independent of the coordinate system, and try to express the answer as f . "It varies." From, for instance, $f = A \cdot R$, and A and R are vectors we know - however we proved it in Chapter 11 of Vol. I - that f is a scalar. We know that f is a scalar without investigating whether it changes with changes in coordinate system. It can't, because it's a dot product of two vectors. Similarly, if we know that A is a vector, and we have three numbers A_1 , A_2 , and A_3 , and we find out that

$$A \cdot B = A_1 B_1 + A_2 B_2 + A_3 B_3 = S \quad (2.12)$$

where S is the same for any coordinate system, then it must be that the three numbers B_1 , B_2 , B_3 are the components B_x , B_y , B_z of some vector B .

Now it's time to talk of the temperature field. Suppose we take two points P_1 and P_2 , separated by the small interval ΔR . The temperature at P_1 is T_1 and at P_2 is T_2 . Then $\Delta T = T_2 - T_1$. The temperature at these real, physical points certainly do not depend on what axes we choose for measuring the coordinates. In particular, ΔT is a number independent of the coordinate system. It is a scalar.

If we choose some convenient set of axes, we could write $T_1 = T(x, y, z)$ and $T_2 = T(x + \Delta x, y + \Delta y, z + \Delta z)$, where Δx , Δy , and Δz are the components of the vector ΔR (Fig. 2-5). Remembering Eq. (2.7), we can write

$$\Delta T = \frac{\partial T}{\partial x} \Delta x + \frac{\partial T}{\partial y} \Delta y + \frac{\partial T}{\partial z} \Delta z. \quad (2.13)$$

The left side of Eq. (2.13) is a scalar. The right side is the sum of three products times Δx , Δy , and Δz , which are the components of a vector. It follows that the three numbers

$$\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z}$$

are also the x , y , and z -components of a vector. We write this new vector with the symbol ∇T . The symbol ∇ (called "del") is not upside down ∂ , but is supposed to remind us of differentiation. Perhaps read " ∇T " in various ways: "del-T" or "gradient of T ," or "grad T ."

$$\text{grad } T = \nabla T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z} \right)^T. \quad (2.14)$$

Using the notation, we can rewrite Eq. (2.13) in the more compact form

$$\Delta T = \nabla T \cdot \Delta R. \quad (2.15)$$

In Fig. 2-5 this equation says that the difference in temperature between two nearby points is the dot product of the gradient of T and the vector displacement between the points. The form of Eq. (2.15) also illustrates clearly our point above that ∇T is indeed a vector.

Perhaps you are still not convinced? Let's prove it in a different way (and—though if you look carefully, you may be able to reconstruct it's today the form of ∇ a longer winded form). We shall show that the components of ∇T transform like just the three x , y , z -components of ΔR do. If they do, ∇T is a vector according to our original definition of a vector in Chapter 1 of Vol. 1. We pass to a new coordinate system x' , y' , z' , and in this new system we calculate $\partial T/\partial x'$, $\partial T/\partial y'$, and $\partial T/\partial z'$. To make things a little easier, we let $x' = x - x_0$ so that we can forget about the x -coordinate. (You can check out the more general case for yourself.)

We take an $x'y'z'$ -system rotated an angle θ with respect to the xz -system, as in Fig. 2-6(a). For a point (x, y, z) the coordinates in the $x'y'z'$ -system are

$$x' = x \cos \theta - y \sin \theta, \quad (2.16)$$

$$y' = -x \sin \theta + y \cos \theta. \quad (2.17)$$

Or, solving for x and y ,

$$x = x' \cos \theta + y' \sin \theta, \quad (2.18)$$

$$y = x' \sin \theta - y' \cos \theta. \quad (2.19)$$

If any pair of numbers transforms via these equations in the same way that x and y do, they are the components of a vector.

Now let's look at the difference in temperature between the two nearby points P_1 and P_2 , shown as in Fig. 2-6(b). If we calculate with the $x'y'z'$ -coordinates, we would find

$$\Delta T = \frac{\partial T}{\partial x'} \Delta x' \quad (2.20)$$

with $\Delta x' = \Delta x$.

* In computation, the expression (a, b, c) represents a vector with components a , b , and c . If you like to use the unit vectors i , j , and k you may write

$$av = a \frac{\partial v}{\partial x} + b \frac{\partial v}{\partial y} + c \frac{\partial v}{\partial z}$$



Fig. 2-5. The vector ΔR , whose components are Δx , Δy , and Δz .

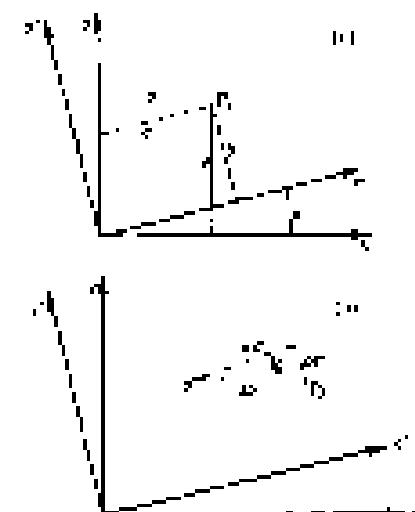


Fig. 2-6. (a) Transformation to a rotated coordinate system. (b) Special case of an interval ΔR parallel to the x -axis.

If we choose some convenient set of axes, we could write $T_1 = T(x, y, z)$ and $T_2 = T(x + \Delta x, y + \Delta y, z + \Delta z)$, where Δx , Δy , and Δz are the components of the vector ΔR (Fig. 2-5). Recalculating Eq. (2.7), we can write

$$\Delta T = \frac{\partial T}{\partial x} \Delta x + \frac{\partial T}{\partial y} \Delta y + \frac{\partial T}{\partial z} \Delta z. \quad (2.13)$$

The left side of Eq. (2.13) is a scalar. The right side is the sum of three products times Δx , Δy , and Δz , which are the components of a vector. It follows that the three numbers

$$\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z}$$

are also the x , y , and z -components of a vector. We write this new vector with the symbol ∇T . The symbol ∇ (read "del") is at upside down a , and is supposed to remind us of differentiation. People read ∇T in various ways: "del-T" or "gradient of T " or "grad T ".

$$\text{grad } T = \nabla T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z} \right)^T. \quad (2.14)$$

Using the notation, we can rewrite Eq. (2.13) in the more compact form

$$\Delta T = \nabla T \cdot \Delta R. \quad (2.15)$$

In Fig. 2-5 this equation says that the difference in temperature between two points is the dot product of the gradient of T and the vector displacement between the points. The form of Eq. (2.15) also illustrates clearly our point above that ∇T is indeed a vector.

Do larger vectors still transform? Let's prove it in a different way (and, though if you look carefully, you may be able to reconstruct it's today the form of a longer winded form). We shall show that the components of ∇T transform just the same way that components of \vec{R} do. If they do, ∇T is a vector according to our original definition of a vector in Chapter 1 of Vol. 1. We pass to a new coordinate system x' , y' , z' , and in this new system we calculate $\partial T/\partial x'$, $\partial T/\partial y'$, and $\partial T/\partial z'$. To make things a little easier, we let $x' = x - x_0$ so that we can forget about the x -coordinate. (You can check out the more general case for yourself.)

We take an $x'y'z'$ -triangle, shown in Fig. 2-6(a) with respect to the x -system, as in Fig. 2-6(b). For a point (x, y, z) the coordinates in the prime system are

$$x' = x \cos \theta - y \sin \theta, \quad (2.16)$$

$$y' = -x \sin \theta + y \cos \theta. \quad (2.17)$$

Or, solving for x and y ,

$$x = x' \cos \theta + y' \sin \theta, \quad (2.18)$$

$$y = x' \sin \theta - y' \cos \theta. \quad (2.19)$$

If any pair of numbers transforms via these equations in the same way that x and y do, they are the components of a vector.

Now let's look at the difference in temperature between the two nearby points P_1 and P_2 , shown as in Fig. 2-6(b). If we calculate with the $x'y'z'$ coordinates, we would write

$$\Delta T = \frac{\partial T}{\partial x'} \Delta x' \quad (2.20)$$

with $\Delta x' = \Delta x$.

* In computation, the expression (a, b, c) represents a vector with components a , b , and c . If you like to use the unit vectors i , j , and k you may write

$$wv = i \frac{\partial T}{\partial x} + j \frac{\partial T}{\partial y} + k \frac{\partial T}{\partial z}$$

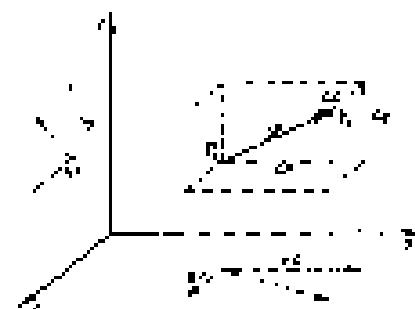


Fig. 2-5. The vector ΔR , whose components are Δx , Δy , and Δz .

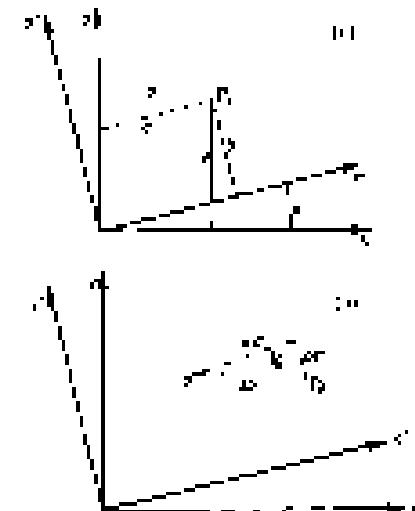


Fig. 2-6. (a) Transformation to a rotated coordinate system. (b) Special case of an interval ΔR parallel to the x -axis.

This would be a continuation of the point system, part 7. We would have gotten

$$\Delta T = \frac{\partial T}{\partial x} \Delta x + \frac{\partial T}{\partial y} \Delta y. \quad (2.21)$$

Looking at Fig. 2.4(b), we see that

$$\Delta x = \Delta r \cos \theta \quad (2.22)$$

and

$$\Delta y = \Delta r \sin \theta. \quad (2.23)$$

Note Δx is negative when θ is positive. Substituting these in Eq. (2.21), we find that

$$\Delta T = \frac{\partial T}{\partial x} \Delta x + \Delta y \frac{\partial T}{\partial y} \quad (2.24)$$

$$= \left(\frac{\partial T}{\partial x} \cos \theta + \frac{\partial T}{\partial y} \sin \theta \right) \Delta r. \quad (2.25)$$

Comparing Eqs. (2.25) with (2.20), we get that

$$\frac{\partial T}{\partial r} = \frac{\partial T}{\partial x} \cos \theta + \frac{\partial T}{\partial y} \sin \theta. \quad (2.26)$$

This equation says that $\partial T/\partial r$ is obtained from $\partial T/\partial x$ and $\partial T/\partial y$, just as x is obtained from x' and y' in Eq. (2.18). But $\partial T/\partial x$ is the *x-component* of a vector. The same kind of arguments would show that $\partial T/\partial y$ and $\partial T/\partial z$ are *y- and z-components*. So ∇T is definitely a vector. It is a vector field, derived from the scalar field T .

2.4 The operator ∇

Now we can do something that is extremely exciting and ingenious—and characteristic of the things that make mathematics beautiful. The argument that $\partial T/\partial r$, $\partial T/\partial x$, ∇T , is a vector did not depend upon this scalar field, T , or r ; it didn't matter. All the arguments would go through if T were replaced by any other field. Since the transformation equations are the same no matter what T differences, we could just as well call the T and replace Eq. (2.26) by the operator equation

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial x'} \cos \theta - \frac{\partial}{\partial y'} \sin \theta. \quad (2.27)$$

We leave the operators $\cos \theta$ and $\sin \theta$ “hungry for something to differentiate.”

Since the derivative operator transforms like the components of a vector should, we can call them components of a vector operator. We can write the

$$\nabla = \left(\frac{\partial}{\partial x'}, \frac{\partial}{\partial y'}, \frac{\partial}{\partial z'} \right), \quad (2.28)$$

which means, of course,

$$\nabla_x = \frac{\partial}{\partial x}, \quad \nabla_y = \frac{\partial}{\partial y}, \quad \nabla_z = \frac{\partial}{\partial z}. \quad (2.29)$$

We have extracted the gradient away from the T —there is the wonder!

You must always remember, of course, that ∇ is an operator—alone. It means nothing. ∇ by itself means nothing; what does it mean if we multiply it by a scalar—say T ?—to get the product $T\nabla$? (One can always multiply a vector by a scalar.) It still doesn't mean anything. (It's *noncommutative*.)

$$T \frac{\partial}{\partial x}, \quad (2.30)$$

where T is not necessarily the *last* term of a product. However, according to the *rule of vectors* we would still call $T\nabla$ a vector.

Now let's multiply ∇ by a scalar on the right side, and we have the product (∇T). In ordinary algebra

$$\nabla A = AT. \quad (2.11)$$

but we have to remember that covariant algebra is a little different from ordinary vector algebra. With operators we must always keep the sequence right, so that the operator comes first. This will become clear if you just remember that the operator ∇ obeys the same commutator as has derivative calculus. What is to be differentiated must be placed on the right of the ∇ . For one is important.

Keeping in mind this point, in order to understand that ∇T is an operator, but the product $T\nabla$ is no longer a hungry operator, the operator is completely stabilized. It becomes a physical vector having a meaning. It represents the spatial rate of change of T . The components of ∇T is the rate T changes in the direction. When we are directional of the vector ∇T , we know that the rate of change of T in the direction is the component of ∇T in that direction (see Eq. 2.12). It follows that the direction of ∇T is that in which it has the largest possible magnitude—in other words, the direction in which T changes the fastest. The gradient of T has the slope of the steepest uphill slope on T .

2.5 Operators with ∇

Consider any other algebra with the scalar operator T^A . Let's try combining it with a vector. We can combine two vectors by making a dot product. We could make the products

$$\text{for example } \nabla \cdot T \quad \text{or} \quad \nabla \cdot (T \text{ times})$$

The first one doesn't mean anything yet, because T is still an operator. What it might ultimately mean would depend on what T is taken to operate on. The second product is somewhat absurd. ($A \cdot B$ is always a scalar.)

Let's say the dot product of ∇ with a vector field \mathbf{A} known, say \mathbf{A} . We write out the components:

$$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}. \quad (2.12)$$

or

$$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x'} + \frac{\partial A_y}{\partial y'} + \frac{\partial A_z}{\partial z'}. \quad (2.13)$$

This sum is invariant under a coordinate transformation. If we were to choose a different system (indicated by primes), we would have

$$\nabla' \cdot \mathbf{A}' = \frac{\partial A'_x}{\partial x'} + \frac{\partial A'_y}{\partial y'} + \frac{\partial A'_z}{\partial z'}, \quad (2.14)$$

which is the same number as would be gotten from Eq. (2.12), even though it looks different. That is,

$$\nabla' \cdot \mathbf{A}' = \nabla \cdot \mathbf{A} \quad (2.15)$$

for every point in space. So $\nabla \cdot \mathbf{A}$ is a scalar field, which must represent some physical quantity. You should realize that the combinations of derivatives in $\nabla \cdot \mathbf{A}$ is rather special. There are all sorts of other combinations like $\nabla \times \mathbf{A}$, which are neither scalars nor components of vectors.

The scalar quantity $\nabla \cdot \mathbf{A}$ (or $\nabla \cdot \mathbf{v}$) is extremely useful in physics. It has been given the name "divergence." For example,

$$\nabla \cdot \mathbf{A} = \text{rate } \mathbf{A} \text{ "dissipates" } \quad (2.16)$$

In the next section we can associate a physical significance to $\nabla \cdot \mathbf{A}$. We shall, however, postpone that until later.

¹We think of A as a scalar quantity that depends on position in space, and not entirely like mathematical functions of three variables. When A is "differentiable" with respect to x, y , and z , it is an infinitely differentiable function, and " ∇ " the mathematical expression for it must then be expanded into a function of the appropriate variables.

First we wish to see what else we can cook up with the vector cross $\nabla \times \mathbf{A}$. What about a cross product? We must expect that

$$\nabla \times \mathbf{A} = \text{vector.} \quad (2.37)$$

If it's a vector whose components we can write by the usual rules for cross products (see Eq. 2.36)

$$(\nabla \times \mathbf{A})_x = A_y \partial_z - A_z \partial_y = \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial y}. \quad (2.38)$$

Similarly,

$$(\nabla \times \mathbf{A})_y = A_z \partial_x - A_x \partial_z = \frac{\partial A_y}{\partial z} - \frac{\partial A_x}{\partial y}, \quad (2.39)$$

and

$$(\nabla \times \mathbf{A})_z = A_x \partial_y - A_y \partial_x = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial x}. \quad (2.40)$$

The combination $\nabla \times \mathbf{A}$ is called "the curl of \mathbf{A} ". The reason for the name and the physical meaning of the combination will be discussed later.

Summarizing, we have three kinds of combinations with ∇ :

$$\nabla T = \text{scalar } T = \text{scalar,}$$

$$\nabla \cdot \mathbf{B} = \nabla \cdot \mathbf{B} = \text{scalar,}$$

$$\nabla \times \mathbf{A} = \text{curl } \mathbf{A} = \text{vector.}$$

Using these combinations, we can write down the spatial variations of fields in a conductor; why?... a way that is general, i.e. that doesn't depend on any particular set of axes.

As an example of a law of nature for sufficiently-pure \mathbf{E} , we might get a situation where eq. above which contains the same laws of electromagnetism that we have in words in Chapter 1. That is called Maxwell's equations.

Maxwell's equations

$$\begin{aligned} (1) \quad & \nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \\ (2) \quad & \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \\ (3) \quad & \nabla \cdot \mathbf{B} = 0 \\ (4) \quad & \nabla \times \mathbf{B} = \frac{\partial \mathbf{E}}{\partial t} + \frac{\mathbf{j}}{\mu_0} \end{aligned} \quad (2.41)$$

where ρ is the "electrostatic charge density," is the amount of charge per unit volume, and \mathbf{j} is the "electric current density," is the rate at which charge flows through a unit area per second. I hope (but I'm not sure) you've seen the complete classical theory of the electromagnetic field. You see what an elegantly simple form we can get with our new toolkit!

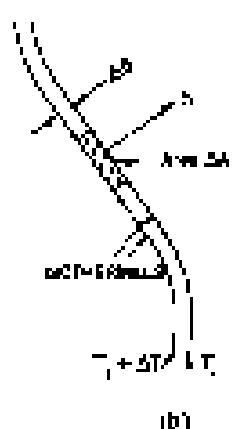
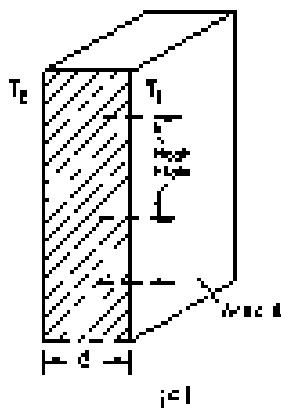


Fig. 2-7. (a) Heat flow through a slab. (b) An infinitesimal disk parallel to an interface surface in a large block.

2.4 The differential equation of heat flow

Let me give another example of a law of physics written in vector notation. This law is not a pristine one, not for many metals and a number of other substances that conduct heat it is quite accurate. You know that if you take a slab of material and heat one face to temperature T_0 and cool the other at a different temperature T_1 , the heat will flow through the material. In Fig. 2-7(a) the heat flow is proportional to the area A of the faces, and to the temperature $\Delta T = T_0 - T_1$ in inverse proportion to d , the distance between the faces. (For a given temperature difference, the thicker the slab the greater the heat flow.) Letting q be the thermal one per unit area per unit time through the slab, we write

$$q = k(T_0 - T_1) \frac{A}{d}. \quad (2.42)$$

The constant k of proportionality (k appa) is called the thermal conductivity.

What will happen to a more complicated case? Say in an odd-shaped block of material in which the temperature varies in peculiar ways? Suppose we look at a tiny piece of the block and imagine a slab like that of Fig. 2-7(a), on a microscopic scale. We orient the faces parallel to the mathematical surface, as in Fig. 2-7(b). Is Eq. (2.42) in correct for the small slab?

If the area of the small slab is ΔA , the heat flow per unit time is

$$\Delta Q = -k \nabla T \frac{\Delta A}{\Delta x}, \quad (2.47)$$

where Δx is the thickness of the slab. Now $k/\Delta x$ was defined earlier as the magnitude of k , whose direction is the heat flow. The heat flow will be from $T_{x+\Delta x}$ toward T_x , and so it will be perpendicular to the boundaries, as shown in Fig. 2-7(b). Above, $\partial T/\partial x$ is just the rate of change of T with position. And since the position change is perpendicular to the boundaries, $\partial T/\partial x$ is the maximum rate of change. I.e., therefore, out the magnitude $\|\nabla T\|$. Now since the direction of ∇T is opposite to that of k , we can write (2.47) as a vector \mathbf{q}_1 , where:

$$\mathbf{q}_1 = -k \nabla T. \quad (2.48)$$

(This minus sign is necessary because \mathbf{q}_1 does "downhill" in temperature.) Equation (2.48) is the differential equation of heat conduction in bulk materials. You can now type a proper computer program. Each variable is a vector of size just a number. It is the generalization to arbitrary cases of the second relation (2.42) for rectangular slabs. Later we should learn to write all sorts of elementary physics relations (Eq. 2.48) in the much more dignified vector notation. This notation is useful not only because it makes the equations more compact. It also makes more clearly the physical content of the equations without reference to any arbitrarily chosen coordinate system.

2-1 Second derivatives of vector fields

So far we have had only first derivatives. What are second derivatives? We could have several answers:

- (a) $\nabla \cdot (\nabla T)$
 - (b) $\nabla \times (\nabla T)$
 - (c) $\nabla(\nabla \cdot \mathbf{A})$
 - (d) $\nabla \cdot (\nabla \times \mathbf{A})$
 - (e) $\nabla \times (\nabla \times \mathbf{A})$
- (2.49)

You can check that there are *at least* the possible combinations.

Let's look first at the second one, (c). It has the same form as

$$\mathbf{A} \times (\mathbf{A} \mathbf{T}) = (\mathbf{A} \times \mathbf{A}) \mathbf{T} = 0,$$

since $\mathbf{A} \times \mathbf{A}$ is always zero. So we should have

$$\text{curl}(\text{grad } T) = \nabla \times (\nabla T) = 0. \quad (2.46)$$

We can see how this equation comes about if we go through once with the components:

$$\begin{aligned} [\nabla \times (\nabla T)]_x &= \tau_x(\nabla T)_x - \tau_y(\nabla T)_y \\ &= \frac{\partial}{\partial z} \left(\frac{\partial T}{\partial x} \right) + \frac{\partial}{\partial y} \left(\frac{\partial T}{\partial x} \right), \end{aligned} \quad (2.50)$$

which is zero (by Eq. 2.4). It goes the same for the other components. So $\nabla \times (\nabla T) = 0$, for any temperature distribution T that is an *scalar* function.

Now let us take another example. Let us see whether we can find another zero. The dot product of a vector with a cross product which contains that vector is zero.

$$A \cdot (A \times B) = 0. \quad (2.48)$$

Because $A \times B$ is perpendicular, or to A , and so has no component in the direction of A . The same conclusion applies to all of (2.45), so we have

$$\nabla \cdot (\nabla \times A) = \nabla \cdot (\text{curl } A) = 0. \quad (2.49)$$

Again, it is easy to show that it is zero by carrying through the operations with components.

Now we are going to state a famous mathematical theorem, but we will not prove. They are very interesting and useful theorems for physicists to know.

In a physical problem we frequently find that the curl of some quantity ϕ of the vector field A is zero. Now we know from (Eq. 2.46) that the curl of a gradient is zero, which is easy to remember because of the way the vectors work. It could certainly be, for example, that A is the gradient of some quantity because then its curl would necessarily be zero. Unfortunately it's not true that if the curl is zero, then A is always the gradient of something. There is some scalar field ψ (perhaps) such that A is equal to $\nabla \psi$. In other words, we have the

Theorem:

$$\text{If} \quad \nabla \times A = 0 \\ \text{then there is a} \quad \psi \\ \text{such that} \quad A = \nabla \psi. \quad (2.50)$$

There is a similar theorem. If the divergence of A is zero, we saw in Eq. 2.47 that the divergence of ∇ and of something is always zero. If you could choose a scalar field B for which $\text{div } B$ is zero, then you can conclude that B is the curl of some vector field C .

Theorem:

$$\text{If} \quad \nabla \cdot D = 0 \\ \text{then there is a} \quad C \\ \text{such that} \quad D = \nabla \times C. \quad (2.51)$$

By looking at the possible combinations of two ∇ operators, we have found that two of them always give zero. Now we look at the cases that are not zero. Take the one condition $\nabla \cdot (\nabla T)$, which was last in our list. It is not, in general, zero. We write out the components:

$$\nabla T = \nabla_x T + \nabla_y T + \nabla_z T.$$

Then

$$\begin{aligned} \nabla \cdot (\nabla T) &= (\nabla \cdot \nabla_x T) + (\nabla \cdot \nabla_y T) + (\nabla \cdot \nabla_z T) \\ &= \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}. \end{aligned} \quad (2.52)$$

which you'll in general want to be some number. It's a scalar field.

You see that we do not need to keep the parentheses, but can write, without any chance of confusion:

$$\nabla \cdot (\nabla T) = \nabla \cdot \nabla T = (\nabla \cdot \nabla)T = \nabla^2 T \quad (2.53)$$

We call ∇^2 a new operator. It is a scalar operator. Because it appears often in physics, it has been given a special name—the Laplacian.

$$\text{Laplacian: } \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \quad (2.54)$$

Since the Laplacian is a scalar operator, we may operate with it on a vector—by which you mean the same vectorial on each component in rectangular coordinates:

$$\nabla^2 \mathbf{A} = (\nabla^2 A_x, \nabla^2 A_y, \nabla^2 A_z).$$

Let's look at one more possibility: $\nabla \times (\nabla \times \mathbf{A})$, where was (6) in the list (2.45). Now the curl of the curl can be written differently if we use the vector identity (2.6):

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}(\mathbf{A} \cdot \mathbf{B}). \quad (2.55)$$

In order to use this formula, we should replace \mathbf{A} and \mathbf{B} by the operator ∇ and put $\mathbf{C} = \mathbf{A}$. If we do this, we get

$$\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \mathbf{A}(\nabla \cdot \nabla). \quad (2.56)$$

Wait a minute! Something is wrong. The first two terms are vectors all right. (The operator one is valid), but the last term doesn't come out to anything—it's still an operator. (The trouble is that we haven't been careful enough about keeping the order of our terms straight). If you look again at Eq. (2.55), however, you see that we could equally well have written it as

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) + \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}. \quad (2.56)$$

The second term looks better. It's easier to make the substitution in (2.56). We get

$$\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) + (\nabla \cdot \nabla)\mathbf{A}. \quad (2.57)$$

This form looks all right. It is, in fact, correct, as you can verify by computing the components. The last term is the Laplacian, so we can equally well write

$$\nabla \times (\nabla \times \mathbf{A}) + \nabla(\nabla \cdot \mathbf{A}) = \nabla^2 \mathbf{A}. \quad (2.58)$$

We have had something to say about all of the combinations in one line of double ∇ 's, except for (a), $\nabla(\nabla \cdot \mathbf{A})$. It is a possible vector field, but this is nothing special to say about it. It's just some vector field which may occasionally come up.

It will be convenient to have a table of such combinations.

- | | |
|-----|--|
| (a) | $\nabla \cdot (\nabla T) = \nabla^2 T$ — a scalar field |
| (b) | $\nabla \times (\nabla T) = 0$ |
| (c) | $\nabla(\nabla \cdot \mathbf{A}) = \mathbf{a}$ vector field |
| (d) | $\nabla \cdot (\nabla \times \mathbf{A}) = 0$ |
| (e) | $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$ |
| (f) | $(\nabla \cdot \nabla)\mathbf{A} = \nabla^2 \mathbf{A}$ — a vector field |
- (2.59)

You may notice that we haven't tried to invert a new vector operator ($\nabla \times \nabla$). Do you see why?

2-8 Pitfalls

We have been applying our knowledge of ordinary vector algebra to the algebra of the operator ∇ . We dare to be careful, though, because it is possible to go astray. There are two pitfalls which we shall mention, although they will not come up in this course. What would you say about the following expression, for example, the inverse of the two scalar functions ϕ and ψ (3.3):

$$(\nabla \phi) \times (\nabla \psi)^{-1}$$

You might want to say: it must be zero because it's just like

$$(A\phi) \times (A\psi),$$

which is now because the cross product of two square vectors $A \times A$ is always zero. But in our example the two operators ∇ are different. One is the derivative operator, the other is the gradient operator, or a different function, ϕ . So it might be represented there by the same symbol, ∇ , they must be considered as different operators. Clearly, the direction of $\nabla\phi$ depends on the function, ϕ , so it is not likely to be parallel to $\nabla\psi$.

$$(\nabla\psi) \times (\nabla\phi) \neq 0 \quad (\text{generally})$$

Fortunately, we won't have to deal with expressions like this we have said doesn't change the fact that $\nabla \times \nabla\phi = 0$ for any scalar field, because here both ∇ 's operate on the same function.

The last number ten rule (again we need not get into it in this context) is the following: The rules can we have outlined here are simple and also when we use rectangular coordinates. For example, if we have $\nabla^2\psi$ and we want the x -component it is

$$\nabla^2\psi_x = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \psi_x = \nabla^2\psi_x \quad (2.60)$$

The same expression would not work if we were to ask for the radial component of $\nabla^2\psi$. The radial component of $\nabla^2\psi$ is not equal to $\nabla^2\psi_r$. The reason is that when we are dealing with the algebra of vectors, the directions of the vectors are all quite definite. But when we are dealing with vector fields, their directions are different at different places. If we try to describe a vector field by say, polar coordinates, what we call the "radial" direction varies from point to point. So we run into a lot of trouble when we want to differentiate the components. For example, even for a constant vector field, the radial component changes from point to point.

It is usually safest and simplest just to stick to rectangular coordinates and avoid trouble, but there is one exception: **POLAR COORDINATES:** Since the Laplacian ∇^2 is a scalar, we can write it in any coordinate system we want to (for example in polar coordinates). It is since it is a differential operator, we should use it only on vectors whose components are in a fixed direction, that means rectangular coordinates. So we don't express ∇^2 of our vector fields in terms of their x , y , and z -components when we write our vector differential equations, but in components.

Vector Integrals; Curves

3-1 Vector Integrals; the Line Integral of $\nabla\phi$

We found in Chapter 2 that there were various ways of taking derivatives of fields. Some gave vector fields, some gave scalar fields. Although we developed many different formulas, everything in Chapter 2 could be summarized in one rule: the operations $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ are the three components of a vector operator ∇ . We would like to get some understanding of the "significance" of the derivatives of fields. We will then have a better feeling for what a vector field represents.

We have already discussed the meaning of the gradient, equation 17 on 2 readers. Now we turn to the meanings of the divergence and curl operators. The interpretation of these quantities is best done in terms of certain vector integral and algorithms relating such integrals. These equations, known uniformly, are obtained from vector algebra by some very common rules, so you will just have to learn them as something new. Of these integral formulas, one is physically trivial, but the other two are not. We will derive them and explain their implications. The equations we shall study are really mathematical theorems; they will be useful not only for interpreting the meaning and the control of the divergence and the curl, but also in working out general physical theories. These mathematical theorems are for the theory of fields, what the theorem of the conservation of energy is to the mechanics of particles. Let me, therefore, stress that these are important for a deeper understanding of physics. You will find, though, that they are not very useful for solving problems—except in the simplest cases. It is important, however, that in the beginning of our subject there will be some simple problems which can be solved with the three integral formulas we are going to meet. We will see, however, in the problems you hunger, that you will longer use these simple methods.

We take up first an integral formula involving the gradient. The relation contains a very simple idea: Since the gradient represents the rate of change of a field quantity, if we integrate that rate of change, we should get the total change. Suppose we have the scalar field $\phi(x, y, z)$. At any two points (1) and (2), the function ϕ will have the values $\phi(1)$ and $\phi(2)$, respectively. [This uses a convenient notation, in which (2) represents the point (x_2, y_2, z_2) and $\phi(2)$ means the same thing as $\phi(x_2, y_2, z_2)$.] If γ is a curve joining (1) and (2), as in Fig. 3-1, the following relation is true:

Theorem 1.

$$\phi(2) - \phi(1) = \int_{(1)}^{(2)} (\nabla\phi) \cdot d\mathbf{r} \quad (3-1)$$

The integral is a line integral, from (1) to (2) along the curve γ , of the dot product of $\nabla\phi$ —a vector—with $d\mathbf{r}$ —another vector which is an infinitesimal line element of the curve γ (selected away from (1) and toward (2)).

First, we should review what we mean by a line integral. Consider a scalar function $f(x, y, z)$, and the curve γ joining two points (1) and (2). We mark off the curve at a number of points and join them together by straight-line segments, as shown in Fig. 3-2. Each segment has the length Δs_i , where i is an index that runs 1, 2, 3, ... By the line integral

$$\int_{(1)}^{(2)} f d\mathbf{r}$$

3-2 Vector Integrals; the Line Integral of $\nabla\phi$

3-2 The Line of a vector field

3-3 The flux from a cube; Gauss' theorem

3-4 Heat conduction; the diffusion equation

3-5 The circulation of a vector field

3-6 The circulation around a square; Stokes' theorem

3-7 Curl-free and irrotational free fields

3-8 Summary

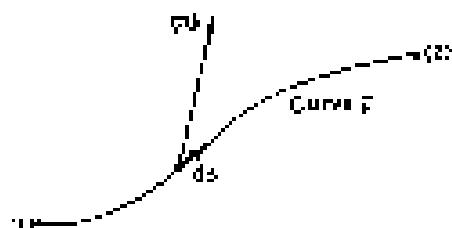


Fig. 3-1. The terms used in Eq. (3-1). The vector $n\delta\mathbf{r}$ is evaluated at the line element $\delta\mathbf{r}$.

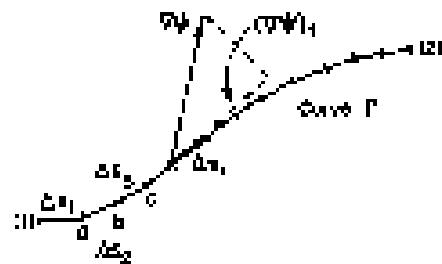


Fig. 3-2. The line integral is the limit of a sum.

we mean the limit of the sum

$$\sum_i f_i \Delta x_i$$

where f_i is the value of the function at the i th segment. The limiting value is what the sum approaches as we take many and many segments (in a sensible way, so that the largest $\Delta x_i \rightarrow 0$).

The integral in our theorem, Eq. (3.7), means the same thing, although it looks a little different. Instead of f_i , we have another scalar—the component of $\nabla \psi$ in the direction of Δx . If we write (3.7) for the trapezoidal approach, it is clear that:

$$(\nabla \psi) \cdot \Delta x = (\nabla \psi) \cdot ds. \quad (3.2)$$

The integral in Eq. (3.1) means the sum of such terms.

Now let's see why Eq. (3.1) is true. In Chapter 1, we showed that the component of $\nabla \psi$ along a small displacement Δx was the rate of change of ψ in the direction of Δx . Consider the line segment Δx_1 from (1) to point a in Fig. 3-2 according to our definition.

$$\Delta x_1 = \psi(a) - \psi(1) = (\nabla \psi)_1 \cdot \Delta x_1. \quad (3.3)$$

At x_1 , we have

$$\psi(0) = \psi(1) + (\nabla \psi)_1 \cdot \Delta x_1, \quad (3.4)$$

where, of course, $(\nabla \psi)_1$ means the gradient evaluated at the segment Δx_1 , and $(\nabla \psi)_2$ the gradient evaluated at Δx_2 . If we add Eqs. (3.3) and (3.4) we get

$$\psi(0) - \psi(1) = (\nabla \psi)_1 \cdot \Delta x_1 + (\nabla \psi)_2 \cdot \Delta x_2. \quad (3.5)$$

You can see that if we keep adding such terms, we get the result:

$$\psi(0) - \psi(1) = \sum_i (\nabla \psi)_i \cdot \Delta x_i. \quad (3.6)$$

The left-hand side doesn't depend on how we chose our intervals; if (1) and (2) are large, always the same—so we might take the limit of the right-hand side. We have therefore proved Eq. (3.1).

You can see from the proof that just as the equality doesn't depend on how the points x_1, x_2, x_3, \dots are chosen, similarly it doesn't depend on what we choose for the curve C to join (1) and (2). Our theorem is correct for any curve from (1) to (2).

One remark on notation: You will see later that it is common if we write for convenience

$$(\nabla \psi) \cdot dx = \nabla \psi \cdot ds. \quad (3.7)$$

With this notation, our theorem is

$$\text{Therefore, } \int_C \psi(0) - \psi(1) = \int_{S_1}^{S_2} \nabla \psi \cdot ds. \quad (3.8)$$

3-3. The Flux of a Vector Field

Before we consider our divergence theorem—a theorem about flux over surfaces—we would like to study a related idea which has an easily understood physical significance in the case of heat flow. We have defined the vector k , which represents the rate of flow through a unit area in a unit time. Suppose that inside a closed shell of volume V we have some closed surface S which encloses the volume V (Fig. 3-3). We would like to find out how much heat is leaving out of this volume. Written, of course, this is by calculating the total heat flow out of the surface S .

We write dA for the area of an element of the surface. The symbol stands for a two-dimensional differential. If, for instance, the open happened to be in the xy -plane we would have

$$dA = dx dy.$$



Fig. 3-3. The closed surface S contains the volume V . The unit vector n is the outward normal to the surface element dA and k is the heat-flow vector at the surface element.

Later we shall have integrals over volume and for these it is convenient to consider a differential volume that is a little cube. So when we write dV we mean

$$dV = dxdydz$$

Some people like to write $d^3\mathbf{r}$ instead of $dxdydz$, themselves, but it is kind of a second-order quantity. They would also write dV instead of dV . We will use the simple notation, and assume that you can remember that dV has two dimensions and dV has three.

The heat flow \mathbf{q} through the surface element $d\mathbf{S}$ is the heat times the component of \mathbf{q} perpendicular to $d\mathbf{S}$. We have already defined \mathbf{n} as a vector pointing outward at right angles to the surface (Fig. 3-2). The component of \mathbf{q} that we want is

$$\mathbf{q}_n = \mathbf{q} \cdot \mathbf{n} \quad (3.9)$$

heat flow out through $d\mathbf{S}$ is then

$$d\Phi = \mathbf{q}_n d\mathbf{S} \quad (3.10)$$

To get the total heat flow through any surface we sum the contributions from all the elements of the surface. In other words, we integrate (3.10) over the whole surface:

$$\text{Total heat flow outward through } S = \int_S \mathbf{q}_n d\mathbf{S} \quad (3.11)$$

We are also going to call this surface integral "the flux of \mathbf{q} through the surface." Originally the word flux meant flow, so that the surface integral just means the flow of \mathbf{q} through the surface. We may think of the "average density" of \mathbf{q} as \mathbf{q} and the surface integral of \mathbf{q} is the total heat current directed out of the surface, i.e., is, the "current density per unit time (rate per second).

We would like to generalize this idea to the case where the surface does not intersect the flow of anything; for instance, it might be the electric field. We can certainly still integrate the normal component of the electric field over an area S without. Although it is not the flow of anything, we still call it the "flux." We say

$$\text{Flux of } \mathbf{E} \text{ through the surface } S = \int_S \mathbf{E} \cdot \mathbf{n} d\mathbf{S} \quad (3.12)$$

We generalize the word "flux" to mean the "surface integral of the normal component of a vector." We will also use the same definition even when the surface contains a nonclosed curve, as in Fig. 3-3.

Returning to our special case of heat flow, let us take a situation in which $\mathbf{q}(x)$ is constant. For example, imagine some material in which there is no heat loss or further heat energy is generated or absorbed. Then if there is a net heat flow out of a given surface, the heat content of the volume inside must decrease. So, in circumstances in which heat would be conserved, we say that

$$\int_S \mathbf{q}_n d\mathbf{S} = -\frac{dQ}{dt}, \quad (3.13)$$

where Q is the heat inside the surface. The heat flux out of S is equal to minus the rate of change of Q with respect to time of the total heat Q inside S . This interpretation is possible because we are speaking of heat flow and also because we supposed that the heat was conserved. We could just, of course, speak of the total heat inside the volume if heat were being generated there.

Now we shall prove an interesting fact about the flux of any vector. You may think of the heat flow result, if you like, but what we say will be true for any vector field \mathbf{q} . Imagine that we have a closed surface S that encloses the volume V . We now separate the volume into two parts by some kind of a "cut," as in Fig. 3-4. Now we have two closed surfaces and volumes. The volume V_1 is enclosed in the surface S_1 , which is made up of part of the original surface S , and of the surface of the cut, S_{21} . The volume V_2 is enclosed by S_2 , which is made up of the rest of the original surface S and closed off by the cut S_{12} . Now consider the

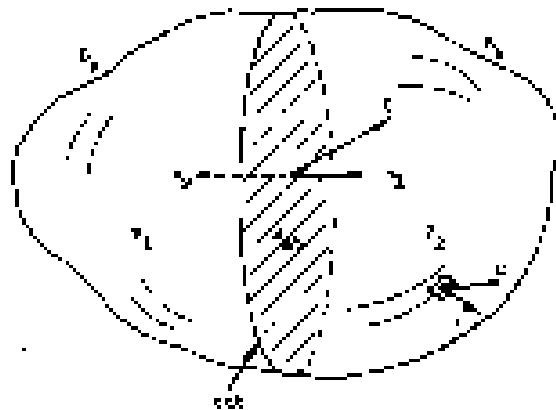


Fig. 3-4. A volume V contained inside the surface S is divided into two pieces by a "cut" at the surface S_{12} . We now have the volume V_1 enclosed in the surface $S_1 = S_1 - S_{12}$ and its volume V_2 enclosed in the surface $S_2 = S_2 - S_{12}$.

Following question: Suppose we calculate the flux $\int_C \mathbf{C} \cdot d\mathbf{s}$ through surface S_1 , and add on it the flux $\int_C \mathbf{C} \cdot d\mathbf{s}$ through surface S_2 . Does the sum equal the flux through the whole surface S ? Can we ignore S_{12} ? The answer is yes. The flux through the part of the surface S_1 common to both S_1 and S_2 just exactly cancels out. For the flux of the vector field \mathbf{C} in V_1 , we can write

$$\text{Flux through } S_1 = \int_{S_1} \mathbf{C} \cdot \mathbf{n}_1 d\mathbf{s} + \int_{S_{12}} \mathbf{C} \cdot \mathbf{n}_1 d\mathbf{s}, \quad (3.14)$$

and for the flux out of V_2 ,

$$\text{Flux through } S_2 = \int_{S_2} \mathbf{C} \cdot \mathbf{n}_2 d\mathbf{s} + \int_{S_{12}} \mathbf{C} \cdot \mathbf{n}_2 d\mathbf{s}. \quad (3.15)$$

Note that in the second integral we have written \mathbf{n}_2 for the outward normal for S_2 , when it belongs to S_1 , and \mathbf{n}_1 when it belongs to S_2 , as shown in Fig. 3-4. Clearly, $\mathbf{n}_1 = -\mathbf{n}_2$ so that

$$\int_{S_{12}} \mathbf{C} \cdot \mathbf{n}_1 d\mathbf{s} = - \int_{S_{12}} \mathbf{C} \cdot \mathbf{n}_2 d\mathbf{s}. \quad (3.16)$$

If we now add Eqs. (3.14) and (3.15), we see that the sum of the fluxes through S_1 and S_2 is in fact the sum of two integrals which, taken together, give the flux through the original surface $S = S_1 + S_2$.

We see that the flux through the complete outer surface of can be considered as the sum of the fluxes from the two parts into which the volume was broken. We can similarly subdivide space—say by cutting V in n two pieces. You see that the same arguments apply. So the only way of dividing the original volume, if indeed it generally true that the flux through the outer surface, which is the *original* integral, is equal to a sum of the fluxes out of all the tiny interior pieces.

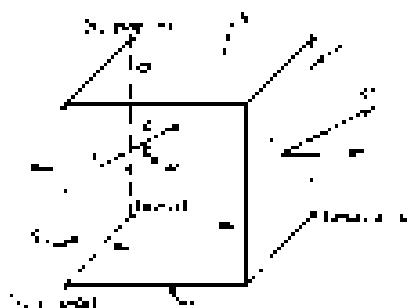


Fig. 3-5. Computation of the flux of \mathbf{C} out of a small cube.

3-5 The flux from a cubic Gauss' theorem

We now face the special case of a small cube and find an interesting formula for the flux out of it. Consider a cube whose edges are all up with the axes as in Fig. 3-5. Let us suppose that the coordinates of the corner points lie in integers x_1, x_2, x_3 . Let a_1 be the length of the cube in the x_1 direction, a_2 be the length in the x_2 direction, and a_3 be the length in the x_3 direction. We wish to find the flux of a vector field \mathbf{C} through the surface of the cube. We do this by making a sum of the fluxes through each of the six faces. First, consider the face marked n_1 in the figure. The flux outward on this face is the negative of the x_1 -component of \mathbf{C} , integrated over the area of the face. This flux is

$$- \int_C \mathbf{C} \cdot \mathbf{n}_1 d\mathbf{s}$$

Since we are considering a small cube, we can approximate this integral by the

* The following development applies equally well to any relatively simple integral.

value of C , at the center of the face—which we call the *face value* of C —multiplied by the area of the face, $\Delta x \Delta y$:

$$\text{Flux out of } 1 = -C_1(1) \Delta y \Delta x$$

Similarly, for the flux out of face 2, we write

$$\text{Flux out of } 2 = C_2(2) \Delta y \Delta x.$$

Now $C_1(1)$ and $C_2(2)$ are, in general, slightly different. If they are small enough, we can write

$$C_2(2) = C_1(1) + \frac{\partial C}{\partial x} \Delta x.$$

There are, of course, more terms, but they will involve $(\Delta x)^2$ and higher powers, and so will be negligible if we consider only the limit of small Δx . So the flux through face 2 is

$$\text{Flux out of } 2 = \left[C_1(1) + \frac{\partial C}{\partial x} \Delta x \right] \Delta y \Delta x.$$

Summing up fluxes for faces 1 and 2, we get

$$\text{Flux out of 1 and 2} = \frac{\partial C}{\partial x} \Delta x \Delta y \Delta z.$$

The derivative should really be the value at the center of face 1; that is, at $(x, y = (x\Delta/2), z = (z\Delta/2))$. But at the limit of an infinitesimal cube, we make a negligible error if we evaluate it at the center (x, y, z) .

Applying the same reasoning to each of the other pairs of faces, we have

$$\text{Flux out of 3 and 4} = \frac{\partial C}{\partial y} \Delta x \Delta y \Delta z$$

and

$$\text{Flux out of 5 and 6} = \frac{\partial C}{\partial z} \Delta x \Delta y \Delta z.$$

The total flux through all the faces is the sum of these terms: Δx and Δy ,

$$\int_{\text{faces}} \mathbf{C} \cdot \mathbf{n} d\mathbf{a} = \left(\frac{\partial C}{\partial x} + \frac{\partial C}{\partial y} + \frac{\partial C}{\partial z} \right) \Delta x \Delta y \Delta z,$$

and the sum of the elements is just ΔV . After $\Delta x = \Delta y = \Delta z$, the volume of the cube. So we can say that, in an infinitesimal cube

$$\int_{\text{faces}} \mathbf{C} \cdot \mathbf{n} d\mathbf{a} = (\nabla \cdot \mathbf{C}) \Delta V. \quad (5.17)$$

We have shown that the outward flux from the surface of an infinitesimal cube is equal to the divergence of the vector multiplied by the volume of the cube. We now see the “meaning” of the divergence of a vector. The divergence of a vector at the point P is the flux—an outgoing “flow” of C —per unit volume, in the neighborhood of P .

We have connected the divergence of C to the flux of C out of an infinitesimal volume. For any finite volume we can do the same, as proven elsewhere: the total flux from a volume is the sum of the fluxes out of each part. We can then integrate the divergence over the entire volume. This gives us the theorem that the integral of the normal component of any vector over any closed surface can also be written as the integral of the divergence of the vector over the volume enclosed by the surface. This is often called the Gauss law.

Final word: Theorem,

$$\int_{\text{S}} \mathbf{C} \cdot \mathbf{n} d\mathbf{a} = \int_V \nabla \cdot \mathbf{C} dV, \quad (5.18)$$

where S is any closed surface and V is the volume inside it.

3-4 Heat conduction; the diffusion equation

Let's consider an example of the use of this theorem, just to get familiar with it. Suppose we take again the case of heat flow in, say, a metal. Suppose we have a simple situation in which all the heat has been previously put in and the body is just cooling off. There are no sources of heat, so that heat is conserved. Then how much heat is there inside some chosen volume at any time? It must be decreasing by just the amount that flows out of the surface of the volume. If our volume is a cube, we would write, following Eq. (3.17),

$$\text{Rate of loss} = \int_{\text{Surface}} \dot{q} \cdot \hat{n} d\sigma = -k \cdot A \Delta T. \quad (3.19)$$

This rate must equal the rate of loss of the heat inside the cube. If \dot{q} is heat per unit volume, the heat in the cube is $\rho A V$, and the rate of loss is

$$\frac{d}{dt} (\rho A V) = - \frac{\partial}{\partial x} \Delta T. \quad (3.20)$$

Comparing (3.19) and (3.20), we see that

$$-\frac{\partial}{\partial x} \Delta T = \dot{q}. \quad (3.21)$$

Take careful note of the ΔT in this equation. This form appears often in physics. It expresses a conservation law—here, the conservation of heat. We have imagined the same physics, due in another way in Eq. (3.13). There we have the differential form of a conservation equation, while Eq. (3.13) is the integral form.

We have obtained Eq. (3.21) by applying Eq. (3.11) to an infinitesimal cube. We can also go the other way. For a big volume V bounded by S , Gauss' law says that

$$\int_S \dot{q} \cdot \hat{n} d\sigma = \int_V \nabla \cdot \dot{q} dV. \quad (3.22)$$

Using (3.21), the integral on the right-hand side is found to be just $-k Q/dV$, and we have Eq. (3.13).

Now let's consider a different case. Imagine that we have a block of material and that inside it there is a very tiny hole in which some chemical reaction is taking place and generating heat. Or we could imagine this. There are some wires running through it, say copper, that is being heated by an electric current. We shall suppose that the heat is generated practically at a point, and let H represent the energy liberated per second at that point. We shall suppose that in the rest of the volume heat is conserved, and that the heat generation has been going on for a long time—so that now the temperature is no longer changing anywhere. The problem is: What does the heat mean? How like a current power is the heat? How much heat flow is there at each point?

We know that if we integrate the normal component of \dot{q} over a closed surface that encloses the source, we will always get H . All the heat that is being generated at the point comes out, goes out through the \dot{q} flow, since we have supposed that the flow is steady. We have the difficult problem of finding a vector field which, when integrated over any surface, always gives H . We can, however, find the field rather easily by taking a somewhat special surface. We take a sphere of radius R , centered at the source, and assume that the heat flow is radial (Fig. 3-6). Our intuition tells us that \dot{q} should be radial if the block of material is large, and we don't get too close to the source, and it should also have the same magnitude at x points on the sphere. You see that we are doing a certain amount of guess-work, usually called "physical intuition". In the mathematics, in order to find the answer,

when \dot{q} is to do just spherically symmetric, the integral of the normal component of \dot{q} over the area is very simple, because the normal component is just \dot{q} .



Fig. 3-6. In the region near a point source of heat, the heat flow is radially outward.

the magnitude of δ and b constant. The area over which we integrate is $4\pi R^2$. We have then that

$$\int_{\text{source}} \delta \cdot \mathbf{a} d\mathbf{x} = b \cdot 4\pi R^2 \quad (3.24)$$

(where b is the magnitude of δ). This integral should equal W , the rate at which heat is produced at the source. We get

$$W = \frac{b}{4\pi R} \cdot$$

or

$$b = \frac{W}{4\pi R} \cdot \tau_n \quad (3.25)$$

With b , as usual, a , represents a unit vector in the radial direction. Our result says that the proportionality W and a varies inversely as the square of the distance from the source.

The result we have just obtained applies in the heat flow in the vicinity of a point source of heat. Let's now try to find the equations that hold in the most general kind of heat flow. Here, keeping only the condition that heat is conserved. We will be dealing only with what happens at places outside of any sources or sinks (sites of heat).

The differential equation for the conduction of heat was derived in Chapter 2, according to Eq. (2.44)

$$\delta = -\kappa \nabla T \quad (3.26)$$

(Remember that this relationship is an approximation, but fairly good for some materials like metals). It is applicable, of course, only in regions of the material where there is no generation or absorption of heat. We derived above another relation, Eq. (3.21), that holds when heat is conserved. If we combine that equation with (3.26), we get

$$\begin{aligned} -\frac{\partial \delta}{\partial r} &= \mathbf{v} \cdot \delta = -\mathbf{v} \cdot (\kappa \nabla T), \\ \text{or} \\ \frac{\partial \delta}{\partial r} &= \kappa \mathbf{v} \cdot \nabla T = \kappa \nabla^2 T. \end{aligned} \quad (3.27)$$

If κ is a constant. Your common sense tells you that δ is proportional to heat in a unit volume and $\nabla^2 T = \nabla^2 \delta$ is the Laplacian operator.

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

If we now make your common assumption we can obtain a very interesting equation. We assume that the temperature of the material is proportional to the heat content per unit volume—that is, that the material has a definite specific heat. When this assumption is valid (as it often is), we can write

$$\delta_T = \kappa \Delta T$$

or

$$\frac{\partial \delta}{\partial r} = \kappa \frac{\partial^2 T}{\partial r^2} \quad (3.28)$$

The rate of change of heat is proportional to the rate of change of temperature. The constant of proportionality κ , is, here, the specific heat per unit volume of the material. Using Eq. (3.27) with (3.28), we get

$$\frac{\partial T}{\partial r} = \frac{\kappa}{c_v} \nabla^2 T. \quad (3.29)$$

We find that the coefficient of change of T —at every point—is proportional to the Laplacian of T , which is the second derivative of its spatial dependence. We have a differential equation—in x , y , z , and r —for the temperature T .

The differential equation (3.26) is called the heat diffusion equation. It is often Problem 43.

$$\frac{\partial T}{\partial t} = \alpha \nabla^2 T, \quad (3.29)$$

where α is called the diffusion constant, and is here equal to c/ρ .

The diffusion equation appears in many physical problems: in the diffusion of gases, in the diffusion of solutions, and so others. We have already discussed the physics of some of these phenomena in Chapter 43 of Vol. I. Now you have the complete equation that describes diffusion in the most general possible situation. At some later time we will take up ways of solving this diffusing equation to find how the temperature varies in particular cases. We can break now no longer off this business about water fiddle.

3-5. The circulation of a vector field

We wish now to look at the circulation in some other way we looked at the divergence. We obtained Green's theorem by considering the integral over a surface; although it was not obvious at the beginning that we were going to be dealing with the divergence. We did not know that we were supposed to integrate over a surface in order to get the divergence! It was not at all clear that this would be the result. And so with an apparently logical lack of justification, we shall calculate something else about a vector and show that it is related to it. This time we calculate what is called the circulation of a vector field. If C is any vector field, we take its component along a curve C and take the integral of this component all the way around a complete loop. The integral is called the circulation of the vector field around the loop. We have already established a line integral of C earlier in this chapter. Now we do the same kind of thing for any vector field C .

Let C be any closed loop in space—imagine, of course. An example is given in Fig. 3-7. The line integral of the tangential component of C around the loop is written as

$$\oint_C C \cdot d\mathbf{r} = \oint_C C \cdot ds. \quad (3.30)$$

You should note that the integral is taken all the way around, not from one point to another as we did before. The little circle on the integral sign is to remind us that the integral is to be taken all the way around. This integral is called the circulation of the vector field around the curve C . The name came originally from experiencing the circulation of a liquid. But the term—“circulation”—we have extended to apply to any field even when there is no material “circulating.”

Using the same kind of pure reason with the flux, we can show that the circulation around a loop is the sum of the circulations around two smaller loops. Suppose we break up our curve of Fig. 3-7 into two loops by joining two points (1) and (2) on the original curve by a new line that cuts across, as shown in Fig. 3-8. There are now two loops, T_1 and T_2 . T_1 is made up of C_1 , which is part of the original curve to the left of (1) and (2), plus C_2 , the “short cut.” C_2 is made up of the rest of the original curve plus the short cut.

The circulation around T_1 is the sum of an integral along C_1 and along C_2 . Similarly, the circulation around T_2 is the sum of two parts, one along C_3 and the other along C_4 . The integral along C_2 will cancel, for the curve C_2 , the opposite sign from what it has for C_1 , because the direction of travel is opposite—we must take both our line integrals with the same “sense” of travel.

Following the same kind of argument we used before, you can see that the sum of the two circulations or give just the line integral around the original curve T . The parts due to C_2 cancel. The circulation around the end part plus the circulation around the second part, or the single line around the inner line. We can continue the process of cutting the original loop into any number of smaller loops. When we add the circulations of the smaller loops, there is always a cancellation of the parts that are adjacent portions, and so however is equivalent to the circulation around the original single loop.

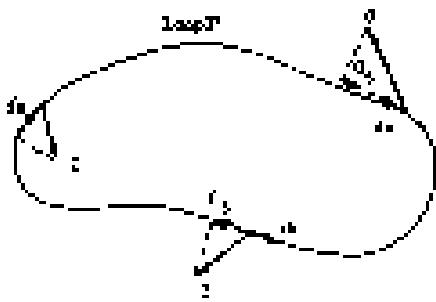


Fig. 3-7. The circulation of C around the curve C is the line integral of C , the tangential component of C .

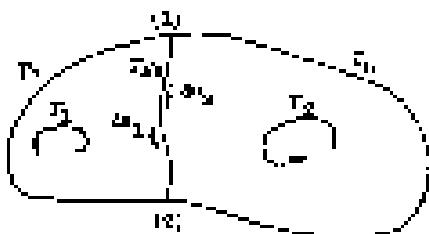


Fig. 3-8. The circulation around the whole loop is the sum of the circulations around the two loops $T_1 = T_a + T_b$ and $T_2 = T_c + T_d$.

Now let us suppose that the original loop is the boundary of some surface. Then, of course, an infinite number of surfaces which have the original loops as the boundary. One result will not, however, depend on which surface we choose. First, we break our original loop into a number of small loops. All lie on the surface we have chosen, as in Fig. 3-9. No matter what the shape of the surface, if we choose our small loops *small enough*, we can assume that each of the small loops will enclose an area which is essentially flat. Also, we can choose our small loops so that each is very nearly a square. Now we can calculate the circulation around the big loop C by finding the circulations around all of the little squares and then adding them up.

3-6 The circulation around a square; Stokes' theorem

How do we find the circulation for each little square? One question is: how is the square oriented to space? We could easily make the calculations if it had a special orientation. For example, if it were in one of the coordinate planes. Since we have not assumed anything about yet about the orientation of the coordinate axes, we can just as well choose the axes so that the four little squares we are concentrating on at the moment lie in the $x-y$ plane, as in Fig. 3-10. If our result is expressed in that orientation, we can say that it will be the same no matter what the particular orientation of the plane.

We want now to find the circulation of the field \mathbf{C} around our little square. It will be easy to do by integrating if we pass the square and through that the vector \mathbf{C} doesn't change much along any one side of the square. (The assumption is better the smaller the squares we are really taking, almost infinitesimal squares.) Starting at the point (x_1, y_1) —the lower-left corner of the figure—we go around it the direction indicated by the arrows. Along the first side—marked (1)—the tangential component is $C_x(1)$ and the distance is Δx . The first part of the integral is $C_x(1) \Delta x$. Along the second leg we get $-C_y(2) \Delta x$. Along the third, we get $-C_x(3) \Delta x$, and along the fourth, $-C_y(4) \Delta x$. The minus signs are necessary because we want the tangential component in the direction of travel. The whole line integral is then

$$\oint \mathbf{C} \cdot d\mathbf{s} = -C_x(1) \Delta x - C_y(2) \Delta y - C_x(3) \Delta x - C_y(4) \Delta y. \quad (3.11)$$

Now let's look at the first and third pieces. Together they are

$$[C_x(1) + C_x(3)] \Delta x. \quad (3.12)$$

You might think that in our approximation the difference is zero. That is true to the first approximation. We can be more accurate, however, and take into account the *rate* of change $\partial C_x / \partial x$. If we do, we may write

$$C_x(3) = C_x(1) + \frac{\partial C_x}{\partial x} \Delta x. \quad (3.13)$$

If we included the next approximation, it would involve terms in $(\Delta x)^2$, but since we still ultimately think of the limit as $\Delta x \rightarrow 0$, such terms can be neglected. Putting (3.13) together with (3.12), we find that

$$[C_x(1) + C_x(3)] \Delta x = -\frac{\partial C_x}{\partial x} \Delta x \Delta y. \quad (3.14)$$

The derivative due to our approximation is evaluated at (x, y) .

Similarly, for the other two terms in the circulation, we may write

$$C_y(2) \Delta y = C_y(4) \Delta y + \frac{\partial C_y}{\partial y} \Delta x \Delta y. \quad (3.15)$$

The circulation around our square is then

$$\left(\frac{\partial C_x}{\partial x} + \frac{\partial C_y}{\partial y} \right) \Delta x \Delta y. \quad (3.16)$$

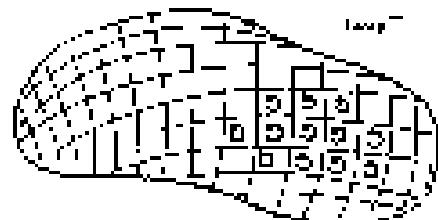


Fig. 3-9. Some surface bounded by the loop C is shown. The surface is divided into a number of small squares, each approximately a square. The circulation around C is the sum of the circulations around the little loops.

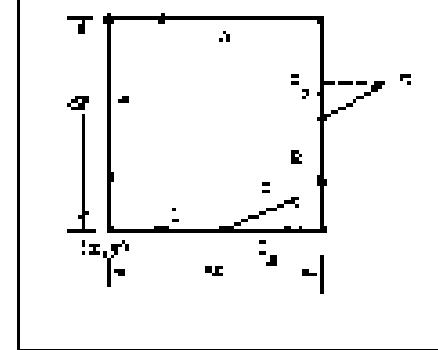


Fig. 3-10. Considering the circulation of \mathbf{C} around a unit square.

which is interesting because the two terms in the parentheses are just the *z*-component of the curl. Also, we may note that $\Delta \cdot \Delta p$ is the area of our square. So we can write our circulation (3-36) as

$$(\nabla \times C)_z \Delta p.$$

But the *z*-component really means the component normal to the surface element. We can, therefore, write the circulation around a differential square as a small vector sum:

$$\oint C \cdot d\ell = (\nabla \times C)_z \Delta p = (\nabla \times C) \cdot n \Delta p. \quad (3-37)$$

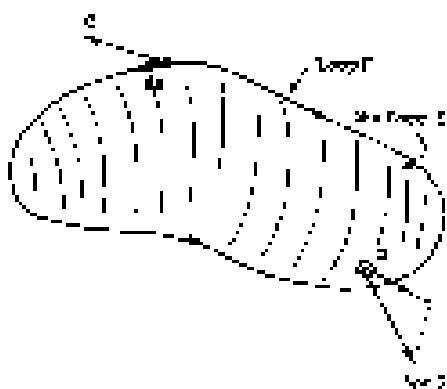


Fig. 3-11. The circulation of C around C is the surface integral of the *z*-component of $\nabla \times C$.

Our result for the circulation of any vector C around an infinitesimal square is the component of the curl of C normal to the surface, times the area of the square.

The circulation around any loop C can now be easily related to the curl of the vector field. We will do the loop with any convenient surface S , as in Fig. 3-11, and add the contributions from a set of "infinitesimal" squares in this surface. The sum can be written as an integral. Our result is a very useful theorem called Stokes' theorem (after M.L. Stokes).

Stokes' Theorem:

$$\oint_C C \cdot d\ell = \int_S (\nabla \times C) \cdot n d\ell, \quad (3-38)$$

where S is any surface bounded by C .

We must now speak about a convention of signs. In Fig. 3-10 the *z* axis would point toward you in a "fisted"→→→"right-handed"→→→system of axes. When we took our line integral with a "positive" sense of rotation, we found that the circulation was equal to the *z*-component of $\nabla \times C$. If we had gone around counter-clockwise, we would have gotten the opposite sign. Now here shall we know, in general, what direction to choose for the positive direction of the "normal" component of $\nabla \times C$? The "positive" normal must always be related to the sense of rotation, as in Fig. 3-11. It is indicated for the general case in Fig. 3-11.

One way of remembering the relationship is by the "right-hand rule". If you make the fingers of your right hand go around the curve C , with the fingertips pointing in the direction of the positive sense of $d\ell$, then your thumb points in the direction of the positive normal to the surface S .

3-7 Curl-free and divergence-free fields

We would like, now, to consider some consequences of our new theorem. Take first the case of a vector whose curl is everywhere zero. Then Stokes' theorem says that the circulation around any loop is zero. Now if we choose two paths (1) and (2) on a closed curve (Fig. 3-12), it follows that the surface integral of the tangential component from (1) to (2) is independent of which of the two possible paths is taken. We can, in fact, deform the integral from (1) to (2) over disjoint paths on the location of these poles - that is to say, it is a *conservative* of position only. This is analogous to what we learned in Chapter 14 of Vol. I, where we proved that if the integral along a closed loop of some quantity is always zero, then the integral can be represented as the difference of a function of the position of the two ends. This last statement we call the *law of a potential*. We just said, further, that this vector field was the gradient of this potential function (see Chap. 14, Sec. 17 of Vol. I).

It follows that any vector field whose curl is zero is equal to the gradient of some scalar function. That is, if $\nabla \times C = 0$, everywhere there is a vector field for which $C = \nabla \phi$ —a useful idea. We can, if we wish, describe this special kind of vector field by means of a scalar field.

Let's show something else. Suppose we have two paths (1) and (2) (Fig. 3-13) we take the gradient. The line integral of this vector along any curve has to be zero. Its line integral from point (1) to point (2) is $(\phi(2) - \phi(1))$. If (1) and (2) → 0

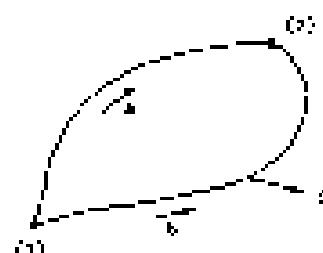


Fig. 3-12. If $\nabla \times C$ is zero, the circulation around the closed curve C is zero. The line integral of $C \cdot d\ell$ from (1) to (2) along C may be the same as the line integral along C .

and the second part, our Theorem 1, Eq. (1.6), tells us that the line integral is zero:

$$\oint_{\text{loop}} \mathbf{v} \cdot d\mathbf{l} = 0.$$

Using Stokes' Theorem, we can conclude the

$$\int_S \nabla \times (\mathbf{v}\mathbf{n}) \, dS = 0$$

over any surface. But if the integral is zero over any surface, the integral must be zero. So:

$$\nabla \times (\mathbf{v}\mathbf{n}) = 0, \text{ always.}$$

We proved the same result in Section 2.7 by vector algebra.

Let's look now at a special case in which we fill in a small loop T with a large surface S , as indicated in Fig. 3-11. We would like, in fact, to see what happens when the loop shrinks down to a point, so that the surface S ends up disappearing from the surface boundary of S . Now, if the vector \mathbf{C} is everywhere finite, the line integral around T must go to zero as we shrink the loop— \mathbf{v} is tangential to roughly perpendicular to the circumference of T , which goes to zero. According to Stokes' theorem, the surface integral of $(\nabla \times \mathbf{C})$ must also vanish. Furthermore, as we did the surface we add in contributes that cancel out what was there before. So we have a new theorem:

$$\int_S (\nabla \times \mathbf{C}) \cdot d\mathbf{S} = 0. \quad (3.39)$$



Fig. 3-11. Going to the limit of a closed surface, we find that the surface integral of $(\nabla \times \mathbf{C})$ must vanish.

Now this is interesting, because we already have a theorem about the surface integral of a vector field \mathbf{C} : such a surface integral is equal to the volume integral of the divergence of the vector, according to Gauss' theorem (Fig. 3-13). Gauss' theorem applied to $\nabla \times \mathbf{C}$ says

$$\int_{\text{volume}} (\nabla \times \mathbf{C}) \cdot d\mathbf{v} = \int_{\text{surface}} \mathbf{v} \cdot (\nabla \times \mathbf{C}) \, d\mathbf{S}. \quad (3.40)$$

So we conclude that the second integral must also be zero:

$$\int_{\text{surface}} \mathbf{v} \cdot (\nabla \times \mathbf{C}) \, d\mathbf{S} = 0, \quad (3.41)$$

and this is true for any vector field \mathbf{C} whatever. Since Eq. (3.41) is true for any volume, it must be true that at every point in space the divergence is zero. We have

$$\nabla \cdot (\mathbf{v} \times \mathbf{C}) = 0, \text{ always.}$$

But this is the same result we got from vector algebra in Section 2-7. Now we begin to see how everything fits together.

3-8 Summary

Let us summarize what we have learned about the vector calculus. There are really two main points of Chapters 2 and 3:

1. The operators $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ can be considered as the three components of a vector operator ∇ , and the formulas which result from vector algebra by treating this operator as a vector are correct.

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right).$$

2. The difference of the values of a scalar field at two points is equal to the line integral of the tangential component of the gradient of that scalar along

any curve of all between the first and second points:

$$\psi(2) - \psi(1) = \int_{\text{any path}}^{(2)} \nabla \psi \cdot d\mathbf{r} \quad (2.42)$$

3. The surface integral of the normal component of an arbitrary vector over a closed surface is equal to the integral of the divergence of the vector over the volume interior to the surface:

$$\int_{\text{closed surface}} \mathbf{C} \cdot \mathbf{n} d\mathbf{S} = \int_{\text{volume}} \nabla \cdot \mathbf{C} dV \quad (2.43)$$

4. The line integral of the tangential component of an arbitrary vector around a closed loop is equal to the surface integral of the normal component of the curl of that vector over any surface which is bounded by the loop:

$$\int_{\text{closed loop}} \mathbf{C} \cdot d\mathbf{r} = \int_{\text{any surface}} (\nabla \times \mathbf{C}) \cdot \mathbf{n} d\mathbf{S} \quad (2.44)$$

Electrostatics

4-1 Statics

We begin now our detailed study of the theory of electromagnetism. All of electromagnetism is contained in the Maxwell equations.

Maxwell's equations:

$$\nabla \cdot E = \frac{\rho}{\epsilon_0}, \quad (4.1)$$

$$\nabla \times E = -\frac{\partial B}{\partial t}, \quad (4.2)$$

$$\nabla \cdot B = 0, \quad (4.3)$$

$$\nabla \times B = \frac{\mu_0}{\epsilon_0} J + \frac{J}{c}, \quad (4.4)$$

The situations described by these equations can be very complicated. We will consider first relatively simple situations, and soon how to handle them; next we take up more complicated ones. The easiest circumstance to treat is one in which nothing depends on the time, except the work done. All charges are permanently fixed —静止—, or if they do move, they move at a steady rate in a circuit (so ρ and J are constant in time). In these circumstances, all of the terms in the Maxwell equations which are derivatives of the field are zero. In this case, the Maxwell equations become:

Electricity:

$$\nabla \cdot E = \frac{\rho}{\epsilon_0}, \quad (4.5)$$

$$\nabla \times E = 0. \quad (4.6)$$

Magnetostatics:

$$\nabla \times B = \frac{J}{\epsilon_0 c^2}, \quad (4.7)$$

$$\nabla \cdot B = 0. \quad (4.8)$$

You will notice an interesting thing about this set of four equations. It can be separated into two parts. The electric field E appears only in the first two, and the magnetic field B appears only in the second two. The two fields are not fully uncoupled. This means that electricity and magnetism are distinct phenomena so long as charges and currents are static. The interdependence of E and B does not appear until there are changes in charges or currents, as when a condenser is charged, or a magnet moved. Only when there are sufficiently rapid changes so that the time derivatives in Maxwell's equations become significant, will E and B depend on each other.

Now if you look at the equations of statics you will see that the study of the two subjects we call **electrostatics** and **magnetostatics** is just from the point of view of learning about the mathematical properties of vector fields. Electrostatics is a field example of a vector field with zero divergence and a given direction. Magnetostatics is a field example of a field with zero divergence and a given curl. The more conventional way you may be thinking about electrostatics is a way of presenting

4-1 Statics

4-2 Coulomb's law; superposition

4-3 Electric potential

4-4 $E = -\nabla \phi$

4-5 The flux of E

4-6 Gauss' law; dot divergence of E

4-7 Field of a sphere of charge

4-8 Field lines; equipotential surfaces

*Review Chapters 13 and 14, Vol. I,
Work and Potential Energy*

$$\epsilon_0 = \frac{1}{4\pi G} = 8.85 \times 10^{-12} \text{ coulombs}^2/\text{newton-meters}^2$$

The theory of electromagnetism is vector, first with electrodynamics and then we learn about the divergence, magnetostatics and the curl are taken up later. Finally, electricity and magnetism are put together. We have drawn to that will the incomplete history of vector calculus. Now we shall apply it to the special case of Electrodynamics, the law of E given by the first pair of equations.

We will begin with the simplest situation—cases in which the positions of all charges are specified. If we have only to study electromagnetism at this level (as we shall do in the next two chapters), it would be very simple to form, *exact* laws. Everything can be derived from Coulomb's law and some integration, as you will see. In many real electrostatic problems, however, we do not know initially where the charges are. We know only that they have distributed themselves in ways that depend on the properties of matter. The positions of the charges also depend on the A field, which in turn depends on the positions of the charges. These things can get quite complicated. If, for instance, a charged body is brought near a conductor or insulator, the electrons and protons in the body, either move or will move around. The charge density ρ (Eq. (4.5)) may not be zero, but we know about it from the charge that we brought up; but there will be other parts of the charge that have moved around in the conductor. And all of the charges move or move into account. One can get into some rather subtle and interesting problems. So although this chapter is to be on electrostatics, it will not cover the more beautiful and subtle parts of the subject. It will treat only the situation where we can assume that the positions of all the charges are known. Naturally, you should be able to do that case before you try to handle the other cases.

4-2 Coulomb's law: superposition

It would be logical to use Eqs. (4.3) and (4.6) as our starting points. It will be easier, however, if we can somewhere else and come back to these equations. The results will be equivalent. We will start with a law that we have called above Coulomb's law, which says that between two charges at rest there is a force directly proportional to the product of the charges and inversely proportional to the square of the distance between. The force is along the straight line from one charge to the other.

$$\text{Coulomb's law: } F_1 = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}^2} \hat{v}_{12} = -F_2. \quad (4.9)$$

F_1 is the force on charge q_1 , v_{12} is the unit vector in the direction to q_2 . Since q_1 and q_2 is the distance between q_1 and q_2 . Therefore F_2 on q_2 is equal and opposite to F_1 .

The constant of proportionality, for historical reasons, is written as $1/4\pi\epsilon_0$. In the system of units which we use—the cgs system—it is defined as exactly 10^{-10} times the speed of light squared. Now since the speed of light is approximately 3×10^8 meters per second, the constant is approximately 9×10^9 , and the unit turns out to be newton-meter²/coulomb² or volt-meter per coulomb.

$$\begin{aligned} \frac{1}{4\pi\epsilon_0} &= 10^{-10} \quad (\text{by definition}) \\ &= 9.0 \times 10^9 \quad (\text{by experiment}). \end{aligned} \quad (4.10)$$

(Unit: newton-meter²/coulomb²,
or volt-meter per coulomb).

When there are more than two charges present—the only really interesting place—~~we~~ we supplement Coulomb's law with one other fact of nature: the force on one charge is the vector sum of the Coulomb forces from each of the other charges. This is so called "the principle of superposition." There is nothing to it in electrodynamics. If we combine the Coulomb law and the principle of superposition, there is nothing else. Equations (4.3) and (4.6)—the electrostatic equations—say no more than that.

When applying Coulomb's law, it is convenient to introduce the idea of an electric field. We say that the field $E(1)$ is the force per unit charge at \mathbf{r}_1 (the position of point charges). Dividing Eq. (4.9) by q_1 , we have for one-unit charge located at 1:

$$E(1) = \frac{1}{4\pi\epsilon_0} \frac{q_1}{r_{12}^2} \mathbf{e}_{12} \quad (4.11)$$

Now we consider that $E(1)$ describes something about the point (1) over if q_1 were not there—assuming that all other charges keep their same positions. We say, $E(1)$ is the electric field at the point (1).

The electric field E is a vector, so by Eq. (4.11) we really have three equations—one for each component. Writing out explicitly the x component, Eq. (4.11) becomes

$$E_x(x_1, y_1, z_1) = \frac{q_1}{4\pi\epsilon_0} \left[\frac{x_1 - x_2}{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \right] \mathbf{i} \quad (4.12)$$

and similarly for the other components.

If there are many charges present, the field E at any point (1) is a sum of the contributions from each of the other charges. Each term in the sum will look like Eq. (4.12) or (4.13). Letting q_j be the magnitude of the j th charge, and r_{ij} the displacement from q_j to the point (1), we write

$$\mathbf{E}(1) = \sum_j \frac{1}{4\pi\epsilon_0} \frac{q_j}{r_{ij}^2} \mathbf{e}_{ij}, \quad (4.13)$$

which means, of course,

$$E_i(x_1, y_1, z_1) = \sum_j \frac{1}{4\pi\epsilon_0} \frac{q_j(x_1 - x_j)}{(x_1 - x_j)^2 + (y_1 - y_j)^2 + (z_1 - z_j)^2} \quad (4.14)$$

and so on.

Often it is convenient to ignore the fact that charges come in packages like electrons and protons, and think of them as being spread out in volumes called "charge distributions," see figure 4.1. This is OK, as long as we are not interested in what is happening on too small a scale. We describe a charge distribution by the "charge density," $\rho(x_1, y_1, z_1)$. If the amount of charge ρ in a small volume dV located at the point (x_1, y_1, z_1) is $d\rho$, then ρ is defined by

$$d\rho = \rho(x) dV. \quad (4.15)$$

To use Coulomb's law with such a description, we replace the sums of Eqs. (4.13) or (4.14) by integrals over all volumes containing charges. Then we have

$$\mathbf{E}(1) = \frac{1}{4\pi\epsilon_0} \int_{\text{all}} \frac{\rho(\mathbf{r}') \mathbf{e}_{12}}{r_{12}^2} dV'. \quad (4.16)$$

Some people prefer to write

$$\mathbf{e}_{12} = \frac{\mathbf{r}_{12}}{r_{12}},$$

where \mathbf{r}_{12} is the vector displacement to (1) from (2), as shown in Fig. 4-1. The integral for \mathbf{E} is then written as

$$\mathbf{E}(1) = \frac{1}{4\pi\epsilon_0} \int_{\text{all}} \frac{\rho(\mathbf{r}') \mathbf{r}_{12}}{r_{12}^3} dV'. \quad (4.17)$$

When we start to evaluate something with these integrals, we usually have to write them out in explicit detail. For the x -component, in either Eq. (4.16) or (4.17), we would have

$$E_x(x_1, y_1, z_1) = \int_{\text{all}} \frac{(x_1 - x') \rho(x_1, y_1, z_1) dx' dy' dz'}{4\pi\epsilon_0 r_{12}^3 [(x_1 - x')^2 + (y_1 - y')^2 + (z_1 - z')^2]^{3/2}}. \quad (4.18)$$

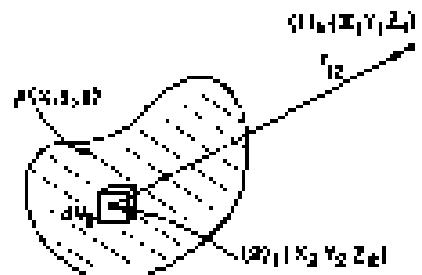


Fig. 4-1. The electric field \mathbf{E} at point (1), from a charge distribution, is obtained by an integral over the distribution. Point (1) could also be inside the distribution.

We are not going to use this formula much. We write it here only to emphasize the fact that we have completely solved all the diagnostic problems in which we know the locations of all of the charges. Given the charges, what are the fields? Answer: This is integral. So there is nothing to the subject: it is just a case of doing complicated integrals over three dimensions—strictly a job for a computing machine!

With our integrals we can find the field produced by a sheet of charge, finite line of charge, finite spherical shell of charge, or from any specified distribution. It is important to realize, as we go on to draw field lines, to talk about potentials, or to calculate divergences, that we already have the answer here. It is really a matter of it being somewhat easier to do an integral by some clever piece of work than by actually carrying it out. The question requires learning all sorts of straightforwards. In practice, it might be easier in most trying to be clever, and always to do the integral directly instead of trying to smart. We are, however, going to try to be smart about it. We shall go on to discuss some other features of the electric field.

4-3 Electric potential

First we introduce the idea of electric potential, which is related to the work done in carrying a charge from one point to another. There is some distribution of charge which produces an electric field. We ask: How much work is needed to carry a small charge from one place to another. The work done against the electrical forces in carrying a charge along some path is the negative of the component of the electrical force in the direction of the motion, integrated along the path. If we carry a charge from point a to point b ,

$$W = - \int_a^b F \cdot d\mathbf{r},$$

where F is the electrical force on the charge at each point, and $d\mathbf{r}$ is the differential vector displacement along the path. (See Fig. 4-2.)

It is more interesting for our purposes to consider the work that would be done in carrying one unit of charge. Then the force on the charge is numerically the same as the electric field. Calling the work done against electrical forces in this case "potential," we have

$$V(\text{unit}) = - \int_a^b E \cdot d\mathbf{r} \quad (4.19)$$

Now, in general, what we get with this kind of an integral depends on the path we take. So if the integral of (4.19) depended on the path from a to b , we could get work out of the field by carrying the charge to b along one path and then back to a on the other. We would go back along the path for which V is smaller and back along the other, paying off more work than we put in.

"There is nothing impossible." In principle, above getting energy out of a field, "Work," in fact, amounts to where it is possible. It could be that to you move a charge you produce forces on the other part of the "machinery." If the "machinery" moves against the force it would lose energy, thereby keeping the total energy in the work constant. For instance, however, there is no such "machinery." We know what the forces look on the geometry of the field lines. They are the Coulomb forces on the charges responsible for the field. If the other charges are fixed in position, so we assume in electrostatics only these last forces can do no work or harm. There is no way to get energy from them—provided, of course, that the principle of energy conservation works for electrostatic situations. We believe that it will work, but let's just review that it must follow from Coulomb's law of force:

We consider the work happens in the field due to a single charge q . Let point a be at the distance r_1 from q , and point b at r_2 . Now we carry a different charge, which we will call the "test" charge, and whose magnitude we choose to



Fig. 4-2. The work done in carrying a charge from a to b is the negative of the integral of $F \cdot d\mathbf{r}$ along the path taken.

so one take from a to b . Let's start with the easiest possible path to evaluate. We carry our test charge first from the left, then up a circle, then down, as shown in part (a) of Fig. 4-3. Since on the return, the path it is difficult to find the work done (otherwise we wouldn't have picked it). First, there is no work done at all on the path from a to b . The field is radial (from Coulomb's law), so it is parallel to the direction of motion. Next, on the path from b to a , the field is in the direction of motion and varies as $1/r^2$. Thus the work done by the test charge in carrying it from a to b would be

$$-\int_a^b \mathbf{F} \cdot d\mathbf{r} = -\frac{q}{4\pi\epsilon_0} \int_{r_0}^R \frac{dr}{r^2} = -\frac{q}{4\pi\epsilon_0} \left(\frac{1}{r_0} - \frac{1}{R} \right). \quad (4.20)$$

Now let's take another easy path. For instance, the one shown in part (b) of Fig. 4-3. It goes for awhile along surface of a circle, then radially outwards, then radially inwards. Every time we go along the circular parts, we do no work. Every time we go along the radial parts, we must just integrate $1/r^2$. Along the first radial stretch, we integrate from r_0 to r_1 , then along the next radial stretch from r_1 to r_2 , and so on. The sum of all these integrals is the same as a single integral directly from r_0 to R . We get the same answer for the path that we did for the first path, as listed. It is clear that we could get the same answer for any path which is made up of an arbitrary number of the same kinds of pieces.

What about smooth paths? What we get the same answer? We discussed this point previously in Chapter 13 of Vol. I. Applying the same arguments used there, we can conclude that work done in carrying a test charge from a to b is independent of the path.

$$\left[\Phi(\text{final}) - \Phi(\text{initial}) \right]_{a \rightarrow b} = - \int_a^b \mathbf{F} \cdot d\mathbf{r}.$$

Since the work done depends only on the endpoints, it can be represented as the difference between two numbers. We can see this in the following way. Let's choose a reference point P_0 and agree to evaluate our integral by using a path that always passes near enough to P_0 . Let $\phi(\mathbf{r})$ stand for the work done against the field by going from P_0 to point a , and let $\phi(\mathbf{r})$ be the work done in going from P_0 to point b (Fig. 4-3). The work in going to a from P_0 (or the way to b) is the negative of $\phi(a)$, so we have that

$$-\int_a^b \mathbf{F} \cdot d\mathbf{r} = \phi(b) - \phi(a). \quad (4.21)$$

Since only the difference in the function ϕ at our points is ever involved, we do not really have to specify the location of P_0 . Once we have chosen some reference point, however, a number ϕ is determined for any point in space; ϕ is then a scalar field. It is a function of x, y, z . We call this scalar function the *electrostatic potential* at any point.

Electrostatic potential:

$$\phi(\mathbf{r}) = - \int_{P_0}^{\mathbf{r}} \mathbf{E} \cdot d\mathbf{r}. \quad (4.22)$$

For convenience, we will often take the reference point at infinity. Then, for a single charge at the origin, the potential ϕ is given for any point (x, y, z) using Eq. (4.20):

$$\phi(x, y, z) = \frac{q}{4\pi\epsilon_0} \frac{1}{r}. \quad (4.23)$$

The electric field from several charges can be written as the sum of the electric field from the first, from the second, from the third, etc. When we integrate the sum to find the potential we get a sum of integrals. Each of the integrals is the

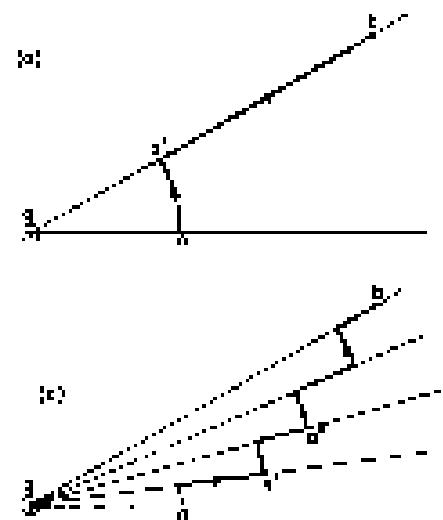


Fig. 4-3. In carrying a test charge from a to b the same work is done along either path.

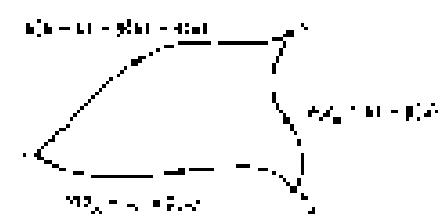


Fig. 4-4. The work done in going along any path from a to b is the negative of the work from some point P_0 to a plus the work from P_0 to b .

potential from one of the charges. We conclude that the potential ϕ from a lot of charges is the sum of the potentials from all the individual charges. There is a superposition principle also for potentials. Using the same kind of arguments by which we found the electric field from a group of charges, and for a distribution of charge, we can get the complete formulas for the potential ϕ at a point we call (1):

$$\phi(1) = \sum_j \frac{1}{4\pi\epsilon_0} \frac{q_j}{r_{1j}}, \quad (4.24)$$

$$\phi(1) = \frac{1}{4\pi\epsilon_0} \int_{V_{ext}}^1 \rho(2) dV_2. \quad (4.25)$$

Remember that the potential ϕ has a physical significance: it is the potential energy which a unit charge would have if brought to the specified place in space from some reference point.

4.4 $E = -\nabla\phi$

We discussed above ϕ ? Forces on charges are given by E , the electric field. The point is that E can be obtained easily from ϕ —it is so easy, in fact, as taking a derivative. Consider two points, one at x and one at $(x + dx)$, but both at the same y and z , and ask how much work is done in carrying a test charge from one point to the other. The path is along the horizontal line from x to $x + dx$. The work done is the difference in the potential at the two points:

$$\Delta W = \phi(x + dx, y, z) - \phi(x, y, z) = \frac{\partial \phi}{\partial x} dx.$$

But the work done against the field for the same path is

$$\Delta W' = -\int E \cdot dx = -E_x dx.$$

We see that

$$E_x = -\frac{\partial \phi}{\partial x}. \quad (4.26)$$

Similarly, $E_y = -\partial \phi / \partial y$, $E_z = -\partial \phi / \partial z$, or, summarizing with the notation of vector analysis,

$$\mathbf{E} = -\nabla\phi. \quad (4.27)$$

This equation is the differential form of Eq. (4.22). Any problem with spherical charges can be solved by computing the potential from (4.24) or (4.25) and using (4.27) to get the field. Equation (4.27) also agrees with what we found from vector calculus that for any scalar field ϕ

$$\int_{\gamma}^{\gamma'} \nabla\phi \cdot d\mathbf{l} = \phi(\gamma') - \phi(\gamma). \quad (4.28)$$

According to Eq. (4.27) the scalar potential ϕ is given by a three-dimensional integral similar to the one we had for E . Is there any advantage to computing ϕ rather than E ? Yes. There is only one integral for ϕ , while there are three integrals for E —because E is a vector. Furthermore, ϕ is usually a little easier to integrate than E_x/E^2 . It turns out in many practical cases that it is easier to calculate ϕ and then take the gradient to find the electric field, than it is to evaluate the three integrals for E . It is merely a potential answer.

There is also a deeper physical significance to the potential ϕ . We have shown that E in Coulomb's law is obtained from $E = -\nabla\phi$ if ϕ is given by (4.24). If ϕ is equal to the gradient of a scalar field, then we know from the vector calculus that the curl of E must vanish:

$$\nabla \times \mathbf{E} = 0. \quad (4.29)$$

But that is just our second fundamental equation of electrostatics, Eq. (4.6). We have shown that Coulomb's law gives an E field that satisfies that condition. So far, our finding is all right.

We had really generalized $\nabla \times E$ was zero before we defined the potential. We had shown that the work done around a closed path is zero. That is, that

$$\oint \mathbf{E} \cdot d\mathbf{s} = 0$$

for any path. We saw in Chapter 3 that for any such field $\nabla \times E$ must be zero everywhere. The electric field in electrostatics is an example of a curl-free field.

You can practice your vector calculus by proving that $\nabla \times E$ is zero in a different way—by computing the components of $\nabla \times E$ at the field at a point shown as given by Eq. (3.11). If you get zero, the superposition principle says you would get zero for the field of any charge distribution.

We should point out an important fact. For any field since the work done is independent of the path and from which a potential. If you think about it, the entire argument we made above to show that the work integral was independent of the path depended only on the fact that the field from a single charge was radial and spherically symmetric. If the non-dependence of the field on the distance was not $1/r^2$ —there could still be a dependence. The existence of a potential, and the fact that the curl of E is zero, comes really only from the symmetry and direction of the electrostatic forces. Because of this, Eq. (4.28)—or (4.29)—can contain only part of the laws of electricity.

4-6 The flux of E

We will now derive a field equation that depends specifically and exactly on the fact that the force law is inverse square. That the field varies inversely as the square of the distance seems, for some people, to be "fairly natural," because "that's the way things spread out." Take a light source with light spreading out, the amount of light that passes through a surface cut out by a cone with its apex at the center of the source—what happens the surface is placed? I must know if there is to be conservation of light energy. The density of light per unit area—the intensity—is proportional to the area cut by the cone, i.e., inversely as the square of the distance from the source. Comparing the electric field should vary inversely as the source of the distance for the same reason. But there is no such thing as the "source power" here. Nothing can say that the electric field measures the flow of something like light which must be conserved. (If we had a "model" of the electric field in which the electric field vector represented the direction and speed, say, the current, of some kind of little "bullets" which were flying out, and if our model required that these bullets were however that they could ever disappear once it was shot out of a source, then we might say that we can "see" that the inverse square law is necessary. On the other hand, there would necessarily be some mathematical way to express this "bullet" idea. If the electric field were like conserved bullets going out, then it would vary inversely as the source of the distance and we would be able to describe that behavior by an equation—which is purely mathematical. Now this is not true in this simple way, so long as we do not say that the electric field is made out of bullets, but rather that we are using a model to help us find the right mathematics.)

Suppose, instead, that we imagine for a moment, that the electric field did represent the flow of something that was conserved—everywhere, that is, except at charges. (It has to start somewhere!) We imagine that whenever it is flows out of a charge into the space around. If E were the vector of such a flow (as \mathbf{J} is for heat flow), it would have a $1/r^2$ dependence near a point charge. Now we wish to use this model to find out how to state the inverse square law in a deeper or more subtle way, rather than simply saying "inversely proportional." You may wonder why we should want to avoid the simple statement of such a simple law, and want instead to imply the same thing subtlety in a different way. Because it will turn out to be useful.

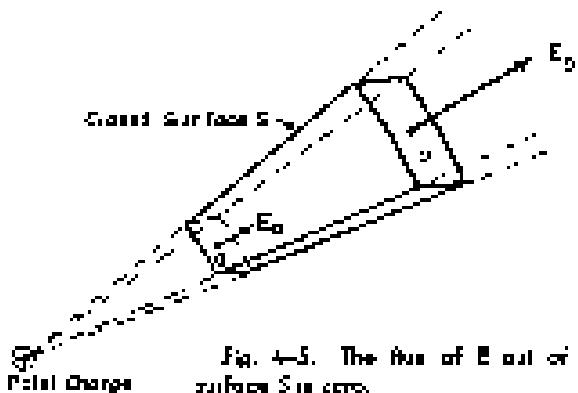


Fig. 4-5. The flux of E out of the surface S is zero.

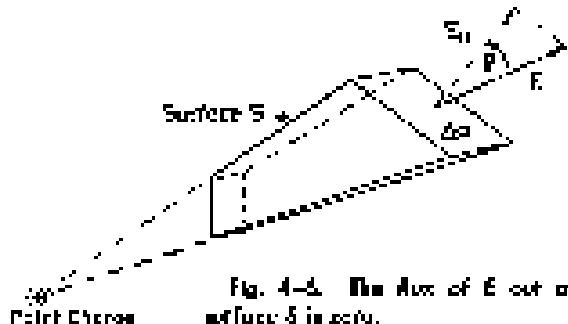


Fig. 4-6. The flux of E out of the surface S is zero.

We ask, "What is the "flux" of E out of an arbitrary closed surface in the neighborhood of a point charge?" First let's take an easy surface—the one shown in Fig. 4-5. If the E field is like it is now, the net flow out of this box should be zero. That is what we get; if by the "flux" from this surface we mean the surface integral of the normal component of E —that is, the flux of E on the radial faces, the normal component is zero. On the spherical faces, the normal component E_n is just the magnitude of E times the sine of the smaller face solid angle for the larger side. The magnitude of E increases as $1/r^2$, but the surface area is proportional to r^2 , so the product is independent of r . The flux of E into face a is just canceled by the flux out of face b . The total flux out of S is zero, which is to say that for this surface:

$$\int_S E_n d\sigma = 0. \quad (4-3)$$

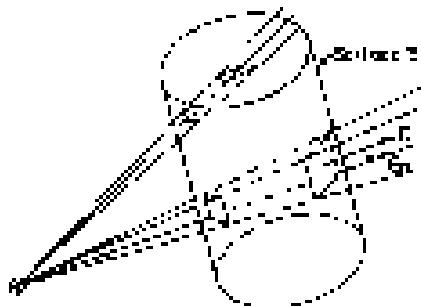


Fig. 4-7. Any volume can be thought of as completely made up of infinitesimal truncated cones. The flux of E from one end of each truncated segment is equal and opposite to the flux from the other end. The total flux over the surface S is therefore zero.

Next we note that the two end surfaces may be closed with respect to the radial line without changing the integral (4-3). Although it is true in general, for our purposes it is only necessary to show that this is true when the end surfaces are small, so that they subtend a small angle from the source—in fact, no retinocular angle. In Fig. 4-6 we show a surface S whose "ends" are tilted, but whose "ends" are tilted. The end surfaces represent small in the figure, but you are to imagine the situation for very small end surfaces. Since the field E will be sufficiently uniform over the surface that we can use just its value at the center, when we tilt the surface by an angle θ , the area is increased by the factor $1/\cos \theta$. But E_n , the component of E normal to the surface, is decreased by the factor $\cos \theta$. The product $E_n d\sigma$ is unchanged. The flux out of the whole surface S is still zero.

Now it is easy to see how the flux out of a volume enclosed by a surface S must be zero. Any volume can be thought of as made up of pieces, like that in Fig. 4-6. The surface will be subdivided completely into pairs of end surfaces, and opposite fluxes in and out of these end surfaces cancel by pairs. The total flux out of the surface will be zero. This idea is illustrated in Fig. 4-7. We have the completely general result that the total flux of E out of any surface S at the field of a point charge is zero.

Our proof works only if the surface S does not surround the charge. What would happen if the point charge were inside the surface? We could still divide our surface into pairs of ones that are matched by radial lines through the charge, as shown in Fig. 4-8. The fluxes through the two surfaces are still equal—by the same argument as before—but they have the same sign. The flux out of a surface that contains a charge is not zero. Then what is it? We can find out by a little trick. Suppose we "remove" the charge from the "inside" by surrounding the charge by a little surface S' totally inside the original surface S , as shown in Fig. 4-9. Now the volume enclosed between the two surfaces S and S' has no charge in it. The total flux out of this volume (including that through S') is zero, by the arguments we have given above. The arguments tell us, in fact, that the flux into the volume through S' is the same as the flux outward through S .

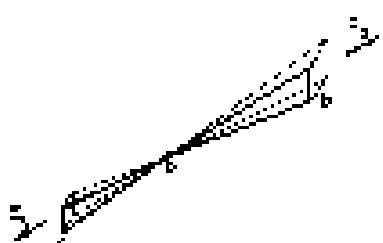


Fig. 4-8. If a charge is inside a surface, the flux out is not zero.

We can choose any surface we want for S' , so let's make it a sphere centered on the charge, as in Fig. 4-10. Then we can easily calculate the flux through it. If the radius of the little sphere is r , the value of E everywhere on its surface is

$$\frac{q}{4\pi r^2 \epsilon_0},$$

and is directed always normal to the surface. We find the total flux through S' if we multiply this normal component of E by the surface area:

$$\text{Flux through the surface } S' = \left(\frac{1}{4\pi \epsilon_0 r^2} q \right) (4\pi r^2) = \frac{q}{\epsilon_0}. \quad (4.31)$$

A number independent of the radius of the sphere! We know then that the flux outward through S is also q/ϵ_0 , or value independent of the shape of S as long as the charge q is inside.

We can state our conclusion as follows:

$$\int_{\text{surfaces } S} E_n dA = \sum_{\text{inside } S} q \quad (4.32)$$

Let's return to our "bullet" analogy and see if it makes sense. Our theorem says that the net flow of bullets through a surface is zero if the surface does not enclose the gun that shoots the bullets. If the gun is located in a surface, inside or outside and charges it is, the number of bullets passing through is the same as is given by the rate at which bullets are generated at the gun. It all seems quite reasonable for current bullets. But then the number will be anything more than we get simply by writing Eq. (4.32)! We can't be surprised in making that "bullet" do anything else can produce this new law. Also, that they start shooting bullets. That is why we prefer to represent the electrostatic field purely vectorially.

4-6 Gauss' law; the divergence of E

The last result, Eq. (4.32), was proved for a single point charge. Now suppose there are two charges, a charge q_1 at one point and a charge q_2 at another. The problem looks more difficult. The electric field whose normal component we integrate for the flux is the sum due to both charges. That is, if E_1 represents the electric field that would have been produced by q_1 alone, and E_2 represents the electric field generated by q_2 alone, the total electric field is $E = E_1 + E_2$. The flux through any closed surface S is

$$\int_S (E_{1n} + E_{2n}) dA = \int_S E_{1n} dA + \int_S E_{2n} dA. \quad (4.33)$$

The flux with both charges present is the flux due to a single charge plus the flux due to the other charge. If both charges are outside S , the flux through S is zero. If you integrate E and q_2 is suddenly there, the first integral goes to zero and the second integral gives q_2/ϵ_0 . If the surface encloses both charges, each will generate contribution and we have that the flux is $q_1 + q_2/\epsilon_0$. The general rule is clearly then the total flux out of a closed surface is equal to the total charge inside, divided by ϵ_0 .

Our result is an important general law of the electrostatic field, called Gauss' law.

$$\text{Gauss' law: } \int_{\text{closed surface } S} E_n dA = \frac{\text{sum of charges inside}}{\epsilon_0}. \quad (4.34)$$

$$\text{or} \quad \int_{\text{closed surface } S} E_n dA = \frac{Q_{\text{in}}}{\epsilon_0}, \quad (4.35)$$

$$\text{where:} \quad Q_{\text{in}} = \sum_{\text{inside } S} q_i. \quad (4.36)$$

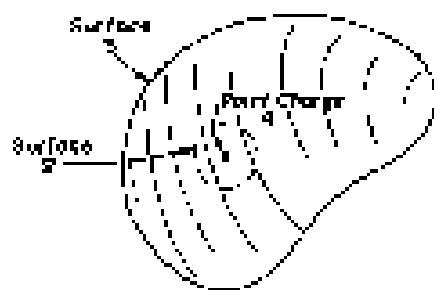


Fig. 4-9. The flux through S is the same as the flux through S' .

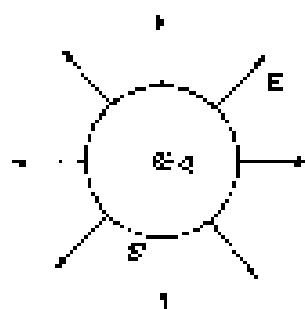


Fig. 4-10. The flux through a spherical surface enclosing a point charge q is q/ϵ_0 .

If we describe the location of charges in terms of a charge density ρ , we can consider that each infinitesimal volume dV contains a "point" charge ρdV . The sum over all charges is then the integral

$$Q_{\text{tot}} = \int_{\text{Volume}} \rho dV. \quad (4.37)$$

From our derivation you see that Gauss' law follows from the fact that the exponent in Coulomb's law is exactly two. A $1/r^2$ field, or any $1/r^n$ field with $n \neq 2$, would not give Gauss' law. So Gauss' law is just an expression in a different form of the Coulomb law of forces between two charges. In fact, working back from Gauss' law, you can derive Coulomb's law. The two are equivalent so long as we keep in mind the rule that the force between charges is radial.

We would now like to write Gauss' law in terms of derivatives. To do this, we apply Gauss' law to an infinitesimal cubical surface. We saw in Chapter 3 that the flux of E out of such a cube is $\nabla \cdot E$ times the volume dV of the cube. The charge inside of dV , by the definition of ρ , is equal to ρdV . So Gauss' law gives

$$\nabla \cdot E dV = \frac{\rho dV}{\epsilon_0},$$

or

$$\nabla \cdot E = \frac{\rho}{\epsilon_0}. \quad (4.38)$$

The differential form of Gauss' law is the first of the fundamental field equations of electrodynamics (Eq. 4.38). We have now shown that the two equations of electrostatics, Eqs. (4.5) and (4.6), are equivalent to Coulomb's law of force. We will now see the nice example of the use of Gauss' law. (We will soon refer to many more examples.)

4-7 Field of a sphere of charge

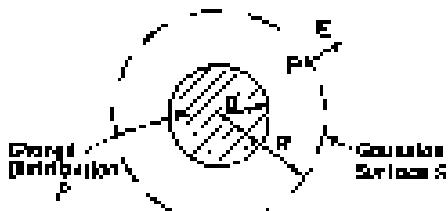


Fig. 4-11. Using Gauss' law to find the field of a uniform sphere of charge.

One of the difficult problems we had when we studied the theory of gravitational attraction was to prove that the force on a shell by a solid sphere of matter was the same as the surface of the sphere as it would be if the matter were concentrated on the surface. For many years Newton didn't make public his theory of gravitation, because he couldn't be sure this theory was true. He proved the theorem in Chapter 13 of Vol. I by doing the integral for the potential and then finding the gravitational force by using the gradient. Now we can prove the theorem in a much simpler fashion. Only this time we will prove the corresponding theorem for a uniform sphere of electrical charge. (Since the laws of electrodynamics are the same as those of gravitation, the same proof could be done for the gravitational field.)

We ask: What is the electric field E at a point P anywhere outside the surface of a sphere filled with a uniform distribution of charge? Since there is no "spur" direction, we can assume that E is everywhere directed away from the center of the sphere. We consider an imaginary surface that is spherical and concentric with the sphere of charge, and that passes through the point P (Fig. 4-11). For this surface, the flux outward is

$$\oint E_i dA = E \cdot 4\pi R^2.$$

Gauss' law tells us that this flux is equal to the total charge Q of the sphere (over ϵ_0):

$$E \cdot 4\pi R^2 = \frac{Q}{\epsilon_0},$$

or

$$E = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2}. \quad (4.39)$$

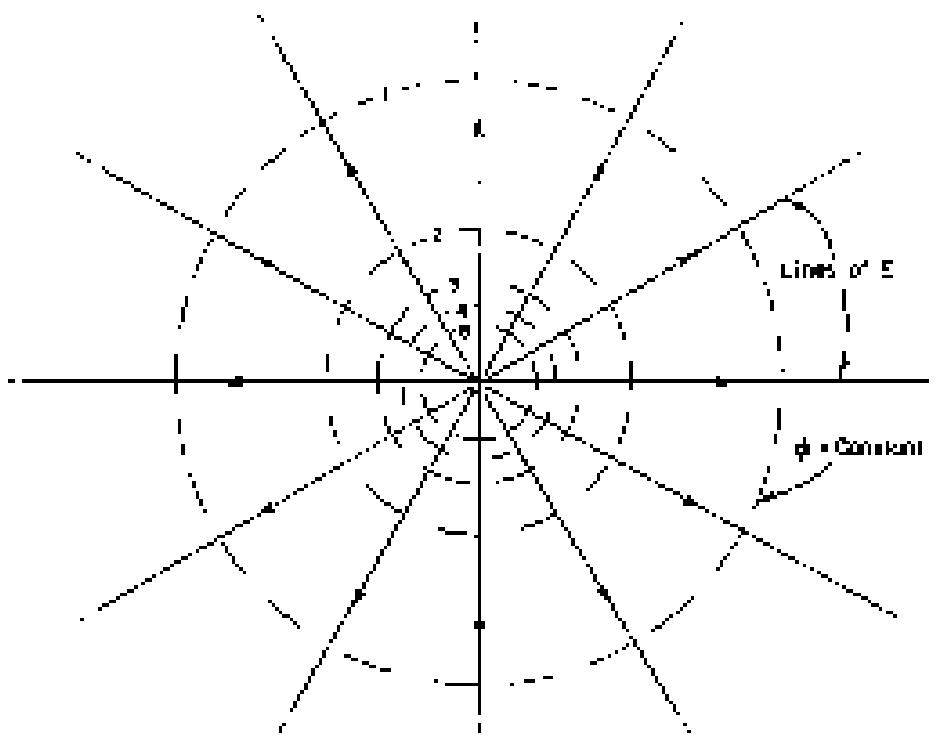


Fig. 4-12. Field lines and equipotential surfaces for a positive point charge.

which is the same formula we would have had a point charge q . We have proved Newton's problem more easily than by going the trigonal. This, of course, is false kind of reasoning—it has taken you some time to be able to understand Gauss' law, so you may think the new method has really been saved. But let you have used the theorem more and more, it begins to pay. It is a question of efficiency.

4-8 FIELD LINES; EQUIPOTENTIAL SURFACES

We would like now to give a geometrical description of the electric field. The two laws of electrostatics, one that the field is proportional to the charge inside and the other that the electric field is the gradient of a potential, can also be represented geometrically. We illustrate this with two examples.

First, we take the field of a point charge. We draw lines in the direction of the field—lines which are always tangent to the field, as in Fig. 4-12. These are called field lines. The lines show everywhere the orientation of the electric vector. But we also wish to represent the magnitude of the vector. We can do this now, for the strength of the electric field will be represented by the "density" of the lines. By this density we mean the number of lines per unit area through a surface perpendicular to the lines. With these two rules we can draw a picture of the electric field. For a point charge, the density of the lines must decrease as $1/r^2$. But the area of a spherical surface is proportional to the lines, so if radius increases as r^2 , so if we always keep the same number of lines for all distances from the charge, the density will remain proportional to the magnitude of the field. We can guarantee that there are the same number of lines at every distance if we make the lines be continuous. Since once a line is emitted from the charge, it never stops. In terms of the field lines, Gauss' law says that lines should start only at plus charges and stop at minus charges. The number where there is a charge q will be equal to q/e_0 .

Now, we can find a similar geometrical picture for the potential ϕ . The easiest way to represent the potential is to draw surfaces on which ϕ is a constant. We call them *equipotential surfaces*—surfaces of equal potential. Now, repeat the geometrical

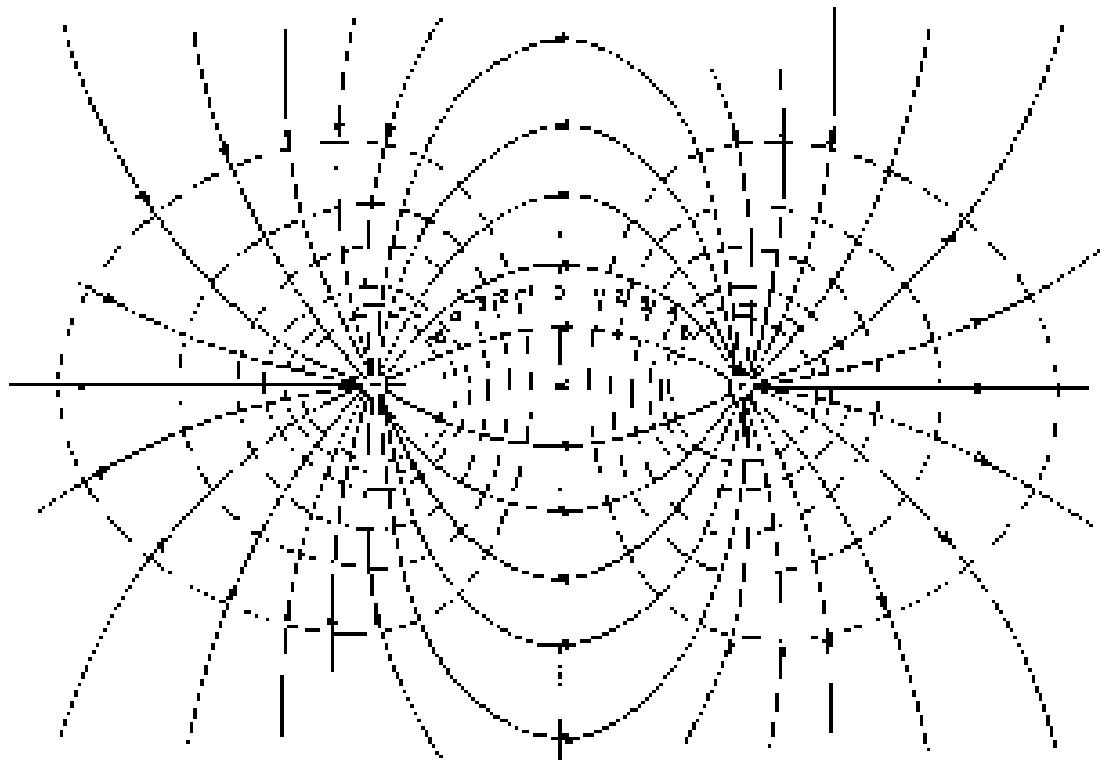


Fig. 4-12. Field lines and equipotential's for two equal and opposite point charges.

cal relationship of the equipotential surfaces to the field lines? The electric field is the gradient of the potential. The gradient is in the direction of the most rapid change of the potential, and is therefore perpendicular to an equipotential surface. If \mathbf{E} were not perpendicular to the surface, it would have a component to the surface. The potential would be changing in the surface, but then it wouldn't be an equipotential! The equipotential surfaces make them be everywhere at right angles to the electric field lines.

For a point charge q by itself, the equipotential surfaces are spheres centered on the charge. We have shown in Fig. 4-12 the interaction of these spheres with a plane through the charge.

As a second example, we consider the field near two equal charges, a positive one and a negative one. To get the field is easy. The field is the superposition of the fields from each of the two charges. So, we can take two point charges (Fig. 4-12) and superimpose them. It's possible! But we would have field lines meeting each other, and this is not possible, because \mathbf{E} can't have two directions at the same point. The disadvantage of the head-on picture is now obvious. By geometrical arguments it is impossible to analyze it a very simple way where the field lines go. From the two independent pictures, we can't get the combined picture. The principle of superposition, a simple and deep principle about electric fields, does not have, at the field-line picture, an easy representation.

The field-line picture has its uses, however, so we might still like to draw the picture for a pair of equal (but opposite) charges. If we calculate the field from Eq. (4.15) and the potentials from (4.23), we can draw the field lines and equipotential's. Figure 4-13 shows the result. But we first had to solve the problem *mathematically*!

4 Note sheet Units	
Quantity	Units
F	Newton
G	newton-m
L	meter
W	Joule
$e = Gm^2$	newton-meter ²
$V_{\text{ext}} \sim FL^2/Q^2$	newton-meters ² /coulomb ²
$N = 8\pi G$	newton-meters ²
$\phi \sim W/Q$	joule/coulomb = V
$E \sim \phi/r$	N/Coulomb
$V_{\text{ext}} \sim \pi L^2 Q^2$	newton-meters/coulomb ²

Application of Gauss' Law

5-1 Electrostatics in Gauss' law plus...

There are two laws of electrostatics: that the flux of the electric field from a volume is proportional to the charge inside—Gauss' law, and that the circulation of the electric field is zero. \mathbf{E} is a gradient. From these two laws, all the predictions of electrostatics follow. But, in my view, this is unattractive in one sense, because there is only one with a certain amount of ingenuity, is another. In this chapter we will work through a number of calculations which can be made with Gauss' law alone. We will prove theorems and describe some effects, particularly in conductors, that can be understood very easily from Gauss' law. Gauss' law by itself cannot give the solution of any problem because the other law must be obeyed also. So when we use Gauss' law for the solution of particular problems, we will tend to add something to it. We will have to incorporate, for instance, some idea of how the field looks—based, for example, on arguments of symmetry. Or we may have to introduce specifically the idea that the field is the gradient of a potential.

5-2 Equilibrium in an electrostatic field

Consider first the following question: When can a point charge be in stable mechanical equilibrium in the electric field of a bar charge? As an example, imagine three negative charges at the corners of an equilateral triangle in a horizontal plane. Would a positive charge placed at the center of the triangle remain there? It will be simpler if we ignore gravity for the moment, although including it would not change the results. The force on the positive charge is zero, but is the equilibrium stable? Would the charge return to the equilibrium position if displaced slightly? The answer is no.

There are no points of stable equilibrium in any electrostatic field—except right on top of another charge. Using Gauss' law, it is easy to see why. First, since charges can be in equilibrium at any particular point P_1 , the field must be zero. Second, if the equilibrium is to be a stable one, we require that if we move the charge away from P_1 in any direction, there should be a restoring force directed opposite to the displacement. The electric field at all nearby points must be pointing inward—toward the point P_1 . But that is in violation of Gauss' law, if there is no charge at P_1 , as we can easily see.

Consider a tiny imaginary surface that encloses P_1 , as in Fig. 5-1. If the electric field everywhere in the vicinity is pointed toward P_1 , the surface integral of the normal component is certainly not zero. For the case shown in the figure, the flux through the surface must be a negative number. By Gauss' law, this the flux of electric field through any surface is proportional to the total charge inside. If there is no charge at P_1 , the field will have violated previous Gauss' law. It is impossible to balance a positive charge in empty space—at a point where there is no some negative charge. A positive charge can be in equilibrium if it is in the middle of a distributed negative charge. Of course, the negative charge distribution would have to be held in place by other non-electrical forces!

Our result has been negative for a point charge. Does this mean that, for a complicated arrangement of excesses held together in fixed relative positions—with rods, for example? We consider the question for two equal charges fixed on a rod. Is it possible that this combination can be in equilibrium in some electrostatic field? The answer is again no. The total force on the rod cannot be balancing for displacements in every direction.

- 5-1 Electrostatics in Gauss' law plus...
- 5-2 Equilibrium in an electrostatic field
- 5-3 Equilibrium with conductors
- 5-4 Stability of atoms
- 5-5 The field of a bar charge
- 5-6 A sheet of charge; two sheets
- 5-7 A sphere of charge; a spherical shell
- 5-8 Is the field of a point charge exactly $1/r^2$?
- 5-9 The field of a conductor
- 5-10 The field in a cavity of a conductor



Fig. 5-1. If P_1 were in position of stable equilibrium for a positive charge, the electric field everywhere in the neighborhood would point toward P_1 .

Call E the total force on the rod in its position. E is then a vector field. Following the argument just above, we conclude that at a position of static equilibrium, the divergence of E must be a negative number. But the total force on the rod is the first charge times the field at its position, plus the second charge times the field at its position:

$$E = q_1 E_x + q_2 E_y. \quad (3.1)$$

The divergence of E is given by

$$\nabla \cdot E = q_1 (\nabla \cdot E_x) + q_2 (\nabla \cdot E_y).$$

If each of the two charges q_1 and q_2 is in free space, both $\nabla \cdot E_x$ and $\nabla \cdot E_y$ are zero, and $\nabla \cdot E$ is zero—not negative, as would be required for equilibrium. You can see that an extension of the argument shows that no finite combination of any number of charges can have a position of static equilibrium in an electrostatic field in free space.

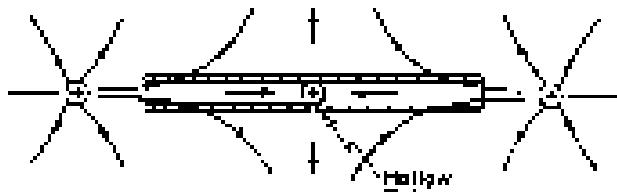


Fig. 3-2 A charge can be in equilibrium if there are mechanical constraints

Now we have not shown that equilibrium is forbidden if there are forces or other mechanical constraints. An example exists: a hollow tube in which a charge can move back and forth freely, but not sideways. It's very easy to derive an electric field that points toward both ends of the tube if it is allowed that the field has point intensity $q_1/q\pi r^2$ near the center of the tube. We simply place positive charges q at each end of the tube, as in Fig. 3-2. There remains now a equilibrium point even though the divergence of E is zero. The charge, of course, would not be in static equilibrium for sideways motion were it not for "non-electrical" forces from the tube walls.

3-3 Equilibrium with conduction

There is no simple spot in the field of a system of fixed charges. What about a system of charged conductors? Can a system of charged conductors produce a field that will have a static equilibrium point for a point charge? (We mean at a point other than on a conductor, of course.) You know that conductors have the property that charges do not freely travel in them. Perhaps when the point charge is displaced slightly, the other charges on the conductor will move in a way that will give a retarding force to the point charge. This answer is still doubtful; we have just given doesn't prove it. The proof for this case is more difficult, and we will only indicate how it goes.

First, we note that when charges redistribute themselves on the conductors, they can only do so if their motion decreases their total potential energy. (Some energy is lost, in fact, as they move in the conductor.) Now we have already shown that if the charges producing a field are stationary, there is, near any zero point P_0 in the field, some direction for which moving a point charge away from P_0 will decrease the energy of the system (since the forces are zero from P_0). Any redistribution of the charges on the conductors can only lower the potential energy until zero, so by the principle of virtual work, their motion will only increase the force in that particular direction away from P_0 , and not toward it.

Our conclusions do not mean that it is not possible to balance a charge by electrical forces. It is possible if one is willing to control the location or the sizes of the supporting charges with sufficient care. You know that a point charge on a point in a gravitational field is unstable, but this does not prove that it cannot be balanced on the end of a string. Similarly, a charge can be held in one spot by electric fields if they are suitable. But not with a *point*—that is, a static—system.

5-4 Stability of atoms

If charges could be held steady at position, it is surely not possible to keep the nuclei from being torn up of the negative charges (electrons) and pulled apart entirely by the law of electrostatics. Such a stable configuration is impossible; it would collapse!

It was suggested that the positive charge of an atom must be distributed uniformly in a sphere, and the negative charges, the electrons, could be at rest outside the positive charge, as shown in Fig. 5-3. This was the first atomic model, proposed by Thomson. But Rutherford concluded from the experiment of Geiger and Marsden that the positive charges were very much concentrated, in what he called the nucleus. Thus Thomson's static model had to be abandoned. Rutherford and Bohr then suggested that the equilibrium might be dynamic, with the electrons revolving in orbits, as shown in Fig. 5-4. The electrons would be kept from falling toward the nucleus by their orbital motion. We already know at least one difficulty with this picture. With such motion, the electrons would be accelerating (because of the circular motion) and would therefore be radiating energy. They would lose kinetic energy required to stay in orbit, and would spiral toward the nucleus. Again unstable!

The stability of the system is now explained in terms of quantum mechanics. The electrostatic forces pull the electron as close to the nucleus as possible, but the electron is compelled to stay several times over a distance given by the uncertainty principle. It is more confined in how small a space, so it would have a great uncertainty in momentum. Our rule assures that it would have a high expected energy, which it would get in excess from the Coulomb attraction. The net result is an electrical repulsion which differs from the model of Thompson — only it is not negative charge that is spread out (because the mass of the electron is so much smaller than the mass of the proton).

5-5 The field of a line charge

Gauss' law can be used to solve a number of electrostatic field problems involving a special symmetry—usually spherical, cylindrical, or plane symmetry. In the remainder of this chapter we will apply Gauss' law to a few such problems. The ease with which these problems can be solved may give the misleading impression that the method is very powerful, and that one should be able to go on to many other problems. It is unfortunately not so. One soon exhausts the list of problems that can be solved easily with Gauss' law. In later chapters we will develop more powerful methods for investigating electrostatic fields.

As our first example, we consider a system with cylindrical symmetry. Suppose that we have a very long, uniformly charged wire. By this we mean that electric charge is distributed uniformly along a indefinitely long straight line, with the charge λ per unit length. We wish to know the electric field. The problem can, of course, be solved by integrating the contribution to the field from every part of the line. We are going to do it using the gauging, by using Gauss' law and some guesswork. First, we assume that the electric field will be directed radially outward from the line. Any axial component from charges on one side would be accompanied by an equal and opposite force from charges on the other side. The result would only be a radial field. It also seems reasonable that the field should have the same magnitude at all points equidistant from the line. This is, however, (it is not necessary to prove, but it is true if space is symmetric)—as we believe it is.

We can use Gauss' law to see thisalog way. We consider an hemispherical surface in the shape of a cylinder intersecting with the line, as shown in Fig. 5-5. Around up a Gauss' law, the total flux of E from this surface is equal to the charge inside divided by ϵ_0 . Since the field is assumed to be zero outside the surface, the inward current density is the magnitude of the field. Let's call it E . Also let the radius of the cylinder be a , and its length be b (in your unit for convenience). The flux through the cylindrical surface is equal to E times the area of the surface, which is $2\pi r^2$. The flux through the two end faces is zero because the electric field is tan-

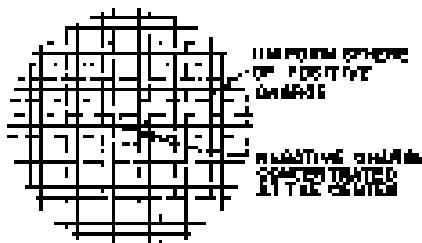


Fig. 5-3. The Thomson model of an atom.

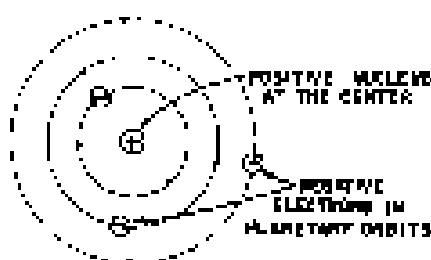


Fig. 5-4. The Rutherford-Bohr model of an atom.

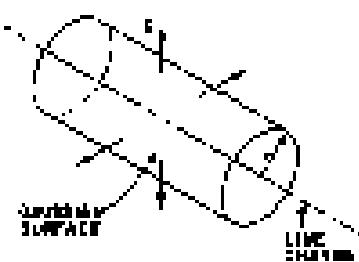


Fig. 5-5. A cylindrical Gaussian surface intersected with a line charge.

parallel to them. The total charge inside our surface is just λ , having the length of the surface as one axis. Gauss' law then gives

$$\begin{aligned} 1. \quad 2\pi r &= \lambda/4\epsilon_0, \\ 2. \quad E &= \frac{\lambda}{2\pi r \epsilon_0}. \end{aligned} \quad (5.2)$$

The electric field of a line charge depends inversely on the first power of the distance from the line.

5-6 A sheet of charges: two sheets

As another example, we can calculate the field from a uniform plane sheet of charge. Suppose that the sheet is infinite in extent and that the charge per unit area is σ . We are going to take another guess. Considerations of symmetry lead us to believe that the field direction is everywhere normal to the plane, and if no other fields from any other charges in the system, the field's magnitude must be the same (in magnitude) on each side. This time we choose for our Gaussian surface a rectangular box that cuts through the sheet, as shown in Fig. 5-6. The two faces parallel to the sheet will have equal areas, say A . The field is normal to these two faces, and parallel to the other four. The total flux is E times the area of the top face, plus E times the area of the opposite face—with no contribution from the other four faces. The total charge enclosed in the box is σA . Equating the sum to the charge outside, we have

$$EA - EA = \sigma A \frac{\epsilon_0}{\epsilon_0},$$

from which

$$E = \frac{\sigma}{2\epsilon_0}. \quad (5.3)$$

A simple but important result.

You may remember that the same result was obtained by an earlier chapter by an integration over the entire surface. Gauss' law gives, in this manner, much more quickly (although it is not as generally applicable as the earlier method).

We emphasize that this result applies only to the field due to the charges on the sheet. If there are other charges in the neighborhood, the total field near the sheet would be the sum of (5.3) and the field of the other charges. Gauss' law would then tell us only this:

$$E + E_s = \frac{\sigma}{\epsilon_0}, \quad (5.4)$$

where E_s and E are the fields directed outward on each side of the sheet.

The problem of two parallel sheets with equal and opposite charge densities, $+\sigma$ and $-\sigma$, is equally simple if we assume again that the distance between them is quite small. Either by superposing two solutions for a single sheet or by constructing a problem like that involving two sheets, it is easily seen that the field is zero outside of the two sheets (Fig. 5-7a). By considering a box that includes only one surface of the sheet, as in (b) or (c) of the figure, it can be seen that the field between the sheets must be twice what it is for a single sheet. The result is

$$E_{\text{between}}(\text{two sheets}) = \sigma/\epsilon_0. \quad (5.5)$$

$$E_{\text{outside}}(\text{two sheets}) = 0. \quad (5.6)$$

5-7 A sphere of charge: a spherical shell

We have already (in Chapter 4) used Gauss' law to find the field outside a uniformly charged spherical region. The same method can also give us the field at points inside the sphere. For example, the computation can be used to obtain a good approximation to the field inside an atomic nucleus. In spite of the fact that the protons in a nucleus repel each other, they are, because of the strong nuclear force, distributed fairly uniformly throughout the body of the nucleus.

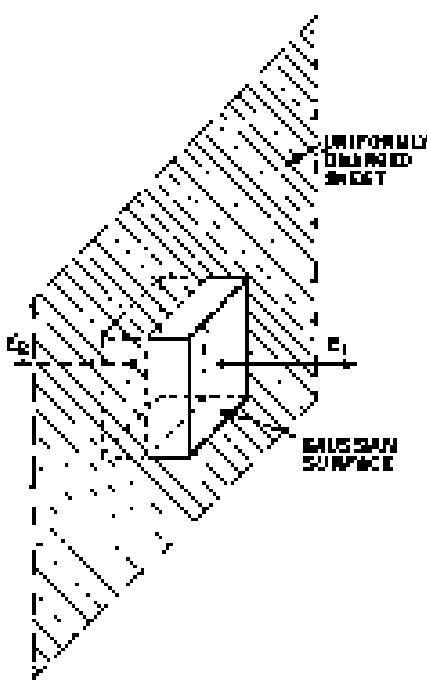


Fig. 5-6 The electric field near a uniformly charged sheet can be found by applying Gauss' law to an imaginary box.

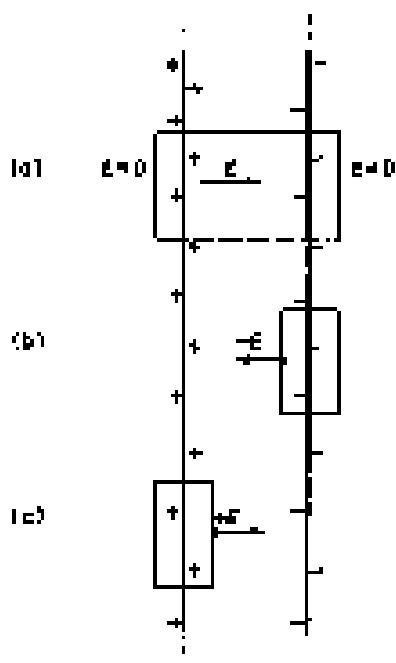


Fig. 5-7 The field between two charged sheets is σ/ϵ_0 .

Suppose that we have a sphere of radius R that uniformly contains charge. Let ρ be the charge per unit volume. Again, using arguments of symmetry, we assume the field to be radial and equal in magnitude at all points at the same distance from the center. To find the field at the distance r from the center, we take a spherical Gaussian surface of radius r ($r < R$), as shown in Fig. 5-8. The flux out of this surface is

$$4\pi r^2 E.$$

The charge inside our Gaussian surface is not volume: it's density ρ , or

$$4\pi r^3 \rho.$$

Using Gauss' law, it follows that the magnitude of the field is given by

$$E = \frac{\rho r}{\epsilon_0} \quad (r < R) \quad (5-7)$$

You can see that the formula gives the proper result for $r = R$. The electric field is proportional to the radius and is directed radially outward.

The arguments we have just given for a uniformly charged sphere can be applied also to a non-spherical shell of charge. Assuming that the shell is everywhere radially symmetric, one goes immediately from Gauss' law that the field outside the shell is zero. (A question is how inside the shell will there be no charge.)

5-8. Is the field of a point charge exactly $1/r^2$?

If we look in a little more detail at how the field inside the shell gets to be zero, we can see more clearly why it is that Gauss' law is true only because the coulomb force depends exactly on the inverse of the distance. Consider any point P inside a uniform spherical shell of charge. Suppose a small rectangular area $d\sigma_1$ and $d\sigma_2$ extends to the surface of the sphere, where it cuts out a small surface area $d\sigma_3$, as in Fig. 5-9. An equally symmetric cone discharging from the opposite side of the shell cuts off the surface area $d\sigma_4$. If the distance from P to each of the areas $d\sigma_1$ is r_1 , and r_2 , the areas are in the ratio

$$\frac{d\sigma_2}{d\sigma_1} = \frac{r_1^2}{r_2^2}.$$

(You can show this by geometry for any point P inside the shell.)

If the surface of the sphere is uniformly charging, the charging on each of the elements of area is proportional to the area, so

$$\frac{d\sigma_4}{d\sigma_1} = \frac{d\sigma_3}{d\sigma_1},$$

Coulomb's law then says that the magnitudes of the fields produced at P by these two surface elements are in the ratio

$$\frac{E_2}{E_1} = \frac{d\sigma_2/d\sigma_1}{r_1^2/r_2^2} = 1.$$

The fields cancel exactly. Since all parts of the surface can be paired off in the same way, the total field at P is zero. But you can see that it would not be so if the exponent of r in Coulomb's law were anything other than 2.

The validity of Gauss' law depends upon the inverse-square law of Coulomb. If the force law were not exactly the inverse square, it would not be true that the field inside a uniformly charged sphere would be zero by zero. For instance, if the force varied more steadily, like, say, the inverse cube of r , the portion of the surface which is nearer to an interior point would produce a field which is larger than that which is further away, resulting in a radial inward field for a positive surface

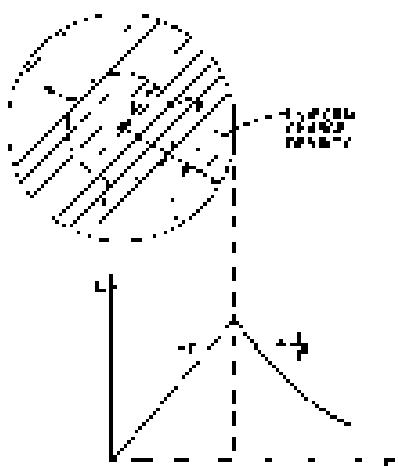


Fig. 5-8. Gauss' law can be used to find the field inside a uniformly charged sphere.

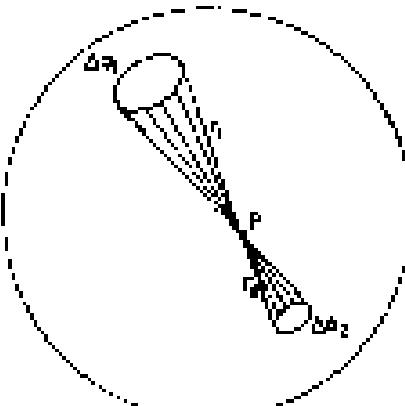


Fig. 5-9. The field is zero at any point P inside a spherical shell of charge.

charge. These conclusions suggest an elegant way of finding out whether the inverse square law is probably correct. We need only determine whether or not the field inside of a uniformly charged spherical shell is practically zero.

It is likely that such a method exists. It is usually difficult to measure a physical quantity to high precision - a one percent result may not be too difficult, but how would one go about measuring say, Coulomb's law to an accuracy of one part in a billion? It is almost certainly not possible with the best available techniques to measure the force between two charged objects with such an accuracy. But by determining only that the electric fields inside a charged sphere are smaller than some value we can make a highly accurate measurement of the correctness of Gauss' law, and hence of the inverse square dependence of Coulomb's law. What one does, in effect, is compare the force law in an ideal inverse-square system. Such comparisons of things that are equal, or nearly so, are usually the basis of the most precise physical measurements.

How shall we measure the field inside a charged sphere? One way is to try to charge an object by touching it to the inside of a spherical conductor. You know that if you touch a small metal ball to a charged object and then touch it to an electron gun the meter will become charged and the pointer will move from zero (Fig. 3-10a). The ball picks up charge because there are electric fields outside the charged sphere that cause charges to accumulate on the little ball. If you do the same experiment by touching the little ball to the inside of the charged sphere, you find that no charge is caused to the electron gun. With such an experiment you can easily show that the field inside is, at most, a few percent of the field outside, and that Gauss' law is at least approximately correct.

It appears that Huguenin's first slip was the due to neglect that the field inside a conducting shell is zero. The result seemed strange to him. When he reported his observation to Priestley, the latter suggested that it might be connected with an inverse square law, since it was known that a spherical shell of matter produced no gravitational field inside. But Coulomb didn't measure the inverse square dependence on it 'till years later, and Gauss' law came even later still.

Gauss' law has been checked carefully by putting an electron gun inside a large sphere and observing whether any deflections occur when the sphere is charged to a high voltage. A null result is always observed. Knowing the accuracy of the experimenter and the sensitivity of the meter, it is possible to compute the smallest field that would be observed. From this number it is possible to place an upper limit on the deviation of the exponent from two. If we write that the electrostatic force depends on $r^{-\alpha}$, we can place an upper bound on α . By this method Maxwell generalized that α was less than 1/10,000. The experiment was repeated and improved upon in 1936 by Plimpton and Laughlin. They found that Coulomb's exponent differs from two by less than one part in a billion.

Now that brings up an interesting question: How accurate do we know this Coulomb's law to be in various circumstances? The experiments we just described measure the dependence of the field on distance for clusters of some tens of millimeters. But what about the distances inside an atom - in the hydrogen atom, for instance, where we believe the electron is attracted to the nucleus by the same inverse square law? It is true that quantum mechanics must be used for the mechanical part of the behavior of the electron, but the force is the usual electrostatic one. In the simplification of the problem, the potential energy of an electron must be known as a function of distance from the nucleus, and Coulomb's law gives a potential whose values decrease with the first power of the distance. How accurately is the exponent known in such small distances? As a result of very careful measurements in 1911 by Faraday and Rutherford on the relative positions of the energy levels of hydrogen, we know that the exponent is correct up to one part in a billion in the atomic scale - that is, at distances of the order of one angstrom (10^{-8} centimeter).

The accuracy of the Rutherford measurements was possible again because of a physical "accident." Two of the states of a hydrogen atom are separated by near-almost identical energies only if the potential varies exactly as $1/r$. A measurement was made of the very slight difference in energies by finding

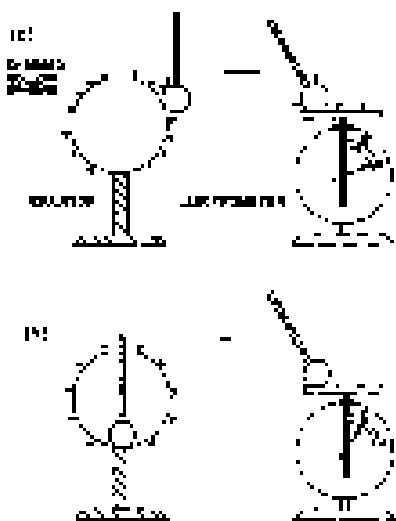


Fig. 3-10. The electric field is zero inside a closed conducting shell.

the frequency ν of the photons that are emitted or absorbed in the transition from one state to the other, using for the energy difference $\Delta E = h\nu$. Compton showed that a 2-wavelength had been impossibly different from what was observed if the exponent in the force law $1/r^2$ differed by as much as one part in a billion.

Is the same exponent correct at still greater distances? From measurements in stellar physics, it is found that there are electrostatic forces in atoms, nuclear distances about 10^{-15} centimeter—and that they fall very approximately as the inverse square. We can look at some of the evidence in a later chapter. Coulomb's law is, we know, still valid at least to some extent, in a range of the order of 10^{-12} centimeter.

How about 10^{-11} centimeters? This question can be investigated by bombarding protons with very energetic electrons and observing how they are scattered. Results indicate again an inverse law for the law of these distances. The electrical force seems to be about 10 times too weak at distances less than 10^{-11} centimeter. Now there are two possible explanations. One is that the Coulomb law does not work at such small distances; the proton is too big, perhaps, two electrons and two protons, and not point charges. Perhaps either the electron or proton, or both, is not a kind of a atom. Most physicists prefer, in fact, that the charge of the proton is smeared. We know that protons interact strongly with mesons. This implies that a proton will, from time to time, exist as a neutron with a π^+ meson around it. Such a configuration would last, on the average, about 10⁻²³ seconds of nuclear charge. We know from reasoning from a sphere of charge that $1/r^2$ all the way into the center. It is quite likely that the proton charge is smeared, but the charge of proton is at 10⁻¹¹ quite incomplete, so it may even be that Coulomb's law fails at very small distances. This question is still open.

One more point: The inverse square law is valid at distances like one meter and also at 10^{-10} m. But is the coefficient 1/4 $\pi\epsilon_0$ the same? The answer is you've had an answer of 1/4 per cent in my book.

We go back now to an important result that we obtained when we spoke of the experimental verification of Gauss' law. You may have wondered how the experiments of Maxwell or of Plimpton and Langford could give such an accuracy unless the spherical conductor they used was a perfect sphere. The accuracy of one part in a billion is really something to believe, and you might well ask what a tiny little tiny sphere, when was not perfect. There are going to be slight irregularities in any real sphere and if there are irregularities, will they not produce field inside? We wish to prove now that it is not necessary to have a perfect sphere. It is possible, in fact, to have then, is no field inside a closed wire coming shell of any shape. In other words, the experiments depended on $1/r^2$, but and nothing to do with the surface being a sphere (unless it is a sphere it is easier to calculate what the fields would be if the earth had been wrong), so we take up the stronger case. To show this, it is necessary to know some of the properties of electrical conductors.

3.9 The fields of a conductor

An electrical conductor is a solid, but containing many "free" electrons. The electrons can move around freely in the metal, but cannot leave the surface. In a metal there are so many free electrons that if any field will set large numbers of them in motion. First, the current of electrons will end up must be constantly set moving by external sources of energy, as the motion of the electrons will cause us they discharge the source producing the initial field. In "electronics" since then, we do not consider any finite amount of current (they will be considered later when we study magnetostatics) as electrons move only until they have arranged themselves to conduct an electric field everywhere inside the metal's body. (This usually happens in a very fraction of a second.) If now were say held left, the field would urge all more electrons to move, the only electric solution is that the field is everywhere zero inside.

Now consider the answer of a charged conducting object. (By "conductor" we mean in the metal itself). Since the metal is a conductor, no electric field can:

between, and on the gradient of the potential ϕ between. That means that E does not vary from point to point. Every conductor is an equipotential surface, and its surface is an equipotential surface. Since it is a conducting material the electric field is everywhere zero, the divergence of E is zero, and by Gauss' law the charge density in the interior of the conductor must be zero.

If there can be no charges in a conductor, how can it ever be charged? What do we mean when we say a conductor is "charged"? Where are the charges? The answer is that they reside on the surface of the conductor, where there are strong forces to keep them from leaving—they are not completely "free." When we study solid-state physics, we shall find that the excess charge of any conductor is on the average within one or two atomic layers of the surface. For our present purposes, it is accurate enough to say that if any charges per unit, or σ , a conductor it all accumulates on the surface, there is no charge in the interior of a conductor.

We note also that the electric field just outside the surface of a conductor must be normal to the surface. There can be no tangential component. If there were a tangential component, the electrons would move along the surface; there is no force perpendicular. Moving in another way, we know that the electric field lines must always go at right angles to an equipotential surface.

We can then, using Gauss' law, relate the field strength just outside a conductor to the density of the charge at the surface. For a Gaussian surface, we take a small cylindrical one half inside and half outside the surface, like the one shown in Fig. 5-11. There is a contribution to the total flux of E only from the side of the box toward the conductor. The field just outside the surface of a conductor is then

Outside a conductor:

$$E = \frac{\sigma}{\epsilon_0}, \quad (5.8)$$

where σ is the local surface charge density.

Why does a sheet of charge on a conductor produce a different field than a sheet of charge? In other words, why is (5.8) twice as large as (5.7)? The reason, of course, is that we have assumed that there are no "other" charges around. There must, in fact, be some to make $E = 0$ in the conductor. The charges in the immediate neighborhood of P will, from the surface in, in fact give a field $E_{ext} = \sigma/\epsilon_0$, and both fields sum outside the surface. But all the rest of the charge on the conductor "cooperates" to produce an additional field at the point P equal in magnitude to E_{ext} . The total field outside goes to zero and the total outside is $2E_{ext} = \sigma/\epsilon_0$.

5-10 The field in a cavity of a conductor

We come now to the problem of the hollow conductor—a conductor with a cavity. There is no field in the cavity, but what about in the cavity? We shall show that if the cavity is empty then there are no fields in it, no matter what the shape of the envelope of the cavity—say for this one in Fig. 5-12. Consider a gaussian surface, like S in Fig. 5-12, that encloses the cavity but stays everywhere in the conducting material. Everywhere on S the field is zero, and there is no flux through S since the total charge inside S is zero. For a spherical shell, one could then argue from symmetry that there could be no charge inside. But, in general, we can only say that there are equal amounts of positive and negative charge on the inner surface of the conductor π . There could be a positive surface charge on one part and a negative one somewhere else, as illustrated in Fig. 5-13. Such a thing cannot be ruled out by Gauss' law.

What really happens, of course, is that any equal and opposite charges on the inner surface would slide around to meet each other, cancelling out completely. We can show that they must cancel completely by using the fact that the circulation of E is always zero (conservation). Suppose there were charges on some parts of the inner surface. We know that there would have to be an equal number of opposite charges somewhere else. Now any line of E would have an end on the

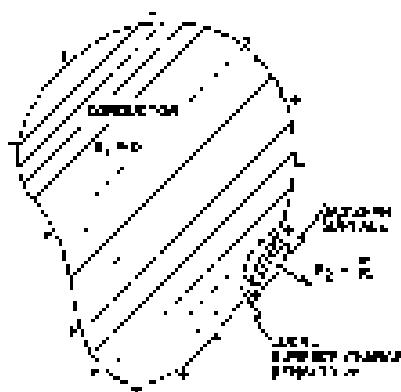


Fig. 5-11. The electric field just outside the surface of a conductor is proportional to the local surface density of charge.

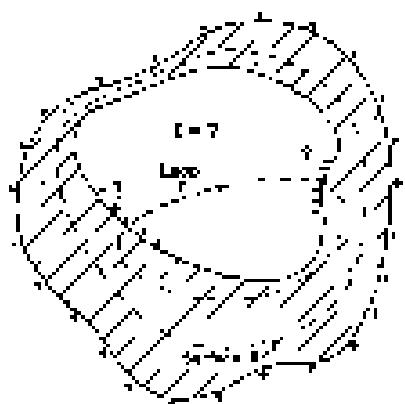


Fig. 5-12. What is the field in an empty cavity of a conductor, for any shape?

positive charge and end on the negative charge (since we are considering only the electric field, not free charges in the cavity). Now integrate a loop Γ that crosses the cavity along a line of force from some positive charge to some negative charge, and returns to its starting point via the conductor (as in Fig. 5-12). The integral along such a line of force from the positive to the negative charge would not be zero. The integral through the metal is zero, since $E = 0$. So we would have

$$\oint \mathbf{E} \cdot d\mathbf{s} \neq 0??$$

But the line integral of \mathbf{E} around any closed loop in an electric field is always zero. So there must be no fields inside the empty cavity, nor any charges on the inside surface.

You should notice carefully one important qualification we have made. We have always said "with an empty cavity." If there charged particles or some other conductors in the cavity, as in an insulator or in a metal conductor insulated from the outside, then there can be fields in the cavity. But then that is not an "empty" cavity.

We have shown that if a cavity is completely enclosed by a conductor, no static distribution of charges outside can ever produce any fields inside. This explains the principle of "shielding" electrical equipment by placing it in a metal can. (See some open cans can be used to show that no static distribution of charges inside a closed conductor can produce any fields outside. Shielding works both ways! In electrostatics—but not in varying fields—the fields on the two sides of a closed conducting shell are completely independent.)

Now you see why it was possible to check Coulomb's law to within a percent precision. The shape of the hollow shell need doesn't matter. It doesn't need to be spherical; it could be square. If Gauss' law is good, the field inside is always zero. Now you also understand why it is safe to sit inside the high voltage terminal of a million-volt Van de Graaff generator without worrying about getting a shock—because of Gauss' law.

The Electric Field in Various Configurations

6.1 Equations of the electrostatic potential

This chapter will describe the behavior of the electric field in a number of different configurations. It will provide some experience with solving the electric field behavior, and will develop some of the other related methods we shall use to study field distributions.

We begin by giving you the whole mathematical framework that describes the structure, the Maxwell equations for electrostatics:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}, \quad (6.1)$$

$$\nabla \times \mathbf{E} = 0. \quad (6.2)$$

In fact, we can combine these two equations. From the second equation, we know at once that we can describe the field as the gradient of a scalar field known as ϕ :

$$\mathbf{E} = -\nabla \phi. \quad (6.3)$$

We may, if you wish, completely bypass any geometric drawing in terms of the potential ϕ . We note, the differential relation that ϕ must obey is still satisfying Eq. (6.2), since

$$\nabla \cdot \mathbf{E} = -\frac{\partial}{\partial r}. \quad (6.4)$$

The divergence of the gradient of ϕ is the same as the divergence of \mathbf{E} :

$$\nabla \cdot \mathbf{E} = \nabla^2 \phi = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} \quad (6.5)$$

and so with Eq. (6.4) as

$$\nabla^2 \phi = -\frac{\rho}{\epsilon_0}. \quad (6.6)$$

The term $\nabla^2 \phi$ is called the *Laplacian*, and Eq. (6.6) is called the *Poisson* or *3D Poisson*. The entire subject of electrostatics, in this mathematical framework, is really a study of the solutions of Eq. (6.6) plus Eq. (6.1). Once ϕ is obtained by solving Eq. (6.6), we can find \mathbf{E} from Eq. (6.3).

We focus on the one special class of problem in which ϕ is given as a function of r , θ , ϕ , in that case the problem is almost trivial. If we already know the sinusoidal ϕ (6.6) for the given ρ , we have shown that \mathbf{E} is known at every point the scientific al point! This

$$\mathbf{E}(r) = \begin{cases} \frac{\partial \phi(r,\theta,\phi)}{\partial r} & \text{if } \rho \neq 0 \\ 0 & \text{if } \rho = 0 \end{cases} \quad (6.7)$$

where $\rho(r)$ is the charge density, $\phi(r,\theta,\phi)$ the relevant function of position (r,θ,ϕ) and \mathbf{E} is the field at between points (2) and (3). The solution of the differential equation (6.6) is nontrivial in many cases (see Chap. 7). The solution of (6.7) should be relatively trivial, provided you are many situations in physics that lead to equations like

$$\frac{d^2 \phi}{dr^2} = -\frac{\rho(r)}{\epsilon_0}, \quad (\text{something}) = \text{something else},$$

and Eq. (6.7) is a solution of the equation for any of these problems.

The solution of the next 10 problems is the majority of your assignment, and after the solutions of all the charges are known, the 3rd section of week 10 is a few examples.

- 6.1 Equations of the electrostatic potential
- 6.2 The diatomic dipole
- 6.3 Remarks on vector equations
- 6.4 The dipole potential as a gradient
- 6.5 The dipole approximation for an arbitrary distribution
- 6.6 The fields of charged conductors
- 6.7 The equation of Gauss
- 6.8 A point charge near a conducting plane
- 6.9 A point charge near a conducting sphere
- 6.10 Conductors: parallel plates
- 6.11 High-voltage breakdown
- 6.12 The field emission microscope

Review Chapter 6. See J. R. Wilson

6-1 The electric dipole

First, take two point charges, $+q$ and $-q$, separated by the distance d . Let the z -axis go through the charges, and pick the origin halfway between, as shown in Fig. 6-1. Then, using (4.24), the potential from the two charges is given by

$$\phi(x, y, z)$$

$$= \frac{1}{4\pi\epsilon_0} \left[\frac{q}{\sqrt{[z - (d/2)]^2 + x^2 + y^2}} + \frac{q}{\sqrt{[z + (d/2)]^2 + x^2 + y^2}} \right]. \quad (6.8)$$

We are not going to write out the formula for the electric field, but we can always calculate it once we have the potential. So we have solved the problem of two charges.

There is an important special case in which the two charges are very close together—which is to say that we are interested in the fields only at distances from the charges large in comparison with their separation. We call such a close pair of charges a *dipole*. Dipoles are very common.

A “dipole” antenna can often be approximated by two charges separated by a small distance if we don’t look at the field too close to the antenna. (We are usually interested in antennas with moving charges; then the equations of statics do not really apply, but for some purposes they are an adequate approximation.)

More important perhaps, are atomic dipoles. If there is an electric field in any material, the electrons and protons feel opposite forces and are displaced relative to each other. In a conductor, you remember, some of the electrons move to the surfaces, so that the field inside becomes zero. In an insulator the electrons cannot move very far; they are pulled back by the attraction of the nucleus. They do, however, shift a little bit. So although an atom, or molecule, remains neutral in an external electric field, there is a very tiny separation of its positive and negative charges and it becomes a microscopic dipole. If we are interested in the fields of these atomic dipoles in the neighborhood of ordinary-sized objects, we are normally dealing with distances large compared with the separations of the pairs of charges.

In some molecules the charges are somewhat separated even in the absence of external fields, because of the form of the molecule. In a water molecule, for example, there is a net negative charge on the oxygen atom and a net positive charge on each of the two hydrogen atoms, which are not placed symmetrically but as in Fig. 6-2. Although the charge of the whole molecule is zero, there is a charge distribution with a little more negative charge to one side and a little more positive charge on the other. This arrangement is certainly not as simple as two point charges, but when you look far away the system acts like a dipole. As we shall see a little later, the field at large distances is not sensitive to the fine details.

Let’s look then at the field of two opposite charges with a small separation d . If d becomes zero, the two charges are on top of each other, the two potentials cancel, and there is no field. But if they are not exactly on top of each other, we can get a good approximation to the potential by expanding the term ϕ in (6.8) in a power series in the small quantity d/r (using the binomial expansion). Keeping terms only to first order in d , we can write

$$\left(z - \frac{d}{2}\right)^2 \approx z^2 - zd,$$

It is convenient to write

$$x^2 + y^2 + z^2 = r^2.$$

Then

$$\left(z - \frac{d}{2}\right)^2 + x^2 + y^2 \approx r^2 - zd - r^2 \left(1 - \frac{zd}{r^2}\right).$$

and

$$\frac{1}{\sqrt{[z - (d/2)]^2 + x^2 + y^2}} \approx \frac{1}{\sqrt{r^2(1 - (zd/r^2))}} \approx \frac{1}{r} \left(1 - \frac{zd}{r^2}\right)^{-1/2}.$$

Using the binomial expansion again for $(1 - (zd/r^2))^{-1/2}$ and throwing away terms with higher powers than the square of d , we get

$$\frac{1}{r} \left(1 + \frac{1}{2} \frac{zd}{r^2} \right).$$

Similarly,

$$\frac{1}{\sqrt{(z + (d/2))^2 + r^2}} = \frac{1}{r} \left(1 - \frac{1}{2} \frac{zd}{r^2} \right).$$

The difference of these two terms gives for the potential

$$\phi(x, y, z) = \frac{1}{4\pi\epsilon_0} \frac{p}{r} \cos\theta. \quad (6.9)$$

The potential, and hence the field, which is its derivative, is proportional to qd , the product of the charge and the separation. This product is defined as the **dipole moment** of the two charges, for which we will use the symbol p (do not confuse with reaction/cm³):

$$p = qd. \quad (6.10)$$

Equation (6.9) can also be written as

$$\phi(x, y, z) = \frac{1}{4\pi\epsilon_0} \frac{p \cos\theta}{r}, \quad (6.11)$$

where $z/r = \cos\theta$, where θ is the angle between the axis of the dipole and the radius vector to the point (x, y, z) —see Fig. 6-1. The potential of a dipole decreases as $1/r^2$ for a given direction from the axis (which is for a point charge it goes as $1/r$). The electric field E of the dipole will then decrease as $1/r^3$.

We can put our formula into a vector form if we define p as a vector whose magnitude is p and whose direction is along the axis of the dipole, pointing from q_+ toward q_- . Then

$$\cos\theta = p \cdot e_r \quad (6.12)$$

where e_r is the unit radial vector (Fig. 6-2). We can also represent the point (x, y, z) by r . Then

$$\text{Dipole potential: } \phi(r) = \frac{1}{4\pi\epsilon_0} \frac{p \cdot e_r}{r^2} = \frac{1}{4\pi\epsilon_0} \frac{p \cdot r}{r^3}. \quad (6.13)$$

This formula is valid for a dipole with any orientation and position if r represents the vector from the dipole to the point of interest.

If we want the electric field of the dipole we can get it by taking the gradient of ϕ . For example, the component of the field is $-\partial\phi/\partial z$. For a dipole located along the z -axis we can use (6.9):

$$\begin{aligned} &= \frac{\partial\phi}{\partial z} = -\frac{2}{4\pi\epsilon_0} \frac{\partial}{\partial z} \left(\frac{z}{r^2} \right) = -\frac{1}{4\pi\epsilon_0} \frac{p \cdot 3z^2}{r^5}, \\ \text{or} \quad E_z &= \frac{p}{4\pi\epsilon_0} \frac{3 \cos^2\theta - 1}{r^5}. \end{aligned} \quad (6.14)$$

The x - and y -components are

$$E_x = \frac{p}{4\pi\epsilon_0} \frac{3xz}{r^5}, \quad E_y = \frac{p}{4\pi\epsilon_0} \frac{3yz}{r^5}.$$

These two can be combined to give one component directed perpendicular to the z -axis, which we will call the transverse component E_t :

$$E_t = \sqrt{E_x^2 + E_y^2} = \frac{p}{4\pi\epsilon_0} \frac{3x}{r^5} \sqrt{x^2 + y^2}$$

$$\text{or} \quad E_t = \frac{p}{4\pi\epsilon_0} \frac{3 \cos\theta \sin\theta}{r^5}. \quad (6.15)$$

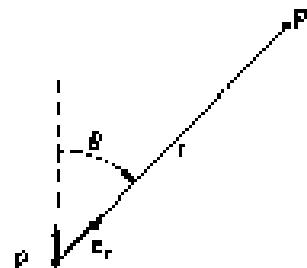


Fig. 6-2. Vector notation for a dipole.

The transverse component B_z is in the x - y plane and points directly away from the axis of the dipole. The total field, of course, is

$$\mathbf{B} = \sqrt{B_x^2 + B_z^2}.$$

The dipole field varies inversely as the cube of the distance from the dipole. On the axis, at $\theta = 0$, it is twice as strong as at $\theta = 90^\circ$. At each of these special angles the electric field has only a component, but of opposite sign at the two places (Fig. 6.4).

6-3 Remarks on vector equations

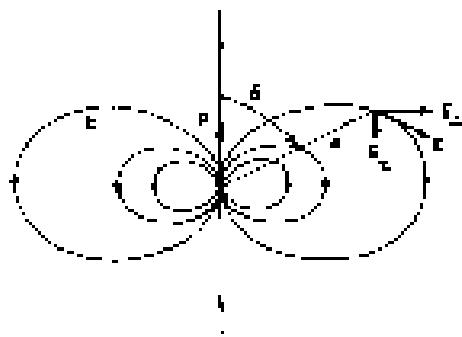


Fig. 6.4. The electric field of a dipole.

This is a good place to make a general remark about vector analysis. The fundamental parts can be expressed by scalar equations in a general form. But in making various calculations and analyses, it is always a good idea to choose the axes in some convenient way. Notice that when we were finding the potential of a dipole we chose the z -axis along the direction of the dipole, rather than at some arbitrary angle. This made the work much easier. But then we wrote the equation in vector form so that this would no longer depend on any particular coordinate system. After that, we were allowed to choose any coordinate system we wished, knowing that the relationships would still hold. It clearly does not make any sense to begin with an arbitrary coordinate system, if you are going to end up where you can change a new system for the particular problem—provided that the result can finally be expressed as a vector equation. So before you do take advantage of the fact that vector equations are independent of any coordinate system.

On the other hand, if you are trying to calculate the divergence of a vector, instead of just looking at $\nabla \cdot \mathbf{E}$ and wondering what it is, don't forget that it can always be expressed as

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}.$$

If you can then work out the x , y , and z -components of the electric field and differentiate them, you will have the divergence. There often seems to be a feeling that there is something inelegant—some kind of deceit involved—in writing out the components; that somehow there ought always to be a way to do everything with the vector operators. There is often no advantage to it. The first time we encounter a particular kind of problem, it usually helps to write out the components to be sure we understand what is meant. There is nothing inelegant about putting numbers into equations, and nothing inelegant about substituting the derivatives for the fancy symbols. In fact, there is often a certain elegance in doing just that. Of course when you publish a paper in a professional journal it will look better—and be more easily understood—if you can write everything in vector form. Besides, it saves space.

6-4 The dipole potential as a gradient

We would like to point out a rather amazing thing about the dipole formula, Eq. (6.11). The potential can also be written as

$$\phi = -\frac{1}{4\pi\epsilon_0} \mathbf{p} \cdot \mathbf{r} \left(\frac{1}{r} \right), \quad (6.15)$$

If you calculate the gradient of $1/r$, you get

$$\nabla \left(\frac{1}{r} \right) = -\frac{x}{r^3} \mathbf{i} - \frac{y}{r^3} \mathbf{j} - \frac{z}{r^3} \mathbf{k},$$

and Eq. (6.16) is the same as Eq. (6.13).

How did we think of that? We just remembered that a/r^2 appeared in the formula for the field of a point charge, and that the field was the gradient of a potential which has a $1/r$ dependence.

There is a physical reason for being able to write the dipole potential in the form of Eq. (6.16). Suppose we have a point charge q at the origin. The potential at the point P at (x, y, z) is

$$\phi_0 = \frac{q}{r}.$$

(Let's leave off the $1/4\pi\epsilon_0$ which we make later; we can stick it in or not.) Now if we move the charge $+q$ up a distance Δz , the potential at P will change a little, by, say, $\Delta\phi_q$. How much is $\Delta\phi_q$? Well, it is just the amount that the potential would change if we were to leave the charge at the origin and move P downwards by the same distance Δz (Fig. 6-2). That is,

$$\Delta\phi_q = -\frac{\partial \phi_0}{\partial z} \Delta z,$$

where by Δz we mean the same as $d/2$. So, using $\phi = q/r$, we have that the potential from the positive charge is

$$\phi_+ = \frac{q}{r} - \frac{\partial}{\partial z} \left(\frac{q}{r} \right) \frac{d}{2}. \quad (6.17)$$

Applying the same reasoning for the potential from the negative charge, we can write

$$\phi_- = -\frac{-q}{r} + \frac{\partial}{\partial z} \left(\frac{-q}{r} \right) \frac{d}{2}. \quad (6.18)$$

The total potential is the sum of (6.17) and (6.18):

$$\begin{aligned} \phi &= \phi_+ + \phi_- = -\frac{\partial}{\partial z} \left(\frac{q}{r} \right) d \\ &= -\frac{\partial}{\partial z} \left(\frac{1}{r} \right) qd. \end{aligned} \quad (6.19)$$

For other orientation of the dipole, we could represent the displacement of the pos. charge by the vector Δr_+ . We should then write Eq. (6.17) as

$$d\phi = -\nabla\phi_0 \cdot d\mathbf{r}_+,$$

where $d\mathbf{r}$ is then to be replaced by $d/2$. Simplifying the derivation as before, Eq. (6.19) would then become

$$\phi = -\nabla \left(\frac{1}{r} \right) \cdot qd.$$

This is the same as Eq. (6.16), if we replace $qd \rightarrow p$, and put back the $1/4\pi\epsilon_0$. Looking at it another way, we see that the dipole potential, Eq. (6.13), can be interpreted as

$$\phi = -p \cdot \nabla \Phi_0, \quad (6.20)$$

where $\Phi_0 = 1/4\pi\epsilon_0 r$ is the potential of a unit point charge.

Although we can always find the potential of a known charge distribution by an integration, it is sometimes possible to save time by getting the answer with a clever trick. For example, one can often make use of the superposition principle. If we are given a charge distribution that can be made up of the sum of two distributions for which the potentials are already known, it is easy to find the desired potential by just adding the two known ones. One example of this is our derivation of (6.20), another is the following.

Suppose we have a spherical surface with a distribution of surface charge that varies as the cosine of the polar angle. The integration for this distribution is fairly messy. But, surprisingly, such a distribution can be analyzed by superposition. For imagine a sphere with a uniform volume density of positive charge, and another sphere with an equal uniform volume density of negative charge,

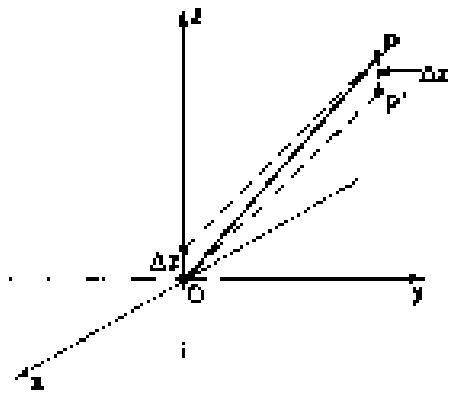


Fig. 6-2. The potential at P from a point charge at Δz above the origin is the same as the potential at $P'(\Delta z$ below $P)$ from the same charge at the origin

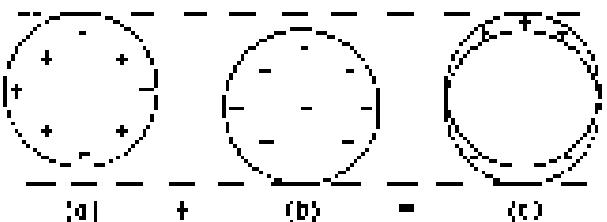


Fig. 6-3. Two uniformly charged spheres, superposed with a slight displacement, are equivalent to a Gaussian distribution of surface charge.

originally supposed to induce a negative test z , charge, sphere. If the left $+z$ sphere is displaced slightly with respect to the negative sphere, the tests of the elongated sphere would remain neutral, but a little positive dipole will appear on one side, and some negative charge will appear on the opposite side, as illustrated in Fig. 6-3. If the relative displacement d of the two spheres is small, the net charge is equivalent to a Gaussian charge (or a spherical surface), and the test charge density ρ_d is proportional to the cosine of the polar angle.

Now we want the potential from the distribution, we do not need to do an integral. We know that the potential from each of the spheres of charge q is for points outside the sphere the same as from a point charge. The two z -aligned spheres are like two point charges. The potential is just that of two z 's.

In this very simple case, but a charge distribution more general than a sphere, we can repeat this procedure.

$$\phi = \phi_{\text{point}}$$

where $\phi_{\text{point}} = kq/r$. This charge q is just that of a single sphere moment is

$$q = \frac{4\pi r^2}{3} \rho_d$$

It can also be shown that inside the sphere the field is constant with the value

$$E = \frac{2k}{3r_0}$$

In fact the angle from the positive z -axis, the electric field inside the sphere is in the original coordinate system. The example we have just considered is not as artificial as may appear; we will encounter it again in the theory of dielectrics.

6-6. The dipole approximation for an arbitrary distribution

The electric field appears in various dielectrics both in vacuum and in exterior. Suppose that we have an object that has a complicated distribution of charge like the water molecule (Fig. 6-2); and we are interested in it in the Coulomb way. Now it's not easy to get a fully satisfactory complete expression for the fields which is appropriate for distances large compared with the size a to be used.

We can think of a charge in symmetry to your charge q at a considerable distance, as shown in Fig. 6-7. (We can, if you like, replace q by $-q$; if you wish.) The two charges, let's call them q and q' , are separated by a distance d . From among q , q' , we can neglect

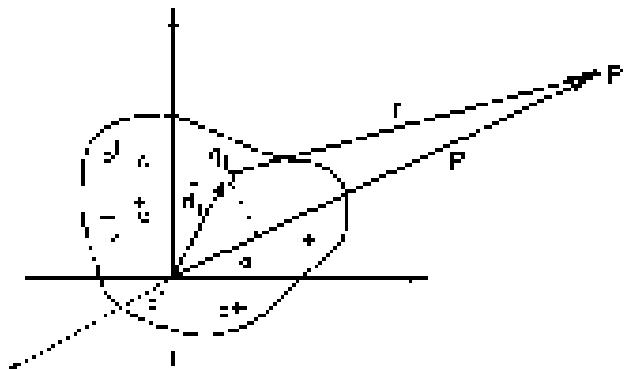


Fig. 6-7. Computation of the potential of a point P at a large distance from a set of charges.

In the middle of the pump of charges, ϕ_0 at is the potential at height R , but not at R , where it is much larger than the maximum ϕ_0^* . The potential from the whole distribution is given by

$$\phi = \frac{1}{4\pi\epsilon_0} \sum_i \frac{q_i}{r_i}, \quad (6.21)$$

where r_i is the distance from P to the charge q_i (the length of the vector $R - \vec{r}_i$). Now if the distance from the charge to P , the sum of all r_i 's, is enormous, each of the r_i 's can be approximated by R ; then term becomes q_i/R , and we can take $1/R$ out as a factor in front of the summation. This gives us the simple result

$$\phi = \frac{1}{4\pi\epsilon_0 R} \sum_i q_i = \frac{Q}{4\pi\epsilon_0 R}, \quad (6.22)$$

where Q is just the total charge of the whole object. This we find that far enough from any lump of charge, the field looks like a point charge. The result is not too surprising.

But what if there are equal numbers of positive and negative charges? Then the total charge Q of the object is zero. Taking my original case; in fact as we show, objects are usually neutral. The water molecule is neutral, but the charges are not all at one point, so if we are close enough we should be able to see some effect of the opposite charges. We need a better approximation than (6.21) for the potential from an arbitrary distribution of charge. A neutral object (Section 6.2) has $dipole$ fields but we are no longer interested in \vec{P} . We must make some approximation for ϕ at any point P at a large distance, as will differ from ϕ_0 in an excellent approximation by the proportion $1/d$ on R , as can be seen from Fig. 6.7. You should begin that ϕ is really to be over the d shown in the figure. In other words, if d is the first term in the expansion of R , then our next approximation to ϕ is

$$\phi = \bar{\phi} - d/\epsilon_0 \omega, \quad (6.23)$$

which we may write as $(1/d)\phi_0$, which shows $d/\epsilon_0 \omega$ is subtracted to our approximation to ϕ .

$$\bar{\phi}_i \approx \frac{1}{R} \left(1 + \frac{d_i}{\epsilon_0 \omega} \right). \quad (6.24)$$

Substituting this in (6.21), we get for the potential is

$$\phi = \frac{1}{4\pi\epsilon_0} \left(\frac{Q}{R} + \sum_i \omega_i \frac{d_i}{\epsilon_0 \omega} + \dots \right). \quad (6.25)$$

The three terms include the terms of higher order in $d/\epsilon_0 \omega$, but we have neglected. These, as well as the ones we have already obtained, are successive terms in an expansion of ϕ about $1/R$, about $1/d$ in powers of $d/\epsilon_0 \omega$.

The first term in (6.25) is what we get before; it drops out if the object is neutral. The second term depends on $1/d\omega$, just as for a dipole, but here, d we ignore

$$P = \sum_i q_i \vec{d}_i \quad (6.26)$$

as a property of the charge distribution. The second term of the potential (6.25) is

$$\delta = \frac{1}{4\pi\epsilon_0} \frac{d}{\omega^2}. \quad (6.27)$$

peculiarly difficult to work out. The quantity p is called the *dipole moment* of the distribution. It is a geometrical notion as its definition, and reduces to d for the special case of two point charges.

Our result is true far enough away from any pieces of charges, but if the whole neutral, the potential is a *long range* potential. It decreases as $1/d^2$ and varies as $\log d$ and its complete dependence on the dipole moment p of the distribution of charge. It is for these reasons that dipole fields are important, since the simple case of point charges is quite rare.

The water is electrically neutral, for example, and it has a rather strong dipole moment. The electric fields that result from this moment are responsible for some of the important properties of water. For many molecules, for example CO_2 , the dipole moment vanishes because of the symmetry of the molecule. For them we should expand still more accurately, containing another term in the potential which decreases as $1/R^3$, $1/m$, which is called a quadrupole potential. We will discuss such cases later.

6-6 The fields of charged conductors

We have now finished with the examples we wish to cover of situations in which the charge distributions is known from the start. It has been a problem without serious complications involving at most some integration. We turn now to an entirely new kind of problem: the determination of the fields near charged conductors.

Suppose that we have a situation in which a total charge Q is placed on an arbitrary conductor. Now we will not be able to say exactly where the charges are. They will spread out in some way on the surface. How can we know how the charges have distributed themselves on the surface? They must distribute themselves so that the potential of the surface is constant. If the surface were not an equipotential, there would be an electric field inside the conductor, and the charges would keep moving until it became zero. The general problem of this kind can be solved in the following way. We guess at a distribution of charge and calculate the potential. If the potential turns out to be constant everywhere on the surface, the problem is finished. If the surface is not an equipotential, we have guessed the wrong distribution of charges, and should guess again—carefully with an improved guess. This can go on forever, unless we are judicious about the successive guesses.

The question of how to guess at the distribution is mathematically difficult. Nature, of course, has time to do it; the charges push and pull until they all balance themselves. When we try to solve the problem, however, it takes us so long to make each trial that that method is very tedious. With an arbitrary group of conductors and charges the problem can be very complicated, and in general it cannot be solved without either elaborate numerical methods. Such numerical calculations, these days, are set up on a computing machine that will do the work for us, once we don't tell it how to proceed.

On the other hand, there are a lot of little practical cases where it would be just as feasible to find the answer by some simple device. Without trying, now, determine for me a conjecture. Frankly, I have a small list of cases where the answer can be obtained by reasoning it out of Nature by some trick or other. The first trick we will describe involves making use of something we have already obtained: the situations in which charges have specified locations.

6-7 The method of images

We have solved, for example, the field of two point charges. Figure 6-9 shows some of the field lines and equipotential surfaces we obtained by the manipulations in Chapter 5. Now consider the equipotential surfaces in Fig. 6-1. Suppose we were to shape a thin sheet of metal so that it just fits this surface. If we place it right at the surface and adjust its potential to the proper value, no one would ever know it was there, because nothing would be changed.

But notice! We have really solved a new problem. We have a situation in which the surface of a curved conductor with a given potential is placed next a point charge. If the metal sheet we place at the equipotential surface eventually closes on itself, for, in practice, if it goes far enough we have the kind of situation considered in Section 5-16, in which our space is divided into two regions, one inside and one outside a closed conducting shell. We found there that the fields in the two regions are quite independent of each other. So we would see the same fields outside our curved conductor no matter what is inside. We can even fill up

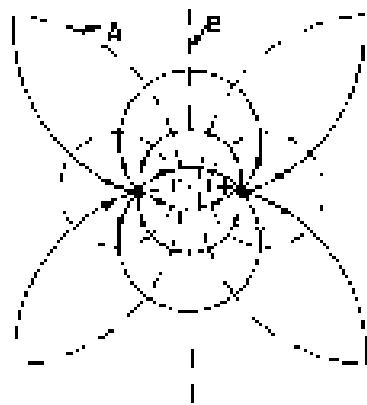


Fig. 6-9. The field lines and equipotential surfaces for two point charges

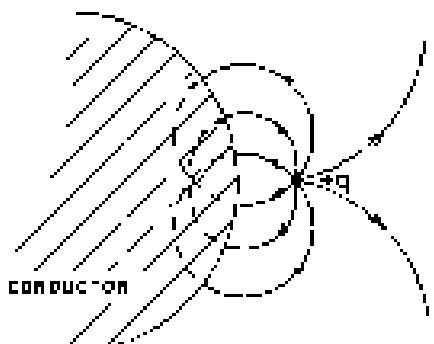


Fig. 6-4. The field outside a conductor shaped like the equipotential A of Fig. 6-9.

the whole inside with conducting material. We have found, therefore, the fields for the arrangement of Fig. 6-9. In the space outside the conductor the field is just like that of two point charges, as in Fig. 6-8. Inside the conductor, it is zero. Also—in it must be—the electric field just outside the conductor is normal to the surface.

Thus we can compute the fields in Fig. 6-9 by applying the field due to $+q$ in its ordinary point charge $-q$ at a suitable point. The point charge we "imagine" existing behind the conducting surface is called an *image charge*.

In books you can find long lists of solutions for hyperboloid-shaped conductors and more complicated looking things, and you wonder how anyone ever solved these terrible shapes. They were solved backwards! Someone solved a simple problem with given charges. He then saw that some equipotential surface showed up in a new shape, and he wrote a paper in which he pointed out that the field outside had particular properties described in a certain way.

6-8 A point charge near a conducting plane

As the simplest application of the use of this method, let's make use of the plane equipotential surface \mathcal{S} of Fig. 6-8. With it, we can solve the problem of a charge in front of a conducting sheet. We just cross out the left-hand half of the surface. The field lines for our solution are shown in Fig. 6-10. Notice that the plane sheet \mathcal{S} was halfway between the two charges, as in problem 1. We have solved the problem of a positive charge q in front of a given conducting sheet.

We have now solved for the total field, but what about the *induced charges* that are responsible for it? There are, in addition to our positive point charge, some induced negative charges on the conducting sheet that have been attracted by the positive charge (from large distances away). Now suppose that for some technical reason—or out of curiosity—you would like to know how the negative charges are distributed on the surface. You can find the surface charge density by using the result we worked out in Section 5-6 with Gauss' theorem. The normal com-

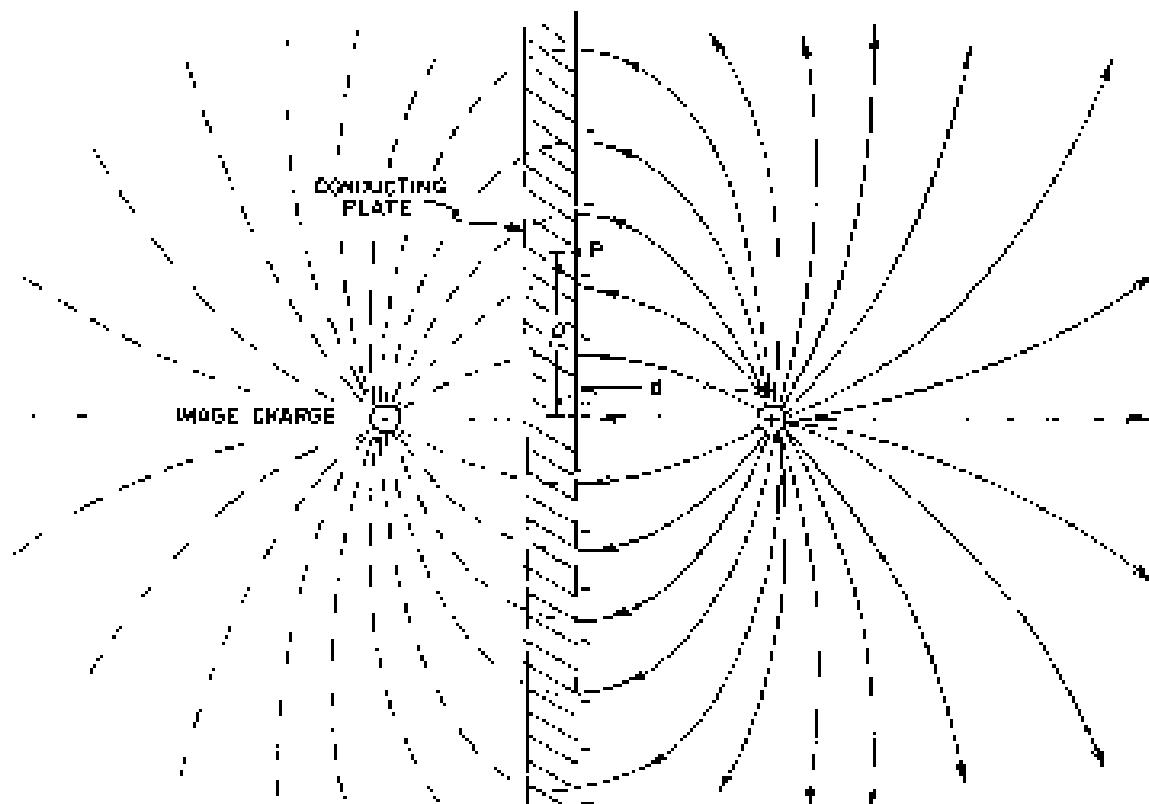


Fig. 6-10. The field of a charge near a plane conducting surface, found by the method of images.

point of the dielectric. The total surface density is equal to the density of surface charge σ , given by ϵ_0 . We can obtain the density of charge at any point on the sphere by looking forwards (from Eq. 6-10) to component of the electric field on the surface. We know that, because we know the field everywhere,

Consider a point P to be at distance r from the point charge. To reach the point charge (Eq. 6-10), the electric field at this point is normal to the surface and is directed inward. The component parallel to the surface of the field from the point charge is

$$E_r = \frac{q}{4\pi\epsilon_0 r^2} \frac{r^2}{r^2 + b^2} \quad (6-21)$$

To the left of P add the electric field produced by the negative surface charge. Then to the right the normal component (and hence all others) on the charge density σ at any point on the surface is

$$\sigma(r) = \sigma_0 \frac{b}{r} = -\frac{q}{4\pi\epsilon_0 r^2} \frac{r^2}{r^2 + b^2} \quad (6-22)$$

An interesting check on our work is to integrate over the whole surface. We find that the total induced charge is $-q$, as it should be.

One further question: Is there force on the point charge? This becomes there is an attraction from the induced negative surface charge on the plate. Now that we know what the surface charge is (from Eq. 6-22) we could compute the force on our positive point charge by integrating. Or we could know that the force acting on the positive charge is exactly the same as it would be were the negative image charge instead of the plate nearer the field. In the capillary case, the result is the same — no twin charges, but a force between the plates which is finite.

$$F = \frac{1}{4\pi\epsilon_0} \frac{q^2}{b^2 r^2} \quad (6-23)$$

We have found out how much more easily than by integrating over all the surface charges.

6-9 A point charge near a conducting sphere

What other problems deserve a little care & simple solution? The next most interesting is a sphere. Let's take the facts about a metal sphere which have charge q near it, as shown in Fig. 6-1. Now we may ask for a simple situation which gives a simple formula for calculating q exactly. If we took account of geometry people have already done, we find the somewhat surprising that the field of two equal point charges has an equipotential field as a sphere. That is we divide the field on R into q and charge $-q$ and pick the right value of R charge $-q$ must be on the upper hemisphere, free fit on sphere. Below, we're done with the following problem.

Assume that you want the equipotential surface to be a sphere of radius r with its center at a distance b from the charge q . Put a image charge of $-q$ at $y = -b$ along the y -axis from the origin to the center of the sphere, and at a distance a/b from the center. The sphere will be at zero potential.

The method used here stems from the fact that a sphere is the locus of points for whom the distances from two points are constant. Referring to Fig. 6-1, the condition at P from q and $-q$ is proportional to

$$\frac{q}{r_1} + \frac{q}{r_2}$$

The point P will thus satisfy a null point equation

$$\frac{r_1^2}{r} = -\frac{q}{r_1} \quad \text{or} \quad \frac{r_2^2}{r} = -\frac{q}{r_2}$$

If we place q' at the distance a' from the center, the ratio a'/r has the constant value a/b . Then if

$$\frac{a'}{r} = \frac{a}{b} \quad (6.31)$$

the sphere is an equipotential. Its potential is, in fact, zero.

What happens if we are interested in a sphere that is not at zero potential? That would be so only if total charge happens accidentally to be q . Of course if it is grounded, the charges induced on it would have to be just that. But what if it is insulated, and we have put no charge on it? Or if we know that the total charge Q has been put on it? Or just that it has a given potential not equal to zero? All these questions are easily answered. We can always find a point charge q'' on the center of the sphere. The sphere will remain an equipotential by superposition; only the magnitude of the potential will be changed.

If we have, for example, a conducting sphere which is initially uncharged and isolated from everything else, and we bring near to it the positive point charge q , the total charge of the sphere will remain zero. The solution is found by requiring a tiny charge q' as before, but in addition adding a charge q'' at the center of the sphere, according

$$q'' = -q' = \frac{q}{b} q. \quad (6.32)$$

The fields everywhere outside the sphere are given by the superposition of the fields of q , q' , and q'' . The problem is solved.

We can see now that there will be a force of attraction between the sphere and the point charge q . It is not zero even though there is no charge on the isolated sphere. What does the attraction come from? When you bring a positive charge up to a conducting sphere, the positive charge attracts negative charges to the side closest to itself, and these positive charges leave the surface of the body. The attraction by the positive charges causes the negative ion the positive charges, thus causing attraction. We can find out how large the attraction is by calculating the force on q by the field produced by q and q'' . The total force is the sum of the attractive force between q and q' ($= -kq_1q_2/q^2$) and the repulsive force between q and q'' ($= kq_1q_2/q^2$) of the distances b .

Those who were interested in followed by the following method, in which I was in. My labeling method is: taking powder, see who contains its self a particle is a "background" (or which may be interesting in the following incident). I imagined spheres, each having a total charge of $-q/2$ of the other with a total charge of $+Q$, are placed side-by-side facing each other. What is the force between them? The potential can be taken from a number of charges. One first approximates each sphere by a charge on its center. These charges will have image charges in the other sphere. The image charges will have charges, etc., etc. The solution is like the picture on the box of baking powder, and it converges pretty fast.

6-10 Condensers; parallel planes

We take up now another kind of a problem involving conductors. Consider two large metal plates which are parallel to each other and separated by a distance small compared with their width. Let's suppose that equal and opposite charges have been put on the plates. The charges on each plate will be attracted by the charges on the other plate, and the charges will spread out uniformly on the inner surfaces of the plates. The plates will have surface charge densities $+\sigma$ and $-\sigma$, respectively, as in Fig. 6-12. From Chapter 5 we know that the field between the plates is $\sigma/2\epsilon_0$, and that the field outside the plates is zero. The plates will have different potentials ϕ_1 and ϕ_2 . For convenience we will call the difference V ; it is often called the "voltage".

$$\phi_1 - \phi_2 = V$$

(Note well first that sometimes people use V for the potential, but we have chosen to use ϕ .)

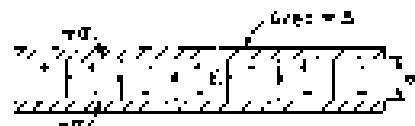


Fig. 6-12 A parallel plate condenser

The potential difference V is the work per unit charge required to bring a small charge from one plate to the other.

$$V = \frac{Q}{C} = \frac{\pi d^2}{4\epsilon_0 A} \cdot \frac{1}{d} Q, \quad (6.11)$$

where A/Q is the surface charge density on each plate, d is the distance between plates, and ϵ_0 is the dielectric constant.

We find that the voltage is proportional to the charge. Another proportionality between V and Q is found for any two conductors in space if there is a charge Q on one and an equal negative charge on the other. The potential difference between them due to the voltage will be proportional to the charge. (We are assuming no other external charges are nearby.)

Very this proves reciprocity, i.e., the superposition principle. Suppose we have the solution to one set of charges, and then we superimpose two such solutions. The charges are distributed, the fields are additive, and the work done in carrying a unit charge from one point to the other is also additive. Therefore the potential difference between two points is proportional to the charges. In particular, the potential difference between two points is proportional to the charges on them, assuming one grid of points. By symmetry of problem on the other side. That is they wrote

$$Q = \epsilon V,$$

where V is a constant. This coefficient of proportionality is called the capacity of the system of two conductors, or called a capacitor. (For a parallel-plate capacitor it is

$$C = \frac{A}{d} \quad (\text{per unit charge}). \quad (6.12)$$

This formula is not exact. Because the field is not truly uniform between the plates, as we assumed. The field does not just suddenly end at the edges of the plates, as shown in Figure 12. The field changes smoothly, as we have learned. There is a little curvature for the field lines; the edges of the plates, where the surface density, σ , is large, force the field lines to curve and bend outward with their impetus at the edges. (This is not illustrated in the sketch which is, however, the subject of techniques which we will not discuss here.) The result of such edge effects is that the average density is somewhat less than the edges of the plates. This means that the capacity of the plates is actually less than we computed. (We can compute exactly the average value of the field, however, in which case $C = A/d$ for the case of a cavity.) (In practice, however, by a factor of 3 or 4 the separation between the plates.)

We have talked about the case of two conductors only. Since, in general, there is more capacity of a single object, why say the more the capacity of a system of conductors? Well, everything is that the potential of a conductor is the sum of individual values. Therefore, there is a rule of superposition of capacities. If you have an infinite system of conductors, the aggregate capacity is ΣC_i , even in infinite system. One can also speak of capacities when there are three or more conductors. A discussion of such cases, however, defers.

Suppose now we wish to have a condenser with a very large capacity. We could get a large capacity by taking a very big distance or very small ϵ_0 . We could put waxed paper between sheets of aluminum foil and call it ϵ_0 . If we did it in plastic, we have a typical dielectric insulator. What you do not know is your insulating strength. It is very hard to characterize it. For example, an insulator has capacity as yet charged ... it may not be charged, that is the charge begins to escape into the air. But if you put the same charge on a conductor whose capacity is very large, the voltage drop would be low. Insulation would be small.

¹ Some people think it is good "superconducting" because it does not need any heat or cooling, or "heatless." We have defined it as the heatless terminology because it will more conveniently lead to the heatless laboratory experiments in the last chapter.

In many applications in electronic circuits, it is useful to have something which can absorb or deliver large quantities of charge without changing its potential much. A capacitor (or "capacitor") does just that. There are also many applications in electronic instruments and in computers where a capacitor is used to get a specified change in voltage in response to a particular change in charge. We have seen a similar application in Chapter 22, Vol. I, where we described the properties of resonant circuits.

From the definition of C , we see that its unit is one coulomb. This unit is also called a farad. Looking at Eq. (6.3), we see that one can express the units of C as farad/meter^2 , which is the unit most commonly used. Typical sizes of capacitors run from one micro-microfarad ($\sim 1 \text{ picofarad}$) to millifarads. Small condensers of a few picofarads are used in high-frequency tuned circuits, and capacitors up to hundreds or thousands of microfarads are found in power-supply filters. A pair of plates one square centimeter in area with a capacitance separation have a capacity of roughly one micro-microfarad.

6-11 High-voltage breakdown

We would like now to discuss qualitatively some of the characteristics of the fields around conductors. If we charge a conductor that is not a sphere, but instead has on it a point or a very sharp end, as, for example, the object sketched in Fig. 6-14, the field around the point is much higher than the field in the other regions. The reason is, qualitatively, that charges try to spread out as much as possible on the surface of a conductor, and the tip of a sharp point is as far away as it is possible to be from most of the surface. Some of the charges on the plate get pushed all the way to the tip. A relatively small amount of charge on the tip can still provide a large surface density; a high charge density means a high field, as we do.

The way to see that the fact is highest at those places on a conductor where the radius of curvature is smallest, is to consider the combination of a big sphere and a little sphere connected by a wire as shown in Fig. 6-15. It is somewhat simplified versions of the conductors of Fig. 6-14. The wire will have little influence on the fields outside, it is there to keep the spheres at the same potential. Now, if each ball has the same field at its surface? If the ball on the left has the radius a and carries a charge Q , its potential is about

$$\phi_1 = \frac{1}{4\pi\epsilon_0} \frac{Q}{a}$$

(Of course the presence of one ball changes the charge distribution on the other, so that the charges are not really spherically symmetric on either. But if we are interested only in an estimate of the fields, we can use the potential of a spherical charge.) If the smaller ball, whose radius is b , carries the charge q , its potential is about

$$\phi_2 = \frac{1}{4\pi\epsilon_0} \frac{q}{b}$$

But $a = \phi_1 \cdot \infty$

$$\frac{Q}{a} = \frac{q}{b}$$

On the other hand, the field at the surface (see Eq. 5.8) is proportional to the surface charge density, which is like the total charge over the radius squared. We get that

$$\frac{\Delta\phi}{\Delta r} = \frac{Q/a^2}{q/b^2} = \frac{b}{a}. \quad (6.5)$$

Therefore the field is higher at the surface of the small sphere. The fields are in the inverse proportion of the radii.

This result is technically very important, because air will break down if the electric field is too great. What happens is that a free charge (electron, or ion) somewhere in the air is accelerated by the field, and if the field is very great, the charge can pick up enough speed before it loses energy due to friction with

$$C_0 = \frac{1}{36\pi} \times 10^{-9} \text{ Farad/meter}$$

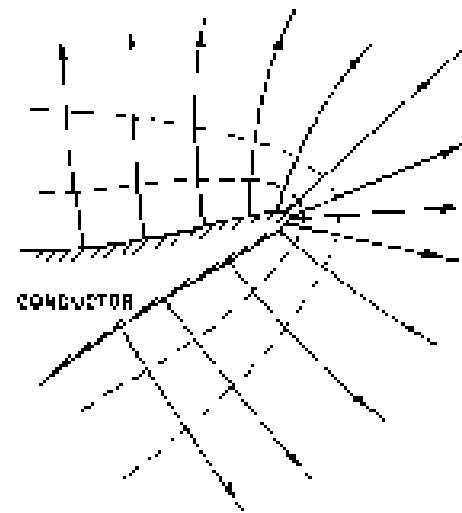


Fig. 6-14. The electric field near a sharp point on a conductor is very high.

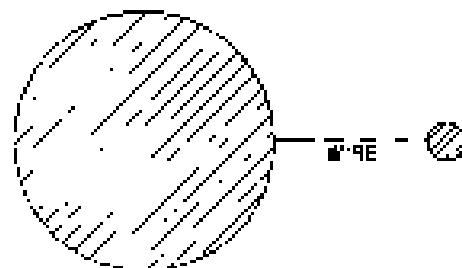


Fig. 6-15. The field of a pointed object can be approximated by that of two spheres of the same potential.

electron off that atom. As a result, more and more ions are produced. Their motion constitutes a discharge, or spark. If you want to charge an object to a high potential and not have it discharge itself by sparks in the air, you must be sure that the surface is smooth, so that there is no place where the field is abnormally large.

6-12 The field-emission microscope

There is an interesting application of the extremely high electric field which surrounds any sharp protuberance or a charged conductor. The field-emission microscope depends for its operation on the high fields produced at a sharp metal point.* It is built in the following way. A very fine needle, with a tip whose diameter is about .0001 cm., is placed at the center of an evacuated glass sphere (Fig. 6-16). The inner surface of the sphere is coated with a thin conducting layer of fluorescent material, and a very high potential difference is applied between the fluorescent coating and the needle.

Let's first consider what happens when the needle is negative with respect to the fluorescent coating. The field lines are highly concentrated at the sharp point. The electric field can be as high as 10 million volts per centimeter. In such intense fields, electrons are pulled out of the surface of the needle and accelerated across the potential difference between the needle and the fluorescent layer. When they arrive there they cause light to be emitted, just as in a television picture tube.

The electrons which arrive at a given point on the fluorescent surface are, to an excellent approximation, those which leave the other end of the radial field line, because the electrons will travel along the field line passing from the point to the surface. Thus we see on the surface spots due to passage of the tip of the needle. More precisely, we see a picture of the sensitivity of the surface of the needle. That is the case with which electrons can leave the surface of the metal tip. If the resolution were high enough, one could hope to resolve the positions of the individual atoms on the tip of the needle. With electrons, that resolution is not possible for the following reasons. First, there is quantum-mechanical diffraction of the electron waves which blur the image. Second, due to the internal motions of the electrons in the metal they have a small sideways initial velocity when they leave the needle, and this random sideways component of the velocity causes some smearing of the image. The combination of these two effects limits the resolution to 25 Å or so.

If, however, we reverse the polarity and introduce a small amount of helium gas into the bulb, much lighter resolutions are possible. When a helium atom collides with the tip of the needle, the intense field there strips an electron off the helium atom, leaving it positively charged. The positive ion is then accelerated outward along a field line to the fluorescent screen. Since the ionization is so much heavier than an electron, the quantum-mechanical wavelengths are much smaller. If the temperature T is not too high, the effect of the thermal velocities is also smaller than in the electron case. With an ionizing voltage of 1000 V, the resolution of the picture is obtained. It has been possible to obtain magnifications up to 2,000,000 times with the best field-ion microscopes, a magnification ten times greater than is obtained with the best electron microscopes.

Figure 6-17 is an example of the results which were obtained with a field-ion microscope, using a tungsten needle. The center of a tungsten atom invites a helium atom at a slightly different rate than the spaces between the tungsten atoms. The pattern of spots on the fluorescent screen shows the arrangement of the individual atoms on the tungsten tip. The reason the spots appear in rings can be understood by visualizing a large box of balls stacked in a rectangular array, representing the atoms in the metal. If you cut an approximately spherical section out of this box, you will see the ring pattern characteristic of the atomic structure. The field-ion microscope provided human beings with the means of seeing atoms for the first time. This is a remarkable achievement, considering the simplicity of the instrument.

* See L. W. Mueller, "The field-ion microscope," *Advances in Electronics and Electron Physics*, 13, 81-176 (1960). Academic Press, New York.

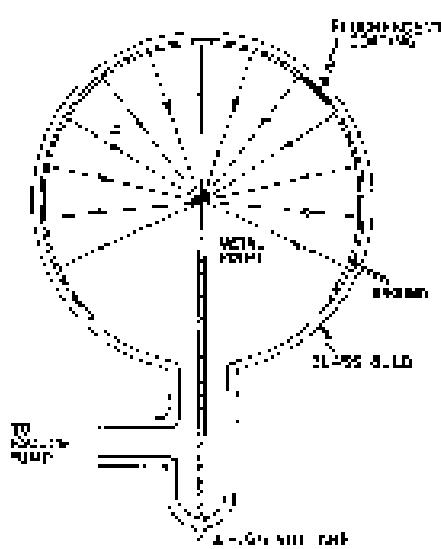


Fig. 6-16. Field-emission microscope.

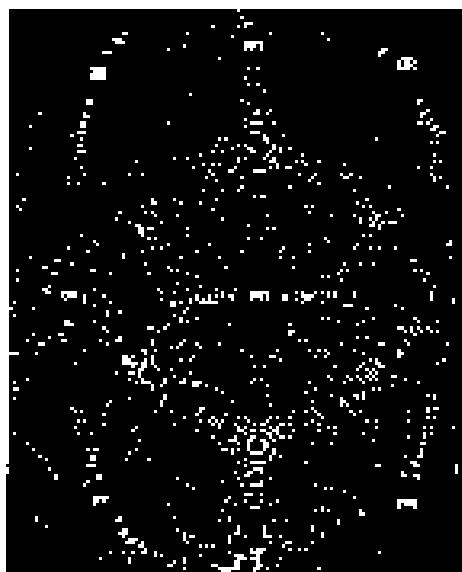


Fig. 6-17. Image produced by a field-ion microscope. [Courtesy of Erwin W. Mueller, Research Prof. of Physics, Pennsylvania State University.]

The Electric Field in Various Circumstances (Continued)

7-1 Methods for finding the electrostatic field

This chapter is a continuation of our consideration of the characteristics of electric fields in various particular situations. We shall first describe some of the more elaborate methods for solving problems with conductors. It is not expected that these more advanced methods can be mastered at this time. Yet it may be of interest to have some idea about the kinds of problems that can be solved, using techniques that may be learned in more advanced courses. Then we take up two examples in which the charge distribution is neither fixed nor is carried by a conductor, but instead is determined by some other law of physics.

As we found in Chapter 6, the problem of the electrostatic field is fundamentally simple when the distribution of charges is specified; it requires only the evaluation of an integral. When there are conductors present, however, complications arise because the charge distribution on the conductors is not initially known; the charge must distribute itself on the surface of the conductor in such a way that the conductor is an equipotential. The solution of such problems is neither difficult nor simple.

We have looked at an indirect method of solving such problems, in which we find the equipotentials for some specified charge distribution and replace one of them by a conducting surface. In this way we can build up a catalog of special solutions for conductors in the shapes of spheres, planes, etc. The use of images, described in Chapter 6, is an example of an indirect method. We shall describe another in this chapter.

If the problem to be solved does not belong to the class of problems for which we can construct solutions by the indirect method, we are forced to solve the problem by a more direct method. The mathematical problem of the direct method is the solution of Laplace's equation,

$$\nabla^2 \phi = 0. \quad (7.1)$$

subject to the condition that ϕ is a suitable constant on certain boundaries—the surfaces of the conductors. Problems which involve the solution of a differential field equation subject to certain boundary conditions are called boundary value problems. They have been the object of considerable mathematical study. In the case of conductors having complicated shapes, there are no general analytical methods. Even such a simple problem as that of a charged cylindrical metal can elicit at both ends a beer can presents formidable mathematical difficulties. It can be solved only approximately, using numerical methods. The only general methods of solution are numerical.

There are a few problems for which Eq. (7.1) can be solved directly. For example, the problem of a charged conductor having the shape of an ellipsoid of revolution can be solved exactly in terms of known special functions. The solution for a thin disc can be obtained by letting the ellipsoid become infinitely oblate. In a similar manner, the solution for a needle can be obtained by letting the ellipsoid become infinitely prolate. However, it must be stressed that the only direct methods of general applicability are the numerical techniques.

Boundary value problems can also be solved by measurement of a physical quantity. Laplace's equation arises in many different physical situations: in steady-state heat flow, in irrotational fluid flow, in current flow in an extended medium,

7-1 Methods for finding the electrostatic field

7-2 Two-dimensional fields; functions of the complex variable

7-3 Plasma oscillations

7-4 Colloidal particles in air; closoxyne

7-5 The electrostatic field of a grid

and in the deflection of an elastic membrane. It is frequently possible to set up a physical model which is analogous to an electrical problem which we wish to solve. By the measurement of a suitable analogous quantity on the model, the solution to the problem of interest can be determined. An example of the analog technique is the use of the electrolytic tank for the solution of two-dimensional problems in electrodynamics. This works because the differential equation for the potential in a uniform conducting medium is the same as it is for a vacuum.

There are many physical situations in which the variations of the physical fields in one direction are zero, or can be neglected in comparison with the variations in the other two directions. Such problems are called two-dimensional; the field depends on two coordinates only. For example, if we place a long charged wire along the z -axis, then for points not too far from the wire the electric field depends on x and y , but not on z ; the problem is two-dimensional. Since in a two-dimensional problem $\partial/\partial z = 0$, the equation for ϕ in free space is

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0. \quad (7.2)$$

Because the two-dimensional equation is comparatively simple, there is a wide range of conditions under which it can be solved analytically. There is, in fact, a very powerful indirect mathematical technique which depends on a theorem from the mathematics of functions of a complex variable, and which we will now describe.

7-2 Two-dimensional fields; functions of the complex variable

The complex variable β is defined as

$$\beta = x + iy$$

(Do not confuse β with the z -coordinate, which we ignore in the following discussion because we assume there is no z -dependence of the fields.) Every point in x and y then corresponds to a complex number β . We can use β as a single (complex) variable, and with it write the usual kinds of mathematical functions $F(\beta)$. For example,

$$F(\beta) = \beta^2,$$

or

$$F(\beta) = 1/\beta^2,$$

or

$$F(\beta) = \beta \log \beta,$$

and so forth.

Given any particular $F(\beta)$ we can substitute $\beta = x + iy$, and we have a function of x and y , with real and imaginary parts. For example,

$$\beta^2 = (x + iy)^2 = x^2 - y^2 + 2ixy. \quad (7.3)$$

Any function $F(\beta)$ can be written as a sum of a pure real part and a pure imaginary part, each part a function of x and y :

$$F(\beta) = U(x, y) + iV(x, y), \quad (7.4)$$

where $U(x, y)$ and $V(x, y)$ are real functions. Thus from any complex function $F(\beta)$ two new functions $U(x, y)$ and $V(x, y)$ can be derived. For example, $F(\beta) = \beta^2$ gives us the two functions

$$U(x, y) = x^2 - y^2, \quad (7.5)$$

and

$$V(x, y) = 2xy. \quad (7.6)$$

Now we come to a marvelous mathematical theorem which is so delightful that we shall leave a proof of it for one of your courses in mathematics. (We should not reveal all the mysteries of mathematics, or that subject would

become zero (null). It is this. For any "ordinary function" (mathematicians will define it better) the functions U and V automatically satisfy the relations:

$$\frac{\partial U}{\partial x} = \frac{\partial V}{\partial y}, \quad (7.7)$$

$$\frac{\partial V}{\partial x} = -\frac{\partial U}{\partial y}. \quad (7.8)$$

It follows immediately that each of the functions U and V satisfy Laplace's equation:

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0, \quad (7.9)$$

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0. \quad (7.10)$$

These equations are clearly true for the functions of (7.5) and (7.6).

Thus, starting with any arbitrary function, we can arrive at two functions $U(x, y)$ and $V(x, y)$, which are both solutions of Laplace's equation in two dimensions. Each function represents a possible electrostatic potential. We can pick one (arbitrary A) and it should represent some electric field problem—in fact, *too* problems, because U and V each represent solutions. We can write down as many solutions as we wish—by just making up functions—but we just have to find the problem that goes with each solution. It may sound backwards, but it's a possible approach.

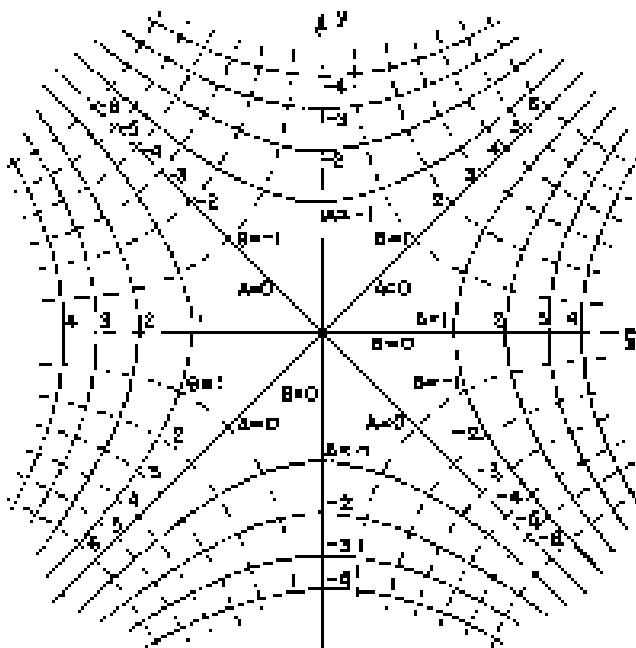


Fig. 7-1. Two sets of elliptical curves which can represent equipotentials in a two-dimensional electrostatic field.

As an example, let's see what physics the function $A(y) = y^2$ gives us. From (7.6) we get the two potential functions of (7.5) and (7.6). To see what problem the function y belongs to, we solve for the equipotential surfaces by setting $U = A$, a constant:

$$x^2 - y^2 = A.$$

This is the equation of a rectangular hyperbola. For various values of A , we get the hyperbolae shown in Fig. 7-1. When $A = 0$, we get the special case of diagonal straight lines through the origin.

Such a set of equipotentials corresponds to several possible physical situations. First, it represents the fine details of the field near the point halfway between two

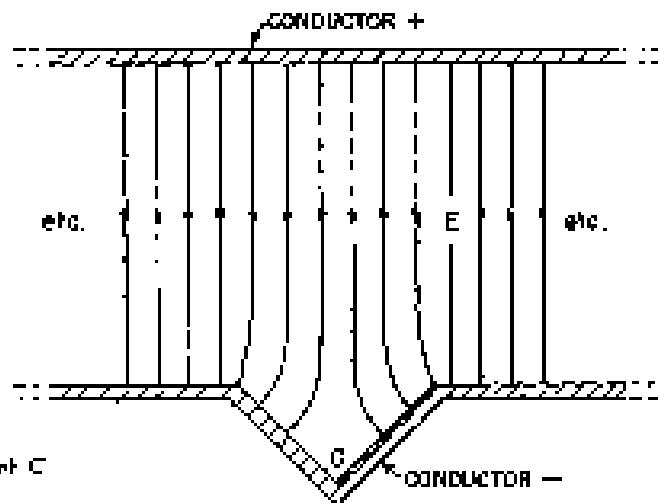


Fig. 7-2. The field near the point C
is the same as that in Fig. 7-1.

equal point charges. Second, it represents the field at an inside right-angle corner of a conductor. If we have two electrodes shaped like those in Fig. 7-2, which are held at different potentials, the field near the corner marked C will look just like the field above the origin in Fig. 7-1. The solid lines are the equipotentials, and the broken lines at right angles correspond to lines of E . Whereas at points of protruberances the electric field tends to be high, it tends to be low in dents or hollows.

The solution we have found also corresponds to that for a hyperbola-shaped electrode near a right-angle corner, or for two hyperbolae at suitable potentials. You will notice that the field of Fig. 7-1 has an interesting property. The x -component of the electric field, E_x , is given by

$$E_x = -\frac{\partial \phi}{\partial x} = -2x.$$

The electric field is proportional to the x distance from the axis. This fact is used to make devices (called quadrupole lenses) that are useful for focusing particle beams (see Section 29-9); the lens field is usually obtained by using four hyperbola-shaped electrodes, as shown in Fig. 7-3. For the electric field lines in Fig. 7-3, we have simply copied from Fig. 7-1 the set of broken-line curves that represent $V = \text{constant}$. We have a bonus! The curves for $V = \text{constant}$ are orthogonal to the ones for $\phi = \text{constant}$ because of the equations (7.7) and (7.8). Whenever we choose a function $F(z)$, we get from 7-1 and 7-2 both the equipotentials and field lines. And you will remember that we have solved either of these problems, depending on which set of curves we call the equipotentials.

As a second example, consider the function

$$F(z) = \sqrt{z}. \quad (7.11)$$

If we write

$$z = x + iy = \rho e^{i\theta},$$

where

$$\rho = \sqrt{x^2 + y^2}$$

and

$$\tan \theta = y/x,$$

then

$$F(z) = \rho^{1/2} e^{i\theta/2} \\ = \rho^{1/2} \left(\cos \frac{\theta}{2} + i \sin \frac{\theta}{2} \right).$$

From which

$$F(z) = \left[\left(\frac{x^2}{\rho} + \frac{y^2}{\rho} \right)^{1/2} + i \right]^{1/2} + i \left[\left(\frac{x^2}{\rho} + \frac{y^2}{\rho} \right)^{1/2} - x \right]^{1/2} \quad (7.12)$$

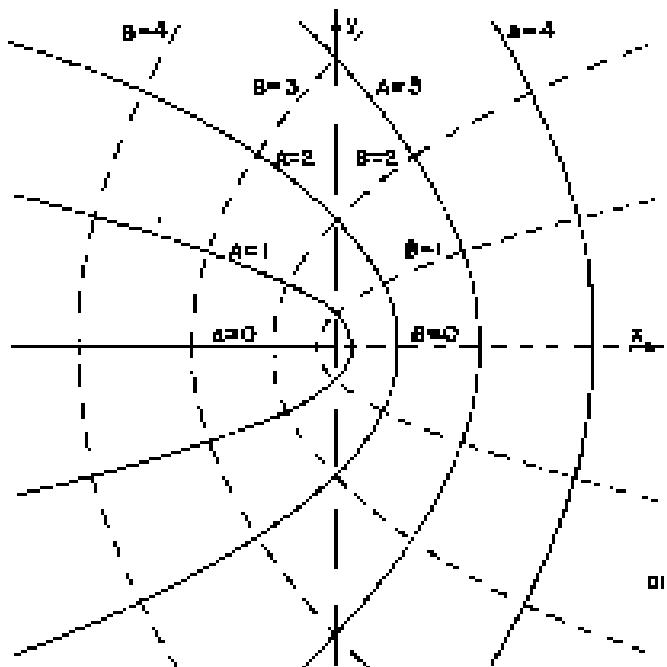


Fig. 7-4. Curves of constant $C(x,y)$ and $V(x,y)$ from Eq. (7.12).

The curves for $C(x,y) = A$ and $V(x,y) = B$, using \mathcal{U} and V from Eq. (7.12), are plotted in Fig. 7-4. Again, there are many possible situations that could be described by these fields. One of the most interesting is the field near the edge of a thin plate. If the line $R = 0$ —to the right of the x -axis—represents a thin charged plate, the field lines near it are given by the curves for various values of A . The physical situation is shown in Fig. 7-5.

Another example fits

$$P(1) = z^{1/2}, \quad (7.13)$$

which yields the field outside a rectangular corner

$$P(z) = \log z, \quad (7.14)$$

which yields the field for a unit charge, and

$$P(z) = 1/z, \quad (7.15)$$

which gives the field for the two-dimensional analog of an electric dipole, i.e., two parallel line charges with opposite volatilities, very close together.

We will not pursue this subject further in this course, but should emphasize that although the complex variable technique is often powerful, it is limited to two-dimensional problems; and also, it is an indirect method.

7.3 Plasma oscillations

We consider now some physical situations in which the field is determined, neither by fixed charges nor by charges on conducting surfaces, but by a combination of two physical phenomena. In other words, the field will be governed simultaneously by two sets of equations: (1) the equations from electromagnetism relating electric fields to charge distribution, and (2) an equation from another part of physics that determines the positions or motions of the charges in the presence of the field.

The first example that we will discuss is a dynamic one in which the motion of the charges is governed by Newton's laws. A simple example of such a situation occurs in a plasma, which is an ionized gas consisting of ions and free electrons distributed over a region in space. The ionosphere—the upper layer of the atmosphere—is an example of such a plasma. The ultraviolet rays from the sun knock

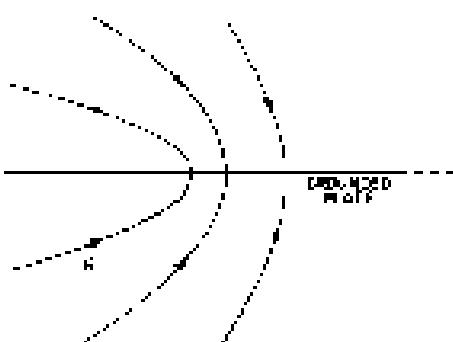


Fig. 7-5. The electric field near the edge of a thin grounded plate.

electrons off the molecules of the air, creating free electrons and ions—an ionized plasma. The positive ions are very much heavier than the electrons, so we may neglect the ionic motion; it is approximately zero at low temperature.

Let n_0 be the density of electrons in the undisturbed, equilibrium state. This must be equal to the density of positive ions, since the plasma is electrically neutral (after unbalance). Now we suppose that the electrons are somehow moved from equilibrium. One thing can happen. If the density of the electrons in one region is increased, they will repel each other and tend to return to their original equilibrium positions. As they return toward their original positions, they pick up kinetic energy, and instead of settling to rest in their equilibrium configuration, they overshoot the mark. They will oscillate back and forth. The situation is similar to what occurs in sound waves, in which the restoring force is the gas pressure. In a plasma, the restoring force is the electrical force on the electrons.

To simplify the discussion, we will consider only the case in which the electrons oscillate in one dimension, say x . Let's suppose that the electrons originally at $x = 0$, at time $t = 0$, displace themselves to equilibrium positions by a small amount of Δx . Since the electrons are nonrelativistic, their density n is, in general, being changed. The change in density is easily calculated. Referring to Fig. 7-6, the electrons initially contained between the two planes $x = 0$ and $x = \Delta x$ are moved and now occupy the interval between $x = -\Delta x$ and $x = \Delta x$. The number of electrons that were originally at $x = 0$ is proportional to $n_0 \Delta x$; the new number are now contained in the space whose width is $\Delta x + 2\Delta x$. The density has changed to

$$n = \frac{n_0 \Delta x}{\Delta x + 2\Delta x} = \frac{n_0}{1 + (2\Delta x/\Delta x)}. \quad (7.16)$$

If the change in density is small, we can write (using the binomial expansion for $(1 + x)^{-1}$)

$$n = n_0 \left(1 - \frac{\Delta x}{\Delta x}\right). \quad (7.17)$$

We assume that the positive ions do not move appreciably (in view of the much larger inertia), so their density remains n_0 . Each electron carries the charge $-e$, so the average charge density at any point is given by

$$\begin{aligned} \rho &= -e(n - n_0) \\ &= -en_0 \frac{\partial}{\partial x}. \end{aligned} \quad (7.18)$$

Let us now consider the differential, small for $\Delta x/\Delta x$.

The charge density is related to the electric field by Maxwell's equations, in particular,

$$\nabla \cdot \mathbf{D} = \frac{\rho}{\epsilon_0}. \quad (7.19)$$

If the problem is indeed one-dimensional (and if there are no other fields but the one due to the displacement of the electrons), the electric field E has a single component, E_x . Equation (7.19), together with (7.18), gives

$$\frac{\partial E_x}{\partial y} = \frac{\rho_0 \partial n_0}{\epsilon_0 \partial x}. \quad (7.20)$$

Integrating Eq. (7.20) gives

$$E_x = \frac{\rho_0}{\epsilon_0} x + K, \quad (7.21)$$

Since $E_x = 0$ when $x = 0$, the integration constant K is zero.

The force on an electron in the displaced position is

$$F_x = -\frac{\partial^2 \mathcal{H}}{\partial x^2} e. \quad (7.22)$$

a restoring force proportional to the displacement s of the electron. This leads to a harmonic oscillation of the electrons. The equation of motion of a displaced electron is

$$m_e \frac{d^2 s}{dt^2} = -\frac{\kappa_0 q_e^2}{\epsilon_0} s. \quad (7.23)$$

We find that s will vary harmonically. Its sinus variation will be $\propto \cos \omega t$, or—using the exponential notation of Vol. I—²

$$e^{i\omega t}. \quad (7.24)$$

The frequency of oscillation ω_p is determined from (7.23):

$$\omega_p^2 = \frac{\kappa_0 q_e^2}{\epsilon_0 m_e}, \quad (7.25)$$

and is called the plasma frequency. It is a characteristic number of the plasma.

When dealing with electron charges many people prefer to express their answers in terms of a quantity ϵ^2 defined by

$$\epsilon^2 = \frac{q_e^2}{4\pi\epsilon_0} \approx 2.3068 \times 10^{-24} \text{ newton-meter}^2. \quad (7.26)$$

Using this convention, Eq. (7.25) becomes

$$\omega_p^2 = \frac{4\pi e^2 n_e}{m_e}. \quad (7.27)$$

which is the form you will find in most books.

Thus we have found that a disturbance of a plasma will set up free oscillations of the electrons about their equilibrium positions at the natural frequency ω_p , which is proportional to the square root of the density of the electrons. The plasma electrons behave like a resonant system, such as those we described in Chapter 23 of Vol. I.

This natural resonance of a plasma has some interesting effects. For example, if our task is to propagate a radiowave through the ionosphere, our finds that it can penetrate only if the frequency is higher than the plasma frequency. Otherwise the signal is reflected back. We could use high frequencies if we wish to communicate with a satellite in space. On the other hand, if we wish to communicate with a radio station beyond the horizon, we must use frequencies lower than the plasma frequency, so that the signal will be reflected back to the earth.

Another interesting example of plasma oscillations occurs in nuclei. In a metal we have a contained plasma of positive ions, and free electrons. The density n_e is very high, so ω_p is also. But it should still be possible to observe the electron oscillations. Now, according to quantum mechanics, a harmonic oscillator with a natural frequency ω_p has energy levels which are separated by the the energy increment $\hbar\omega_p$. If, then, one shoots electrons through, say, an aluminum foil, and makes very careful measurements of the electron energies on the other side, one might expect to find that the electrons sometimes lost the energy $\hbar\omega_p$ to the plasma oscillations. This does indeed happen. It was first observed experimentally in 1956 that electrons which energies of a few hundred to a few thousand electron volts lost energy in incease when scattering from or going through a thin metal foil. The effect was not understood until 1953 when Hobbs and Pines* showed that the observations could be explained in terms of quantum excitations of the plasma oscillations in the metal.

* For some recent work and a bibliography see C. J. Powell and J. R. Szwinger, Phys. Rev. 128, 869 (1962).

7-4 Colloidal particles in an electrolyte

We turn to another phenomenon in which the actions of charges is governed by a potential that arises in part from the same charges. The resulting effects influence in an important way the behavior of colloids. A colloid consists of a suspension in water of small charged particles which, though microscopic, from an atomic point of view are still very large. If the colloidal particles were not charged, they would tend to coagulate into large lumps; but because of their charge, they repel each other and remain in suspension.

Now if there is also some salt dissolved in the water, it will be dissociated into positive and negative ions. (Such a solution of ions is called an electrolyte.) The negative ions are attracted to the colloid particles (assuming their charge is positive) and the positive ions are repelled. We will determine how the ions which surround such a colloidal particle are distributed in space.

To keep the ideas simple, we will again solve only a one-dimensional case. If we think of a colloidal particle as a sphere having a very large radius on an atomic scale, we can then treat a small part of its surface as a plane. (Whenever one is trying to understand a new phenomenon it is a good idea to take a somewhat oversimplified model; then, having understood the problem with that model, one is better able to proceed to tackle the more exact calculation.)

We suppose that the distribution of ions generates a charge density $\rho(x)$, and an electrical potential ϕ , related by the electrostatic law $\nabla^2\phi = -\rho/e$, or, for fields that vary in only one dimension, by

$$\frac{d^2\phi}{dx^2} = -\frac{\rho}{\epsilon_0}. \quad (7.25)$$

Now supposing there were such a potential $\phi(x)$, how would the ions distribute themselves in it? This we can determine by the principles of statistical mechanics. Our problem then is to determine ϕ so that the resulting charge density [from statistical mechanics (7.25)] satisfies (7.25).

According to statistical mechanics (see Chapter 46, Vol. II), particles in thermal equilibrium in a force field are distributed in such a way that the density n of particles at the position x is given by

$$n(x) = n_0 e^{-E(x)/kT}, \quad (7.26)$$

where $E(x)$ is the potential energy, k is Boltzmann's constant, and T is the absolute temperature.

We assume that the ions carry one electronic charge, positive or negative. At the distance x from the surface of a colloidal particle, a positive ion will have potential energy $q\phi(x)$, so that

$$E(x) = q\phi(x).$$

The density of positive ions, n_+ , is then

$$n_+(x) = n_0 e^{-q\phi(x)/kT}.$$

Similarly, the density of negative ions is

$$n_-(x) = n_0 e^{+q\phi(x)/kT}.$$

The total charge density is

$$\begin{aligned} \rho &= n_+ n_- = q n_-, \\ \rho &= q n_0 e^{-q\phi(x)/kT} e^{+q\phi(x)/kT}. \end{aligned} \quad (7.27)$$

Combining this with Eq. (7.25), we find that the potential ϕ must satisfy

$$\frac{d^2\phi}{dx^2} = -\frac{q\epsilon_0}{\epsilon_0} (e^{-q\phi(x)/kT} - e^{+q\phi(x)/kT}). \quad (7.28)$$

This equation is readily solved in general [multiply both sides by $2(\partial\phi/\partial x)$, and integrate with respect to x], but to keep the problem as simple as possible, we will consider here only the limiting case in which the potentials are small or the temperature T is high. The case where ϕ is small corresponds to a dilute solution. For these cases the exponent is small, and we can approximate

$$e^{16\pi\phi/kT} \approx 1 + \frac{8\pi\phi}{kT}. \quad (7.32)$$

Equation (7.31) then gives

$$\frac{d^2\phi}{dx^2} = + \frac{2\pi\sigma g^2}{\epsilon_0 kT} \psi(x). \quad (7.33)$$

Notice that this time the sign on the right is positive. The solutions for ϕ are not oscillatory, but exponential.

The general solution of Eq. (7.33) is

$$\phi = Ax^{-2/D} + Bx^{+2/D}, \quad (7.34)$$

with

$$D^2 = \frac{\epsilon_0 kT}{2\pi\sigma g^2}. \quad (7.35)$$

The constants A and B must be determined from the conditions of the problem. In our case, B must be zero; otherwise the potential would go to infinity for large x . So we have that

$$\phi = Ax^{-2/D}, \quad (7.36)$$

in which A is the potential at $x = 0$, the surface of the colloid particle.

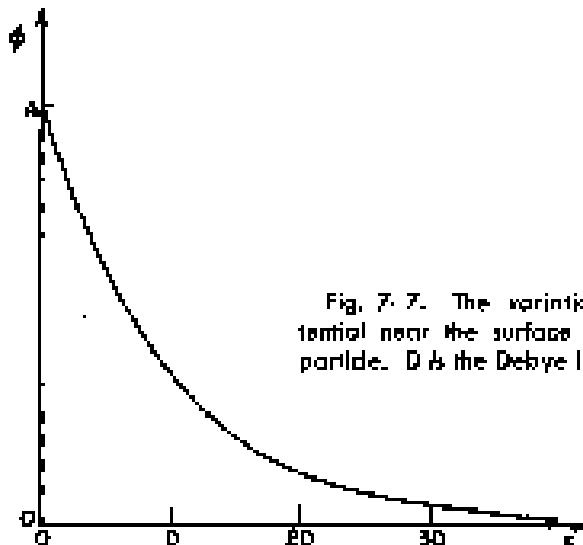


Fig. 7-7. The variation of the potential near the surface of a colloidal particle. D is the Debye length.

The potential decreases by a factor $1/e$ each time the distance increases by D , as shown in the graph of Fig. 7-7. The number D is called the Debye length, and is a measure of the thickness of the double layer surrounding a single charged particle in an electrolyte. Equation (7.36) says that the double layer thins with increasing concentration of the ions (n_i) or with decreasing temperature.

The constant A in Eq. (7.36) is easily obtained if we know the surface charge on the colloid particle. We know that

$$E_s = E_s(0) = \frac{f}{\epsilon_0}. \quad (7.37)$$

But E is also the gradient of ϕ :

$$E_s(0) = - \left. \frac{d\phi}{dx} \right|_0 = + \frac{A}{D}. \quad (7.38)$$

From which we get

$$A = \frac{\sigma D}{\epsilon_0}. \quad (7.39)$$

Using this result in (7.36), we find (by taking $\kappa = 0$) that the potential of the colloidal particle is

$$\phi(0) = \frac{\pi D}{\kappa_1}. \quad (7.40)$$

You will notice that this potential is the same as the potential difference across a condenser with a plate spacing D and a surface charge density σ .

We have said that the colloidal particles are kept apart by their electrical repulsion. But now we see that the field a little way from the surface of a particle is reduced by the ion sheath that collects around it. If the sheath get thin enough, the particles lose a good chance of knocking against each other. They will then stick, and the colloid will coagulate and precipitate out of the liquid. From our analysis, we understand why adding enough salt to a colloid should cause it to precipitate out. The process is called "salting out a colloid."

Another interesting example is the effect that a salt solution has on protein molecules. A protein molecule is a long complicated, and flexible chain of amino acids. The molecule has various charges on it, and it sometimes happens that there is a net charge, say negative, which is distributed along the chain. Because of mutual repulsion of the negative charges, the protein chain is kept stretched out. Now, if there are other similar chain molecules present in the solution, they will be kept apart by the same repulsive effects. We can, therefore, have a suspension of chain molecules in a liquid. But if we add salt to the liquid we change the properties of the suspension. As salt is added to the solution, increasing the Debye distance, the chain molecules can approach one another, and can also coil up. If enough salt is added to the solution, the chain molecules will precipitate out of the solution. There are many chemical effects of this kind that can be understood in terms of electrical forces.

7-5 The electrostatic field of a grid

As our last example, we would like to describe another interesting property of electric fields. It is one which is made use of in the design of electrical instruments, in the construction of vacuum tubes, and for other purposes. This is the character of the electric field near a grid of charged wires. To make the problem as simple as possible let us consider an array of parallel wires lying in a plane, the wires being infinitely long and with equal form spacing between them.

If we look at the field a large distance above the plane of the wires, we see a constant electric field, just as though the charge were uniformly spread over a plane. As we approach the grid of wires, the field begins to deviate from the uniform field; we find at large distances from the grid. We would like to estimate how close to the grid we have to be in order to see appreciable variations in the potential. Figure 7-5 shows a rough sketch of the equipotential surfaces at various distances from the grid. The closer we get to the grid, the larger the variations. As we travel parallel to the grid, we observe that the field fluctuates in a periodic manner.

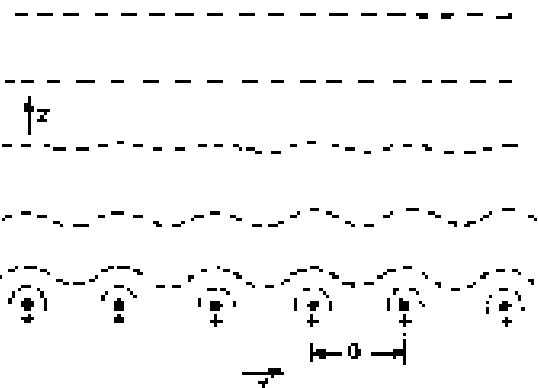


Fig. 7-5. Equipotential surfaces above a uniform grid of charged wires.

Now we have seen (Chapter 50, Vol. I) that any periodic quantity can be expressed as a sum of sine waves (Fourier's theorem). Let's see if we can find a suitable harmonic function that satisfies our field equations.

If the wires lie in the xz -plane and run parallel to the y -axis, then we might try terms like

$$\phi(x, z) = F_n(z) \cos \frac{2\pi n x}{a}, \quad (7.41)$$

where a is the spacing of the wires and n is the harmonic number. (We have assumed long wires, so there should be no variations with y .) A complete solution would be made up of a sum of such terms (for $n = 1, 2, 3, \dots$).

If this is to be a valid potential, it must satisfy Laplace's equation in the region above the wires (where there are no charges). That is,

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial z^2} = 0.$$

Trying this equation on the ϕ in (7.41), we find that

$$-\frac{4\pi^2 n^2}{a^2} F_n(z) \cos \frac{2\pi n x}{a} - \frac{d^2 F_n}{dx^2} \cos \frac{2\pi n x}{a} = 0. \quad (7.42)$$

or that $F_n(z)$ must satisfy

$$\frac{d^2 F_n}{dx^2} = -\frac{4\pi^2 n^2}{a^2} F_n. \quad (7.43)$$

So we must have

$$F_n = A_n e^{-k_n x}, \quad (7.44)$$

where

$$k_n = \frac{n}{2\pi a}. \quad (7.45)$$

We have found that if there is a Fourier component of the field of harmonic n , that component will decrease exponentially with a characteristic distance $x_0 = a/2\pi n$. For the first harmonic ($n = 1$), the amplitude falls by the factor $e^{-2\pi}$ (a large decrease) each time we increase x by one grid spacing a . The other harmonics fall off even more rapidly as we move away from the grid. We see that if we are only a few times the distance a away from the grid, the field is very nearly uniform, i.e., the oscillating terms are small. There would, of course, always remain the "zero harmonic" field

$$\phi_0 = -E_0 z$$

to give the uniform field at large z . For a complete solution, we would combine this term with a sum of terms like (7.41) with F_n from (7.44). The coefficients A_n would be adjusted so that the total sum would, when r is finite, give an electric field that would fit the charge density λ of the grid wires.

The method we have just developed can be used to explain why electrostatic shielding by means of a screen is often just as good as with a solid metal sheet. Except within a distance from the screen a few times the spacing of the screen wires, the fields inside a closed screen are zero. We see why copper screen (higher and cheaper than copper sheet) is often used to shield sensitive electrical equipment from external disturbing fields.

Electrostatic Energy

8-1 The electrostatic energy of charges. A uniform sphere

In the study of mechanics, one of the most interesting and useful discoveries was the law of the conservation of energy. The expressions for the kinetic and potential energies of a mechanical system helped us to discover connections between the states of a system at two different times without having to look into the details of what was occurring in between. We wish now to consider the energy of electrostatic systems. In electricity also the principle of the conservation of energy will be useful for discovering a number of interesting things.

The law of the energy of interaction in electrodynamics is very simple; we have, in fact, already discussed it. Suppose we have two charges q_1 and q_2 separated by the distance r_{12} . There is stored energy in the system, because a certain amount of work was required to bring the charges together. We have already calculated the work done in bringing two charges together from a large distance. It is

$$-\frac{q_1 q_2}{4\pi \epsilon_0 r_{12}}. \quad (8.1)$$

We also know, from the principle of superposition, that if we have many charges present, the total force on any charge is the sum of the forces from the others. It follows, therefore, that the total energy of a system of n number of charges is the sum of terms due to the mutual interaction of each pair of charges. If q_i and q_j are any two of the charges and r_{ij} is the distance between them (Fig. 8-1), the energy of that particular pair is

$$\frac{q_i q_j}{4\pi \epsilon_0 r_{ij}}. \quad (8.2)$$

The total electrostatic energy U is the sum of the energies of all possible pairs of charges:

$$U = \sum_{all \; i, j} \frac{q_i q_j}{4\pi \epsilon_0 r_{ij}}. \quad (8.3)$$

If we have a distribution of charge specified by a charge density ρ , the sum of Eq. (8.3) is, of course, to be replaced by an integral.

We shall concern ourselves with two aspects of this energy. One is the application of the concept of energy to electrostatic problems; the other is the evaluation of the energy in different ways. Sometimes it is easier to compute the work done for some special case than to evaluate the sum in Eq. (8.3) or the corresponding integral. As an example, let us calculate the energy required to assemble a sphere of charge with a uniform charge density. The energy is just the work done in gathering the charges together from infinity.

Imagine that we assemble the sphere by building up a succession of thin spherical layers of infinitesimal thickness. At each stage of the process, we gather a small amount of charge and put it in a thin layer from r to $r + dr$. We continue the process until we arrive at the final radius R (Fig. 8-2). If Q_r is the charge of the sphere when it has been built up to the radius r , the work done in bringing a charge dQ to it is

$$dU = \frac{Q_r dQ}{4\pi \epsilon_0 r}. \quad (8.4)$$

8-1 The electrostatic energy of charges. A uniform sphere

8-2 The energy of a condenser. Charges on charged conductors

8-3 The electrostatic energy of an ionic crystal

8-4 Electrostatic energy in nuclei

8-5 Energy in the electromagnetic field

8-6 The energy of a point charge

Review: Chapter 4, Vol. I, Conservation of Energy
Chapters 13 and 14, Vol. I.
Work and Potential Energy

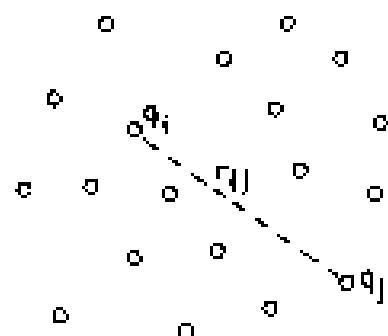


Fig. 8-1. The electrostatic energy of a system of particles is the sum of the electrostatic energy of each pair.

If the density of charge in the sphere is ρ , the charge Q_r is

$$Q_r = \rho \cdot \frac{4}{3} \pi r^3,$$

and the charge dQ is

$$dQ = \rho \cdot 4\pi r^2 dr.$$

Equation (8.4) becomes

$$dU = \frac{4\pi\rho^2 r^4 dr}{3\varepsilon_0}. \quad (8.5)$$

The total energy required to assemble the sphere is the integral of dU from $r = 0$ to $r = a$,

$$U = \frac{4\pi\rho^2 a^5}{15\varepsilon_0}. \quad (8.6)$$

Or if we wish to express the result in terms of the total charge Q of the sphere,

$$U = \frac{3}{5} \frac{Q^2}{4\pi\varepsilon_0 a^2}. \quad (8.7)$$

The energy is proportional to the square of the total charge and inversely proportional to the radius. We can also interpret Eq. (8.7) as saying that the average of $(1/r_{ij})$ for all pairs of points in the sphere is $3/5a$.

8-2 The energy of a condenser. Forces on charged conductors

We consider now the energy required to charge a condenser. If the charge Q has been taken from one of the conductors of a condenser and placed on the other, the potential difference between them is

$$V = \frac{Q}{C}, \quad (8.8)$$

where C is the capacity of the condenser. How much work is done in charging the condenser? Proceeding as for the sphere, we imagine that the condenser has been charged by transferring charge from one plate to the other in small increments dQ . The work required to transfer the charge dQ is

$$dU = V dQ.$$

Taking V from Eq. (8.8), we write

$$dU = \frac{Q dQ}{C}.$$

Or integrating from zero charge to the final charge Q , we have

$$U = \frac{1}{2} \frac{Q^2}{C}. \quad (8.9)$$

This energy can also be written as

$$U = \frac{1}{2} CV^2. \quad (8.10)$$

Recalling that the capacity of a conducting sphere (relative to infinity) is

$$C_{\text{sphere}} = 4\pi\varepsilon_0 a,$$

we can immediately get from Eq. (8.9) the energy of a charged sphere,

$$U = \frac{1}{2} \frac{Q^2}{4\pi\varepsilon_0 a}. \quad (8.11)$$

This, of course, is also the energy of a thin spherical shell of total charge Q and is just $5/6$ of the energy of a uniformly charged sphere, Eq. (8.7).

We now consider applications of the idea of electrostatic energy. Consider the following questions: What is the force between the plates of a condenser? Or what is the torque about some axis of a charged conductor in the presence of another with opposite charge? Such questions are easily answered by using our result Eq. (8.9) for electrostatic energy of a condenser, together with the principle of virtual work (Chapters 4, 13, and 14 of Vol. I).

Let's use this method for determining the force between the plates of a parallel-plate condenser. If we imagine that the spacing of the plates is increased by the small amount Δz , then the mechanical work done from the outside in moving the plates would be

$$\Delta W = F \Delta z, \quad (8.12)$$

where F is the force between the plates. This work must be equal to the change in the electrostatic energy of the condenser.

By Eq. (8.9), the energy of the condenser was originally

$$U = \frac{1}{2} \frac{Q^2}{C}.$$

The change in energy (if we do not let the charge change) is

$$\Delta U = \frac{1}{2} Q^2 \Delta \left(\frac{1}{C} \right). \quad (8.13)$$

Equations (8.12) and (8.13), we have

$$F \Delta z = - \frac{Q^2}{2C^2} \Delta C. \quad (8.14)$$

This can also be written as

$$F \Delta z = - \frac{Q^2}{2C^2} \Delta C. \quad (8.15)$$

The force, of course, results from the attraction of the charges on the plates, but we see that we do not have to worry so much about how they are distributed; everything we need is taken care of in the capacity C .

It is easy to see how the idea is extended to conductors of any shape, and for other components of the force. In Eq. (8.14), we replace Δ by the component we are looking for, and we replace Δz by a small displacement in the corresponding direction. Or if we have an electrode with a pivot and we want to know the torque τ , we write the virtual work as

$$\Delta W = \tau \Delta \theta,$$

where $\Delta \theta$ is a small angular displacement. Of course, $\Delta(1/C)$ must be the change in $1/C$ which corresponds to $\Delta \theta$. We could, in this way, find the torque on the movable plates in a variable condenser of the type shown in Fig. 8-3.

Returning to the special case of a parallel-plate condenser, we can use the formula we derived in Chapter 6 for the capacity:

$$\frac{1}{C} = \frac{\epsilon_0 A}{d}, \quad (8.16)$$

where A is the area of each plate. If we increase the separation by Δz ,

$$\Delta \left(\frac{1}{C} \right) = \frac{\Delta z}{\epsilon_0 d}.$$

From Eq. (8.14) we get that the force between the plates is

$$F = \frac{Q^2}{2\epsilon_0 d^2}. \quad (8.17)$$

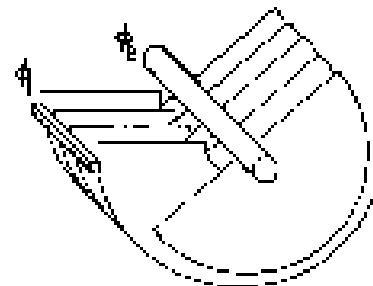


Fig. 8-3. What is the torque on a variable capacitor?

Let's look at Eq. (3.17) a little more closely, and see if we can tell how the force varies. To do this, let's integrate implicitly we write

$$Q = \sigma A.$$

Eq. (3.17) can be rewritten as

$$F = \frac{1}{2} Q \frac{\sigma}{\epsilon_0}.$$

Or, since the electric field between the plates is

$$E_s = \frac{\sigma}{\epsilon_0},$$

then

$$F = \frac{1}{2} Q E_s. \quad (3.18)$$

One would immediately guess that the force acting on one plate is the charge Q on the plate times the field acting on the charge. But we have a surprising answer of ourself. The reason is that E_s is not the field of the charges. If we imagine that the charge at the surface of the plate occupies a thin layer, as indicated in Fig. 3-1, the field will vary from zero at the outer boundary of the layer to E_s in the space outside of the plate. The average field acting on the surface charge is $E_s/2$. That is why the factor one-half is in Eq. (3.18).

You should notice that in computing the virtual work we have assumed that the charge on the conductor was constant. That it was not electrically connected to other objects, and so the total charge could not change.

Suppose we had imagined that the conductor was held at a constant potential difference as we made the virtual displacement. Then we should have taken

$$U = \frac{1}{2} C V^2$$

and in place of Eq. (R.15) we would have had

$$\Delta U = \frac{1}{2} C V^2 \Delta C,$$

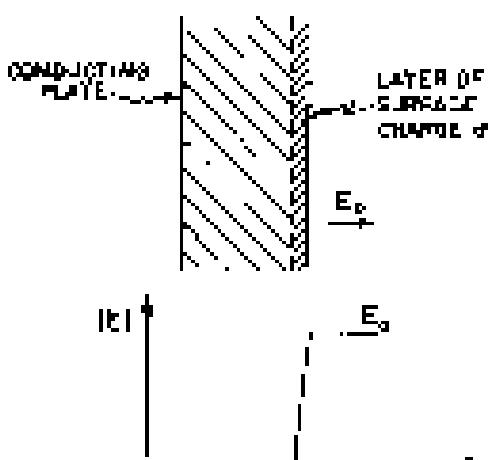
which gives a force equal in magnitude to the one in Eq. (3.18) (because $V = (1/C)$), but with the opposite sign! Surely the force between the condenser plates doesn't reverse in sign as we disassemble it from its charging source. Also, we know that two plates with opposite electrical charges must attract. The principle of virtual work has been incorrectly applied in the second case: we have not taken into account the virtual work done on the charging source. That is, to keep the potential constant at V as the capacity changes, a charge $V \Delta C$ must be supplied by a source of charge. But this charge is supplied at a potential V , so the work done by the electrical system which keeps the potential constant is $V^2 \Delta C$. The mechanical work δU plus this electrical work $V^2 \Delta C$ together make up the change in the total energy $\frac{1}{2} V^2 \Delta C$ of the condenser. Therefore $F \delta u$ is $-V^2 \Delta C$, as before.

3.3 The electrostatic energy of an ionic crystal

We now consider an application of the concept of electrostatic energy in atomic physics. We cannot easily measure the forces between atoms, but we are often interested in the energy differences between one atomic arrangement and another, as, for example, the energy of a chemical change. Since atomic forces are basically electrical, chemical energies are in large part just electrostatic energies.

Let's consider, for example, the electrostatic energy of an ionic lattice. An ionic crystal like NaCl consists of positive and negative ions which repel each other if they get too close. They attract sufficiently until they begin to touch; then there is a resultant force which goes up very rapidly if we try to push them closer together.

For our first approximation, therefore, we imagine a set of rigid spheres that represent the atoms in a salt crystal. The structure of the lattice has been determined by x-ray diffraction. It is a cubic lattice like a three-dimensional



checkerboard. Figure 8-5 shows a cross-sectional view. The spacing of the ions is 2.81 Å ($\sim 2.81 \times 10^{-8}$ cm).

If our picture of this system is correct, we should be able to check it by asking the following question: How much energy will it take to pull all these ions apart—that is, to separate the crystal completely into ions? This energy should be equal to the heat of vaporization of NaCl plus the energy required to dissociate the molecules into ions. This total energy to separate NaCl to ions is determined experimentally to be 7.92 electron volts per molecule. Using the conversion

$$1 \text{ ev} = 1.602 \times 10^{-19} \text{ joule},$$

and Avogadro's number for the number of molecules in a mole,

$$N_A = 6.02 \times 10^{23},$$

the energy of vaporization can also be given as

$$W = 7.92 \times 10^3 \text{ joules/mole}.$$

Physical chemists prefer to use energy units the kilocalorie, which is 4190 joules; so that 1 ev per molecule is 23 kilocalories per mole. A chemist would then say that the dissociation energy of NaCl is

$$W = 183 \text{ kcal/mole}.$$

Can we obtain this chemical energy theoretically by computing how much work it would take to pull apart the crystal? According to our theory, this work is the sum of the potential energies of all the pairs of ions. The easiest way to figure out this sum is to pick out a particular ion and compute its potential energy with each of the other ions. That will give us twice the energy per ion, because the energy belongs to the pair of charges. If we want the energy to be associated with one particular ion, we should take half the sum. But we really want the energy per molecule, which contains two ions, so that the sum we compute will give directly the energy per molecule.

The energy of an ion with one of its nearest neighbors is e^2/a , where $e^2 = q^2/4\pi\epsilon_0$, and a is the inter-ion spacing between ions. (We are considering monovalent ions.) This energy is 5.12 ev, which we already see is going to give us a result of the correct order of magnitude. But it is still a long way from the full sum of terms we need.

Let's begin by summing all the terms from the ions along a straight line. Considering that the ion marked Na in Fig. 8-5 is very special now, we shall consider first those ions on a horizontal line with it. There are two nearest Cl ions with negative charges, each at the distance a . Then there are two positive ions at the distance $2a$, etc. Calling the energy of this sum U_1 , we write

$$\begin{aligned} U_1 &= \frac{e^2}{a} \left(-\frac{1}{1} - \frac{2}{2} - \frac{2}{3} - \frac{2}{4} + \dots \right) \\ &= -\frac{2e^2}{a} \left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots \right). \end{aligned} \quad (8.19)$$

The series converges slowly, so it is difficult to evaluate numerically, but it is known to be equal to $-\ln 2$. So

$$U_1 = -\frac{2e^2}{a} \ln 2 = -1.389 \frac{e^2}{a}. \quad (8.20)$$

Now consider the next adjacent line of ions above. The nearest is negative and at the distance a . Then there are two positives at the distance $\sqrt{2}a$. The next pair are at the distance $\sqrt{3}a$, the next at $\sqrt{5}a$, and so on. So for the whole line we get the series

$$\frac{e^2}{a} \left(-\frac{1}{1} + \frac{2}{\sqrt{2}} + \frac{2}{\sqrt{3}} - \frac{2}{\sqrt{5}} + \dots \right). \quad (8.21)$$

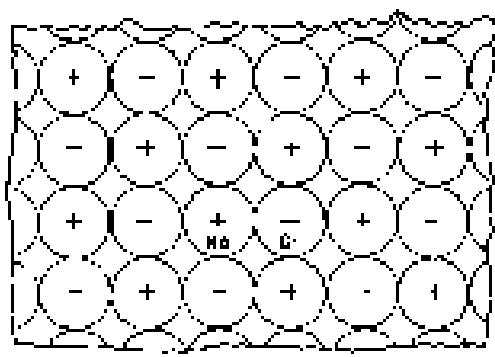


Fig. 8-5

Fig. 8-5. Cross section of a salt crystal on an atomic scale. The checkerboard arrangement of Na and Cl ions is the same in the two cross sections perpendicular to the one shown. (See Vol. I, Fig. 1-7.)

There are, over such lines; above, below, in front, and in back. Then there are the four lines which are the nearest lines on diagonals, and so on and on.

If you work patiently through for all the lines, and then take the sum, you find that the grand total is

$$U = 1.747 \frac{e^2}{a}$$

which is just somewhat more than what we obtained in (9.20) for the first line. Using $e^2/a = 5.12$ ev, we get

$$U = 8.94 \text{ ev.}$$

Our answer is about 10% above the experimentally observed energy. It shows that our idea that the whole lattice is held together by electrical Coulomb forces is fundamentally sound. This is the first time that we have obtained a specific property of a macroscopic substance from a knowledge of atomic physics. We will do much more later. The subject that tries to understand the behavior of bulk matter in terms of the laws of atomic behavior is called solid-state physics.

Now what about the error in our calculation? Why is it not exactly right? It is because of the repulsion between the ions at close distances. They are not perfectly rigid spheres, so when they are close together they are partly squashed. They are not very soft, so they squash only a little bit. Some energy, however, is used in deforming them, and when the ions are pulled apart this energy is released. The actual energy needed to pull the ions apart is a little less than the energy that we calculated; the repulsion helps in overcoming the electrostatic attraction.

Is there any way we can make an allowance for the contribution? We could if we knew the law of the repulsive force. We are not ready to analyze the details of this repulsive mechanism, but we can get some idea of its characteristics from atomic large-scale measurements. From a measurement of the compressibility of the whole crystal, it is possible to obtain a quantitative idea of the law of repulsion between the ions and therefore of its contribution to the energy. In this way it has been found that this contribution must be 1/9.4 of the contribution from the electrostatic attraction and, of course, of opposite sign. If we subtract this contribution from the pure electrostatic energy, we obtain 7.99 ev for the dissociation energy per molecule. It is much closer to the observed result of 7.92 ev, but still not in perfect agreement. There is one more thing we haven't taken into account: we have made no allowance for the kinetic energy of the crystal vibrations. If a correction is made for this effect, very good agreement with the experimental number is obtained. The ideas are then correct; the major contribution to the energy of a crystal like NaCl is electrostatic.

9-4 Electrostatic energy in nuclei

We will now take up another example of electrostatic energy in atomic physics, the electrical energy of atomic nuclei. Before we do this we will have to discuss some properties of the main forces (called nuclear forces) that build the protons and neutrons together in a nucleus. In the early days of the discovery of nuclei—and of the neutrons and protons that make them up—it was hoped that the law of the strong, nonelectrical part of the force between, say, a proton and another proton would have some simple law, like the inverse square law of electricity. For such was had determined this law of force, and the corresponding force between a proton and a neutron, and a neutron and a neutron, it would be possible to describe theoretically the complete behavior of these particles in nuclei. Therefore a big program was started for the study of the scattering of protons, in the hope of finding the law of force between them; but after thirty years of effort, nothing simple has emerged. A considerable knowledge of the force between proton and proton has been accumulated, but we find that the force is as complicated as it can possibly be.

What we mean by "as complicated as it can be" is that the force depends on as many things as it possibly can.

First, the force is not a simple function of the distance between the two protons. At large distances there is no attraction, but at closer distances there is a repulsion. The distance dependence is a complicated function, still imperfectly known.

Second, the force depends on the orientation of the protons' spins. The protons have a spin, and any two interacting protons may be spinning with like or against movements in the same direction or in opposite directions. And the force is different when the spins are parallel from what it is when they are antiparallel, as in (a) and (b) of Fig. 8-6. The difference is quite large; it is not a small effect.

Third, the force is considerably different when the separation of the two protons is in the direction parallel to their spins, as in (c) and (d) of Fig. 8-6, than it is when the separation is in a direction perpendicular to the spins, as in (e) and (f).

Fourth, the force depends, as it does in magnetism, on the velocity of the protons, only much more strongly than in magnetism. And this velocity-dependent force is not a relativistic effect; it is strong even at speeds much less than the speed of light. Furthermore, this part of the force depends on other things besides the magnitude of the velocity. For instance, when a proton is moving near another proton, the force is different when the orbital motion has the same direction of rotation as the spin, as in (e) of Fig. 8-6, than when it has the opposite direction of rotation, as in (f). This is called the "spin-orbit" part of the force.

The force between a proton and a neutron and between a neutron and a neutron are also equally complicated. To this day we do not know the machinery behind these forces—that is to say, any simple way of understanding them.

There is, however, one important way in which the nuclear forces are simpler than they could be. That is that the nuclear force between two neutrons is the same as the force between a proton and a neutron, which is the same as the force between two protons! If, in any nuclear situation, we replace a proton by a neutron (or vice versa), the nuclear interactions are not changed. The "fundamental reason" for this equality is not known, but it is an example of an important principle that can be extended also to the interaction laws of other strongly interacting particles—such as the π -mesons and the "strange" particles.

This fact is nicely illustrated by the locations of the energy levels in similar nuclei. Consider a nucleus like B^{11} (boron-eleven), which is composed of five protons and six neutrons. In the nucleus the eleven particles interact with one another in a most complicated dance. Now, there is one configuration of all the possible interactions which has the lowest possible energy; this is the normal state of the nucleus, and is called the ground state. If the nucleus is disturbed (for example, by being struck by a high-energy proton or other particle) it can be put into any number of other configurations, called excited states, each of which will have a characteristic energy that is higher than that of the ground state. In nuclear physics research, such as is carried on with Van de Graaff generators (for example, in Caltech's Kelling and Sloan Laboratories), the energies and other properties of these excited states are determined by experiment. The energies of the fifteen lowest-known excited states of B^{11} are shown in a two-dimensional graph on the left half of Fig. 8-7. The lowest horizontal line represents the ground state. The first excited state has an energy 2.14 Mev higher than the ground state, the next an energy 4.46 Mev higher than the ground state, and so on. The study of nuclear physics attempts to find an explanation for this rather complicated pattern of energies; there is as yet, however, no complete general theory of such nuclear energy levels.

If we replace one of the neutrons in B^{11} with a proton, we have the nucleus of an isotope of carbon, C^{12} . The energies of the lowest sixteen excited states of C^{12} have also been measured; they are shown in the right half of Fig. 8-7. (The broken lines indicate levels for which the experimental information is questionable.)

Looking at Fig. 8-7, we see a striking similarity between the patterns of the energy levels in the two nuclei. The first excited states are about 2 Mev above the ground states. There is a large gap of about 2.3 Mev to the second excited state, then a small jump of only 0.5 Mev to the third level. Again, between the fourth and fifth levels, a big jump; but between the fifth and sixth a tiny separation of the

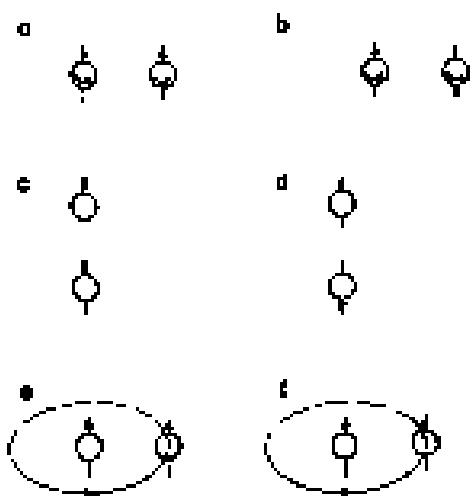


Fig. 8-6. The force between two protons depends on every possible parameter.

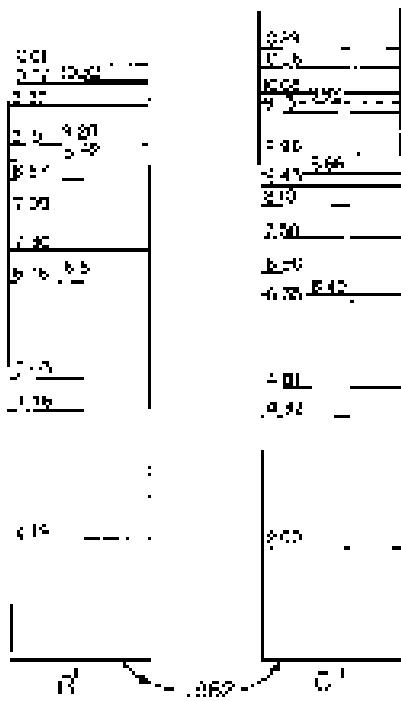


Fig. 8-7. The energy levels of B^{11} and C^{12} (energies in Mev). The ground state of C^{12} is 1.982 Mev higher than that of B^{11} .

order of 0.1 Mev. And so on. After about the tenth level, the correspondence seems to become bad, but can still be seen if the levels are labeled with their other defining characteristics—for instance, their angular momentum and what they do to lose their extra energy.

The striking similarity of the patterns of the energy levels of B^{11} and C^{11} is surely not just a coincidence. It must reveal some physical law. It shows, in fact, that even in the complicated situation in a nucleus, replacing a neutron by a proton makes very little change. This can mean only that the neutron-nucleus and proton-nucleus forces must be nearly identical. Only then would we expect the nuclear configurations with five protons and six neutrons to be the same as with six protons and five neutrons.

Notice that the properties of these two nuclei tell us nothing about the neutron-proton force; there are the same number of neutron-proton combinations in both nuclei. But if we compare two other nuclei, such as C^{14} , which has six protons and eight neutrons, with N^{14} , which has seven of each, we find a similar correspondence of energy levels. So we can conclude that the p-p, n-n, and p-n forces are identical in all their complexities. There is an unexpected principle in the laws of nuclear forces. Even though the force between each pair of nuclear particles is very complicated, the force between the three possible different pairs is the same.

But there are some small differences. The levels do not correspond exactly; also, the ground state of C^{11} has an absolute energy (its mass) which is higher than the ground state of B^{11} by 1.982 Mev. All the other levels are also higher in absolute energy by this same amount. So the forces are not exactly equal. But we know very well that the complete forces are not exactly equal; there is an electrostatic force between two protons because each has a positive charge, while between two neutrons there is no such electrical force. Can we perhaps explain the differences between B^{11} and C^{11} by the fact that the electrical interaction of the protons is different in the two cases? Perhaps so; the remaining minor differences in the levels are caused by electrical effects? Since the nuclear forces are so much stronger than the electrical force, electrical effects would have only a small perturbing effect on the energies of the levels.

In order to check this idea, or rather to find out what the consequences of this idea are, we first consider the difference in the ground-state energies of the two nuclei. To take a very simple model, we suppose that the nuclei are spheres of radius r (to be determined), containing Z protons. If we remember that a nucleus is like a sphere with uniform charge density, we would expect the electrostatic energy (from Eq. 8.7) to be

$$U = \frac{3}{5} \frac{(Zq_r)^2}{4\pi\epsilon_0 r}, \quad (8.22)$$

where q is the elementary charge of the proton. Since Z is five for B^{11} and six for C^{11} , their electrostatic energies would be different.

With such a small number of protons, however, Eq. (8.22) is not quite correct. If we compute the electrical energy between all pairs of protons, considered as points which we assume to be nearly uniformly distributed throughout the sphere, we find that in Eq. (8.22) the quantity Z^2 should be replaced by $Z(Z - 1)$, so the energy is

$$U = \frac{3}{5} \frac{Z(Z - 1)q^2}{4\pi\epsilon_0 r} = \frac{3}{5} \frac{Z(Z - 1)e^2}{r}. \quad (8.23)$$

If we knew the nuclear radius r , we could use (8.23) to find the electrostatic energy difference between B^{11} and C^{11} . But let's do the opposite; let's instead use the observed energy difference to compute the radius, assuming that the energy difference is all electrostatic in origin.

That is, however, not quite right. The energy difference of 1.982 Mev between the ground states of B^{11} and C^{11} includes the rest energies—that is, the energy $m c^2$ of all the particles. In going from B^{11} to C^{11} , we replace a neutron by a proton, which has less mass. So part of the energy difference is the difference in the rest energies of a neutron and a proton, which is 0.784 Mev. The difference,

to be accounted for by electrostatic energy, is thus more than 1.982 Mev; it is
 $1.982 + 0.784 = 2.766$ Mev.

Using this energy in Eq. (8.23), for the radius of either B^{11} or C^{11} we find

$$r = 3.12 \times 10^{-12} \text{ cm.} \quad (8.24)$$

Does this number have any meaning? To see whether it does, we should compare it with atomic size determinations of the radius of these nuclei. For example, we can make another measurement of the radius of a nucleus by seeing how it scatters beta particles. From such measurements it has been found, in fact, that the density of matter in all nuclei is nearly the same, i.e., their volumes are proportional to the number of particles they contain. If we let A be the number of protons and neutrons in a nucleus (a number very nearly proportional to its mass), it is found that its radius is given by

$$r = A^{1/3} r_0 \quad (8.25)$$

where

$$r_0 = 1.2 \times 10^{-12} \text{ cm.} \quad (8.26)$$

From these measurements we find that the radius of a B^{11} (or a C^{11}) nucleus is expected to be

$$r = (1.2 \times 10^{-12})(11)^{1/3} = 2.7 \times 10^{-12} \text{ cm.}$$

Comparing this result with (8.24), we see that our assumptions that the energy difference between B^{11} and C^{11} is electrostatic is fairly good; the discrepancy is only about 15% (not bad for our first nuclear computation!).

The reason for the discrepancy is probably the following. According to the current understanding of nuclei, an even number of nuclear particles—in the case of B^{11} , five neutrons together with five protons—make a kind of core; when one more particle is added to this core, it revolves around the outside to make a new spherical nucleus, rather than being absorbed. If this is so, we should have taken a different electrostatic energy for the additional proton. We should have taken the excess energy of C^{11} over B^{11} to be just

$$\frac{Z_p e^2}{4\pi \epsilon_0 r},$$

which is the energy needed to add one more proton to the outside of the core. This number is just 5/6 of what Eq. (8.21) predicts, so the new prediction for the radius is 5/6 of (8.24), which is in much closer agreement with what is directly measured.

We can draw two conclusions from this agreement. One is that the electrical laws appear to be working at dimensions as small as 10^{-12} cm. The other is that we have verified the remarkable coincidence that the nonelectrical part of the forces between proton and proton, neutron and neutron, and proton and neutron are all equal.

8-5 Energy in the electrostatic field

We now consider other methods of calculating electrostatic energy. They can all be derived from the basic relation Eq. (8.3), the sum, over all pairs of charges, of the mutual energies of each charge-pair. First we wish to write an expression for the energy of a charge distribution. As usual, we consider that each volume element dV contains the element of charge ρdV . Then Eq. (8.3) should be written

$$U = \frac{1}{2} \int_{\text{all space}} \frac{\rho(1)\rho(2)}{4\pi \epsilon_0 r_{12}} dV_1 dV_2. \quad (8.27)$$

Notice the factor $\frac{1}{2}$, which is introduced because in the double integral over dV_1 and dV_2 we have counted all pairs of charge elements twice. (There is no convenient way of writing an integral that keeps track of the pairs so that each pair is counted only once.) Next we notice that the integral over dV_2 in (8.27) is just the potential at (1). That is,

$$\int \frac{\rho(2)}{4\pi\epsilon_0 r_{12}} dV_2 = \phi(1),$$

so that (8.27) can be written as

$$U = \frac{1}{2} \int \rho(1)\phi(1) dV_1.$$

Or, since the point (2) no longer appears, we can simply write

$$U = \frac{1}{2} \int \rho\phi dV. \quad (8.28)$$

This equation can be interpreted as follows. The potential energy of the charge ρdV is the product of this charge and the potential at the same point. The total energy is therefore the integral over $d\rho dV$. But there is again the factor $\frac{1}{2}$. It is still required because we are counting energies twice. The mutual energies of two charges is the charge of one times the potential at it due to the other. Or, it can be taken as the second charge times the potential at it from the first. Thus for two point charges we would write

$$U = q_1\phi(1) = q_1 \frac{q_2}{4\pi\epsilon_0 r_{12}}$$

or

$$U = q_2\phi(1) = q_2 \frac{q_1}{4\pi\epsilon_0 r_{12}}.$$

Notice that we could also write

$$U = \frac{1}{2}(q_1\phi(1) + q_2\phi(1)). \quad (8.29)$$

The integral in (8.28) corresponds to the sum of both terms in the brackets of (8.29). That is why we need the factor $\frac{1}{2}$.

An interesting question is: Where is the electrostatic energy located? One might also ask: Who cares? What is the meaning of such a question? If there is a pair of interacting charges, the combination has a certain energy. Do we need to say that the energy is located at one of the charges or the other, or in both, or in between? These questions may not make sense because we really know only that the total energy is conserved. The idea that the energy is located somewhere is not necessary.

Yet suppose that it did make sense to say, in general, that energy is located at a certain place, as it does for heat energy. We might then extend our principle of the conservation of energy with the idea that if the energy in a given volume changes, we should be able to account for the change by the flow of energy into or out of that volume. You realize that our early statement of the principle of the conservation of energy is still perfectly all right if some energy disappears at one place and appears somewhere else far away without anything passing (that is, without any special phenomena occurring) in the space between. We are, therefore, now discussing an extension of the idea of the conservation of energy. We might call it a principle of the local conservation of energy. Such a principle would say that the energy in any given volume changes only by the amount that flows into or out of the volume. It is indeed possible that energy is conserved locally in such a way. If it is, we would have a much more detailed law than the simple statement of the conservation of total energy. It does turn out that in nature energy is conserved locally. We can find formulas for where the energy is located and how it travels from place to place.

There is also a physical reason why it is imperative that we be able to say where energy is located. According to the theory of gravitation, all mass is a source

of gravitational attraction. We also know, by $E = mc^2$, that mass and energy are equivalent. All energy is, therefore, a source of gravitational force. If we could not locate the energy, we could not locate all the mass. We would not be able to say where the sources of the gravitational field are located. The theory of gravitating would be incomplete.

If we restrict ourselves to electrostatics there is really no way to tell where the energy is located. The complete Maxwell equations of electrodynamics give us much more information (although even then the answer is, strictly speaking, not unique). We will therefore discuss this question in detail again in a later chapter. We will give you now only the result for the particular case of electrostatics. The energy is located in space, where the electric field is. This seems reasonable because we know that when charges are accelerated they radiate electric fields. We would like to say that when light or radio waves travel from one point to another, they carry their energy with them. But there are no charges in the waves. So we would like to locate the energy where the electromagnetic field is and not at the charges from which it comes. We thus describe the energy, not in terms of the charges, but in terms of the fields they produce. We can, in fact, show that Eq. (8.28) is numerically equal to

$$U = \frac{\epsilon_0}{2} \int E \cdot E dV. \quad (8.30)$$

We can then interpret this formula as saying that when an electric field is present, there is located in space an energy whose density (energy per unit volume) is

$$\epsilon = \frac{\epsilon_0}{2} E \cdot E = \frac{\epsilon_0 E^2}{2}. \quad (8.31)$$

This idea is illustrated in Fig. 8-8.

To show that Eq. (8.30) is consistent with our laws of electrodynamics, we begin by introducing into Eq. (8.28) the relation between ρ and ϕ that we obtained in Chapter 6:

$$\rho = -\epsilon_0 \nabla^2 \phi.$$

We get

$$U = -\frac{\epsilon_0}{2} \int \phi \nabla^2 \phi dV. \quad (8.32)$$

Writing out the components of the integrand, we see that

$$\begin{aligned} \phi \nabla^2 \phi &= \phi \left(\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} \right) \\ &= \frac{\partial}{\partial x} \left(\phi \frac{\partial \phi}{\partial x} \right) - \left(\frac{\partial \phi}{\partial x} \right)^2 + \frac{\partial}{\partial y} \left(\phi \frac{\partial \phi}{\partial y} \right) - \left(\frac{\partial \phi}{\partial y} \right)^2 + \frac{\partial}{\partial z} \left(\phi \frac{\partial \phi}{\partial z} \right) - \left(\frac{\partial \phi}{\partial z} \right)^2 \\ &= \nabla \cdot (\phi \nabla \phi) - (\nabla \phi) \cdot (\nabla \phi). \end{aligned} \quad (8.33)$$

Our energy integral is then

$$U = \frac{\epsilon_0}{2} \int (\nabla \phi) \cdot (\nabla \phi) dV = \frac{\epsilon_0}{2} \int \nabla \cdot (\phi \nabla \phi) dV.$$

We can use Gauss' theorem to change the second integral into a surface integral:

$$\int_{V_{ext}} \nabla \cdot (\phi \nabla \phi) dV = \int_{S_{ext}} (\phi \nabla \phi) \cdot n d\sigma. \quad (8.34)$$

We evaluate the surface integral in the case that the surface goes to infinity (so the volume integrals become integrals over all space), supposing that all the charges are located within some finite distance. The simple way to proceed is to take a spherical surface of enormous radius R whose center is at the origin of coordinates. We know that when we are very far away from all charges, ϕ varies as $1/R$ and $\nabla \phi$ as $1/R^2$. (Both will decrease even faster with R if there are no



Fig. 8-8. Each volume element $dV = dx dy dz$ in an electric field contains the energy $(\epsilon_0/2)E^2 dV$.

charge in the distribution is zero.) Since the surface area of the large sphere increases as R^2 , we see that the surface integral falls off as $(1/R)(1/R^2)R^2 = (1/R)$ as the radius of the sphere increases. So if we include all space in our integration ($R \rightarrow \infty$), the surface integral goes to zero and we have that

$$U = \frac{\epsilon_0}{2} \int_{\text{all space}} (\nabla \phi) \cdot (\nabla \phi) dV = \frac{\epsilon_0}{2} \int_{\text{all space}} E \cdot E dV. \quad (8.35)$$

We see that it is possible for us to represent the energy of any charge distribution as being the integral over all energy density located in the field.

8-6 The energy of a point charge

Our new relation, Eq. (8.35), says that even a single point charge q will have some electrostatic energy. In this case, the electric field is given by

$$E = \frac{q}{8\pi\epsilon_0 r^2} \hat{r}.$$

So the energy density at the distance r from the charge is

$$\frac{\epsilon_0 E^2}{2} = \frac{q^2}{32\pi\epsilon_0 r^4}.$$

We can take for an element of volume a spherical shell of thickness dr and area $4\pi r^2$. The total energy is

$$U = \int_{r=0}^{\infty} \frac{q^2}{8\pi\epsilon_0 r^4} 4\pi r^2 dr = -\frac{q^2}{8\pi\epsilon_0} \frac{1}{r} \Big|_{r=0}^{\infty}. \quad (8.36)$$

Now the limit as $r \rightarrow \infty$ gives no difficulty. But for a point charge we are supposed to integrate down to $r = 0$, which gives an infinite integral. Equation (8.35) says that there is an infinite amount of energy in the field of a point charge, although we began with the idea that there was energy only between point charges. In our original energy formula for a collection of point charges (Eq. 8.3), we did not include any interaction energy of a charge with itself. What has happened is that when we went over to a continuous distribution of charge in Eq. (8.27), we counted the energy of interaction of every infinitesimal charge with all other infinitesimal charges. The same account is included in Eq. (8.35), so when we apply it to a finite point charge, we are including the energy it would take to assemble that charge from infinitesimal parts. You will notice, in fact, that we would also get the result in Eq. (8.36) if we used our expression (8.11) for the energy of a charged sphere and let the radius tend toward zero.

We must conclude that the idea of treating the energy in the field is inconsistent with the assumption of the existence of point charges. One way out of the difficulty would be to say that elementary charges, such as an electron, are not points but are really small distributions of charge. Alternatively, we could say that there is something wrong in our theory of electricity at very small distances, or with the idea of the local conservation of energy. These are difficulties with older point of view. These difficulties have never been overcome; they exist to this day. Sometime later, when we have discussed some additional ideas, such as the interaction in an electromagnetic field, we will give a more complete account of these fundamental difficulties in our understanding of nature.

Electricity in the Atmosphere

9-1 The electric potential gradient of the atmosphere

On an ordinary day over flat closed country, or over the sea, as one goes upward from the surface of the ground the electric potential increases by about 100 volts per meter. Thus there is a vertical electric field E of 100 volts/m in the air. The sign of the field corresponds to a negative charge on the earth's surface. This means that midday the potential at the height of your nose is 200 volts higher than the potential at your feet! You might ask: "Why don't we just stick a pair of electrodes one in the air one meter apart and use the 100 volts to power our electric lights?" Or you might wonder: "If there is really a potential difference of 200 volts between my nose and my feet, why is it I don't get a shock when I go out into the street?"

We will answer the second question first. Your body is a relatively good conductor. If you are in contact with the ground, you and the ground will tend to make one equipotential surface. Ordinarily, the equipotentials are parallel to the surface, as shown in Fig. 9-1(a), but when you are there, the equipotentials are distorted, and the field looks somewhat as shown in Fig. 9-1(b). So you still have very nearly zero potential difference between your head and your feet. There are no charges that come from the earth to your head, changing the field. Some of them may be discharged by ions collected from the air, but the current of these is very small because air is a poor conductor.

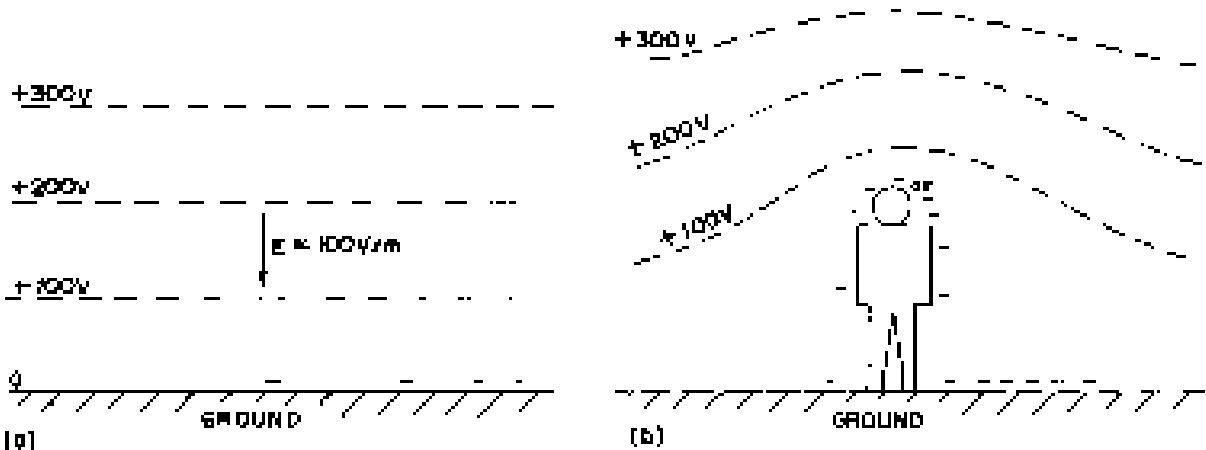


Fig. 9-1. (a) The potential distribution above the earth. (b) The potential distribution near a man on open flat ground.

How can we measure such a field if the field is changed by putting something there? There are several ways. One way is to place an insulated conductor at some distance above the ground and leave it there until it is at the same potential as the air. If we leave it long enough, the very small conductivity in the air will let the charges leak off (or move the conductor until it comes to the potential of its world). Then we can bring it back to the ground, and measure the shift of its potential as we do so. A faster way is to let the conductor be a bucket of water with a small leak. As the water drops out, it carries away any excess charges and the bucket will approach the same potential as the air. (The charges, as you know, reside on the surface, and as the drops come off "pieces of surface" break off.) We can measure the potential of the bucket with an electrometer.

9-2 The electric potential gradient of the atmosphere

9-3 Electric currents in the atmosphere

9-4 The mechanism of charge separation

9-5 The mechanism of charge separation

9-6 Lightning

Reference: Chapman, J. Alan, *Atmospheric Electricity*, Pergamon Press, London (1957).

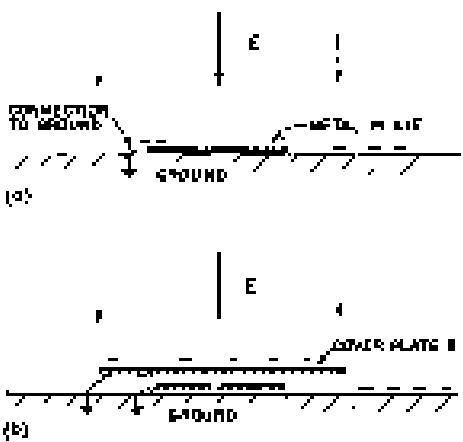


Fig. 9-2. (a) A grounded metal plate will have the same surface charge as the earth. (b) If the plate is connected with a grounded conductor it will have no surface charge.

There is another way to directly measure the potential gradient. Since there is an electric field, there is a surface charge on the earth ($\sigma = \epsilon_0 E$). If we place a flat metal plate at the earth's surface and ground it, negative charges appear on it (Fig. 9-2a). If this plate is now covered by another grounded conducting cover *B*, the charges will appear on the cover, and there will be no charges on the original plate *A*. If we measure the charge that flows from plate *A* to the ground (by, say, a galvanometer in the grounding wire) as we cover it, we can find the surface charge density that was there, and therefore also find the electric field.

Horing suggested how we can measure the electric field in the atmosphere, we now continue our description of it. Measurements show, first of all, that the field continues to exist, but gets weaker, as one goes up to high altitudes. By about 50 kilometers, the field is very small, so most of the potential change (the integral of *E*) is at lower altitudes. The total potential difference from the surface of the earth to the top of the atmosphere is about 400,000 volts.

9-2 Electric currents in the atmosphere

Another thing that can be measured, in addition to the potential gradient, is the current in the atmosphere. The current density is small—about 10 micromilli-amperes/cm² square meter parallel to the earth. The air is evidently not a perfect insulator, and because of this conductivity, a small current—caused by the electric field we have just been describing—goes from the sky down to the earth.

Why does the atmosphere have conductivity? Here and there among the air molecules there is an ion—a molecule of oxygen, say, which has acquired an extra electron, or perhaps lost one. These ions are not steady us single molecules; because of their electric field they usually accumulate a few other molecules around them. Each ion then becomes a little lump which, along with other lumps, drifts in the field—moving slowly upward or downward—making the observed current. Where do the ions come from? It was first guessed that the ions were produced by the radioactivity of the earth. (It was known that the radiation from radioactive materials would make air conducting by ionizing the air molecules.) Particles like β -rays coming out of the atomic nuclei are moving so fast that they tear electrons from the atoms, leaving ions behind. This would imply, of course, that if we were to go to higher altitudes, we should find less ionization, because the radioactivity is all in the dirt on the ground—in the traces of radium, uranium, potassium, etc.

To test this theory, some physicists carried an experiment up in balloons to measure the ionization of the air (Tess, in 1912) and discovered that the opposite was true—the ionization per unit volume increased with altitude! (The apparatus was like that of Fig. 9-3. The two plates were charged periodically to the potential *V*. Due to the conductivity of the air, the plates slowly discharged; the rate of discharge was measured with the electrometer.) This was a most mysterious result—the most dramatic finding in the entire history of atmospheric electricity. It was so dramatic, in fact, that it required a branching off of an entirely new subject—cosmic rays. Atmospheric electricity itself remained less dramatic. Ionization was evidently being produced by something from outside the earth: the investigation of this source led to the discovery of the cosmic rays. We will not discuss the subject of cosmic rays here, except to say that they maintain the supply of ions. Although the ions are being swept away all the time, new ones are being created by the cosmic-ray particles coming from the outside.

To be positive, we must say that besides the ions made of molecules, there are also other kinds of ions. Tiny pieces of dirt, like extremely fine bits of dust, float in the air and become charged. They are sometimes called "nuclei." For example, when a wave breaks in the sea, little bits of spray are thrown into the air. When one of these drops evaporates, it leaves an infinitesimal crystal of NaCl floating in the air. These tiny crystals can then pick up charges and become ions; they are called "tiny ions."

The small ions—these turned by cosmic rays—are the most mobile. Because they are so small, they move rapidly through the air—with a speed of about 1 g-t

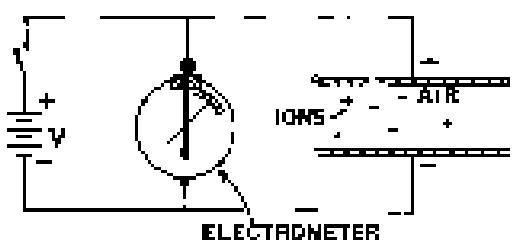


Fig. 9-3. Measuring the conductivity of air due to the motion of ions.

cm/sec in a field of 100 volts/meter, or 1 volt/cm. The much bigger and heavier ions move much more slowly. It turns out that if there are many "nuclei," they will pick up the charges from the small ions. Thus, since the "large ions" move so slowly in a field, the total conductivity is reduced. The conductivity of air, therefore, is quite variable, since it is very sensitive to the amount of "dirt" there is in it. There is much more of such dirt over land—where the winds can blow up dust or where man, through air blends of pollution into the air—than there is over water. It is not surprising that from day to day,风 can move to movement. From place to place, the conductivity near the earth's surface varies enormously. The voltage gradient observed at any particular place on the earth's surface also varies greatly because roughly the same current flows down from high altitudes in different places, and the varying conductivity near the earth results in a varying voltage gradient.

The conductivity of the air due to the drifting of ions also increases rapidly with altitude. For two reasons. First of all, the ionization from cosmic rays increases with altitude. Secondly, as the density of air goes down, the mean free path of the ions increases, so that they can travel farther in the electric field before they have a collision--resulting in a rapid increase of conductivity as one goes up.

Although the electric current-density in the air is only a few microamperes per square meter, there are very many square meters on the earth's surface. The total electric current reaching the earth's surface at any time is very nearly constant at 1800 amperes. This current, of course, is "positive"—it carries plus charges to the earth. So we have a voltage supply of 400,000 volts with a current of 1800 amperes—a power of 700 megawatts!

With such a large current coming down, the negative charge on the earth should soon be discharged. In fact, it should take only about half an hour to discharge the entire earth. But the atmospheric electric field has already lasted more than a half-hour since its discovery. How is it maintained? What maintains the voltage? And between what and the earth? There are many questions.

The earth is negative, and the potential in the air is positive. If you go high enough, the conductivity is so great that horizontally there is no more chance for voltage variations. The air, for the scale of times that we are talking about, becomes effectively a conductor. This occurs at a height in the neighborhood of 50 kilometers. This is not as high as what is called the "ionosphere," in which there are very large numbers of ions produced by photoelectricity from the sun. Nevertheless, for our discussions of atmospheric electricity, the air becomes sufficiently conductive at about 50 kilometers that we can imagine that there is practically a perfect conducting surface at this height, from which the currents come down. Our picture of the situation is shown in Fig. 9-4. The problem is: How is the positive charge maintained there? How is it pumped back? Because if it comes down to the earth, it has to be pumped back somehow. That was one of the greatest puzzles of atmospheric electricity for quite a while.

Each piece of information we can get should give a clue or, at least, tell you something about it. Here is an interesting phenomenon: If we measure the current (which is more stable than the potential gradient) over the sea, for instance, or in careful conditions, and average very carefully so that we get rid of the irregularities, we discover that there is still a daily variation. The average of many measurements over the oceans has a variation with time roughly as shown in Fig. 9-5. The current varies by about -15 percent, and it is largest at 7:00 p.m. in London. The strange part of the thing is that no matter where you measure the current—in the Atlantic Ocean, the Pacific Ocean, or the Arctic Ocean—it is at its peak value when the clocks in London say 7:00 p.m.! All over the world the current is at its maximum at 7:00 p.m. London time and it is at a minimum at 4:00 a.m. London time. In other words, it depends upon the absolute time on the earth, not upon the local time at the place of observation. In one respect this is not mysterious; it checks with our idea that there is a very high conductivity laterally at the top, because that makes it impossible for the voltage difference from the ground to the top to vary locally. Any potential variations should be worldwide, as indeed they are. What we now know, therefore, is that the voltage at the "top" surface is dropping and rising by 15 percent with the absolute time on the earth.

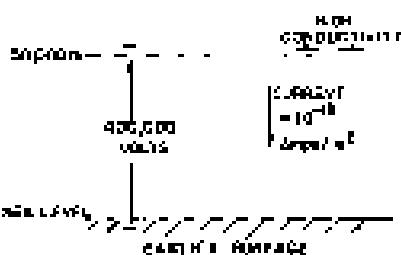


Fig. 9-4. Typical electrical conditions in a clear atmosphere.

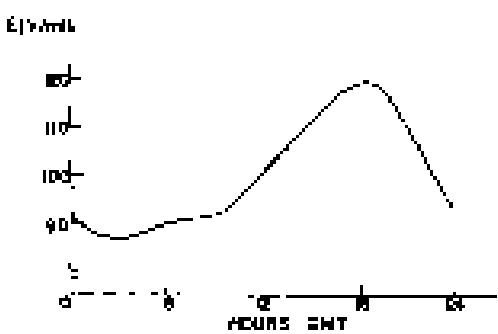


Fig. 9-5. The average daily variation of the atmospheric potential gradient on a clear day over the oceans, referred to Greenwich Time.

9-3 Origin of the atmospheric currents

We must now talk about the source of the large negative currents which must be flowing from the "top" to the surface of the earth to keep charging it up negatively. Where are the batteries that do this? The "battery" is shown in Fig. 9-3. It is the thunderstorm and its lightning. I assume that the bolts of lightning is a sort "discharge" the generator we have been talking about (or you might at first guess). Lightning storms carry negative charges to the earth. When a lightning bolt strikes, it carries it's negative charges to the earth in large amounts. It is the thunderstorms throughout the world that are charging the earth with an average of .800 amperes, which is then being discharged through regions of fair weather.

There are about 1,800 thunderstorms per day all over the earth, and we can think of them as batteries pumping the electricity to the upper layer and down to us by the voltage difference. Then take into account the geography of the earth. There are thunderstorms in the savannah in Brazil, tropical thunderstorms in Africa, and so forth. People have made estimates of how much lightning is striking world wide at any time, and perhaps needless to say, their estimates more or less agree with the voltage difference measurements: the total amount of thunderstorm activity is highest on the whole earth at about 7:00 p.m. in London. However, the thunderstorm estimates are very difficult to make out were made only after it was known that the sunspot should have occurred. These things are very difficult, because we don't have enough observers on the seas and over all parts of the world to count the number of the infections accurately. But these people who think they "got it right" obtain the result, that there is a peak in the activity at 7:00 p.m. Greenwich Mean Time.



Fig. 9-3. The mechanism that generates the atmospheric electric field. (Photo by William L. Widmayer.)

In order to understand how these batteries work, we will look at a thunderstorm in detail. What is going on inside a thunderstorm? We will describe this insofar as it is known. As we get into this marvelous phenomenon of real nature—instead of the idealized spheres of perfect conductors inside of other spheres that we can solve so neatly—we discover that we don't know very much. Yet it is really quite exciting. Anyone who has been in a thunderstorm has enjoyed it, or has been frightened, or at least has had some emotion. And in those places in nature where we get an emotion, we find that there is generally a corresponding complexity and mystery about it. It is not going to be possible to describe exactly how a thunderstorm works, because we do not yet know very much. But we will try to describe a little bit about what happens.

9-4 Thunderstorms

In the first place, an ordinary thunderstorm is made up of a number of "cells" fairly close together, but almost independent of each other. So it is best to analyze one cell at a time. By a "cell" we mean a region with a limit area in the horizontal direction in which all of the basic processes occur. Usually there are several cells side by side, and in each one above the same thing is happening, although perhaps with a different timing. Figure 9-7 indicates in an idealized fashion what such a cell looks like in the early stage of the thunderstorm. It turns out that in a certain place in the air, under certain conditions which we shall describe, there is a general rising of the air, with higher and higher velocities near the top. As the warm, moist air at the bottom rises, it cools and condenses. In the figure the little crosses indicate snow and the dots indicate rain, but because the updraft currents are great enough and the drops are small enough, the snow and rain do not come down at this stage. This is the beginning stage, and not the real thunderstorm yet—in the sense that we don't have anything happening at the ground. At the same time that the warm air rises, there is an entrainment of air from the sides—an imperfect point which was neglected for many years. Thus it is not just the air from below which is rising, but also a certain amount of other air from the sides.

Why does the air rise like this? As you know, when you go up in altitude the air is colder. The ground is heated by the sun, and the re-radiation of heat to the sky comes from water vapor high in the atmosphere; so at high altitude the air is cold—very cold—whereas lower down it is warm. You may say, "Then it's very simple. Warm air is lighter than cold; therefore the combination is automatically unstable and the warm air rises." Of course, if the temperature is different at different heights, the air is unstable thermodynamically. Let it itself fall slowly long, the air would all come to the same temperature. But it is not left to itself; the sun is always shining (during the day). So the problem is indeed not one of thermodynamic equilibrium, but of mechanical equilibrium. Suppose we plot, as in Fig. 9-8—the temperature of the air against height above the ground. In ordinary circumstances we would get a decrease along a curve like the one labeled (a); as the height goes up, the temperature goes down. How can the atmosphere be stable? Why doesn't the hot air below simply rise up into the cold air? The answer is this: if the air were to go up, its pressure would go down, and if we consider a particular parcel of air going up, it would be expanding adiabatically. (There would be no heat coming in or out because in the large dimensions considered here, there isn't time for much heat flow.) Thus the parcel of air would cool as it rises. Such an adiabatic process would give a temperature-height relationship like curve (b) in Fig. 9-8. Any air which rose from below would be cooler than the environment it goes into. Thus there is no reason for the hot air below to rise; if it were to rise, it would cool to a lower temperature than the air already there, would be heavier than the air there, and would just want to come down again. On a good, bright day with very little humidity there is a certain rate at which the temperature in the atmosphere falls, and this rate is, in general, lower than the "maximum stable gradient," which is represented by curve (b). The air is in stable mechanical equilibrium.

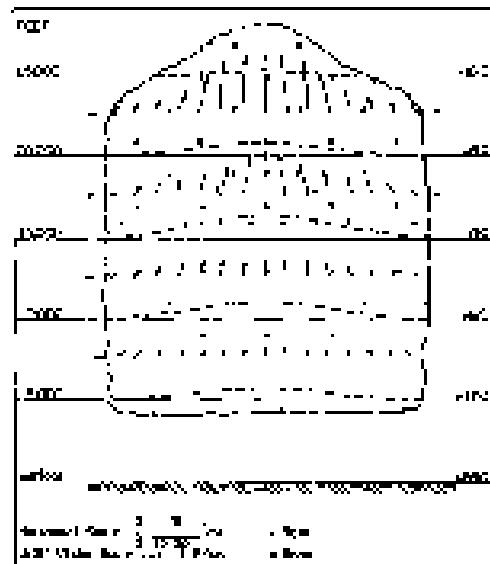


Fig. 9-7. A thunderstorm cell in the early stages of development. [From U.S. Department of Commerce Weather Bureau Report, June 1949.]

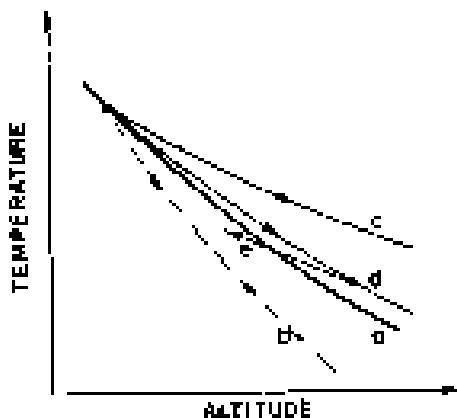


Fig. 9-8. Atmospheric temperature. (a) Static atmosphere; (b) adiabatic cooling of dry air; (c) adiabatic cooling of wet air; (d) wet air with some mixing of ambient air.

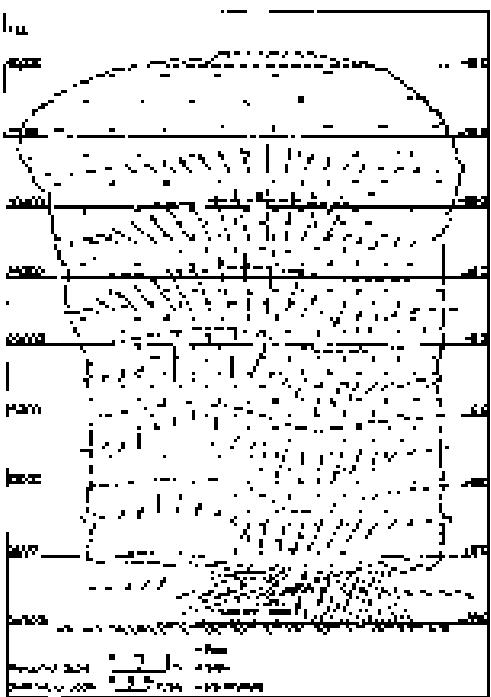


Fig. 9-9. A mature thunderstorm cell.
[From U.S. Department of Commerce Weather Bureau Report, June 1949.]

On the other hand, if we think of a parcel of air that contains a lot of water vapor being carried up into the air, its adiabatic cooling curve will be different. As it expands and cools, the water vapor in it will condense, and the condensing water will liberate heat. Moist air, therefore, does not cool nearly as much as dry air does. So if air that is wetter than the average starts to rise, its temperature will follow a curve like (c) in Fig. 9-8. It will cool off somewhat, but will still be warmer than the surrounding air at the same level. If we have a region of warm moist air and something starts it rising, it will always find itself lighter and warmer than the air around it and will continue to rise until it gets to successive heights. This is the mechanism that makes the air in the thunderstorm cell rise.

For many years the thunderstorm cell was explained simply in this manner, but then measurements showed that the temperature of the cloud at different heights was not nearly as high as indicated by curve (c). The reason is that as the moist air "bubble" goes up, it carries air from the environment and it cools off by it. The temperature-versus-height curve looks more like curve (d), which is much closer to the original curve (a) than to curve (c).

After the convection just described gets under way, the cross section of a thunderstorm cell looks like Fig. 9-9. We have what is called a "mother" thunderstorm. There is a very rapid updraft which, in this stage, goes up to about 10,000 to 15,000 meters—sometimes even much higher. The thunderheads, with their condensation, climb way up out of the general cloud bank, carried by an updraft that is usually about 60 miles an hour. As the water vapor is forced up and condenses, it forms tiny drops which are rapidly cooled to temperatures below zero degrees. They should freeze, but do not freeze immediately—they are "supercooled." Water and other liquids will usually cool well below their freezing points before crystallizing if there are no "nuclei" present to start the crystallization process. Only if there is some small pieces of material present, like a tiny crystal of NaCl, will the water drop freeze into a little piece of ice. Then the equilibrium is such that the water drops evaporate and the ice crystals grow. Thus at a certain point there is a rapid disappearance of the water and a rapid buildup of ice. Also, there may be direct collisions between the water drops and the ice—collisions in which the supercooled water droplets stick to the ice crystals, which causes it to suddenly crystallize. So at a certain point in the cloud expansion there is a rapid accumulation of large ice particles.

When the ice particles are heavy enough, they begin to fall through the rising air—they get too heavy to be supported any longer by the updraft. As they come down, they draw a little air with them and start a downdraft. And surprisingly enough, it is easy to see that once the downdraft is started, it will maintain itself. The air now drives itself down!

Notice that the curve (d) in Fig. 9-8 for the actual distribution of temperature in the cloud is not as steep as curve (c), which applies to wet air. So if we have wet air falling, its temperature will drop with the slope of curve (c) and will go below the temperature of the environment if it goes down far enough, as indicated by curve (e) in the figure. The moment it does this, it is denser than the environment and begins to fall rapidly. You say, "That is perpetual motion. First, you argue that the air should rise, and when you have it up there, you argue equally well that the air should fall!" But it isn't perpetual motion. When the situation is unstable and the warm air should rise, then clearly something has to replace the warm air. It is equally true that cold air coming down would energetically replace the warm air, but you realize that what is coming down is not the original air. The early arguments, that had a particular cloud without entrainment going up and then coming down, had some kind of a puzzle. They needed the twin to maintain the downdraft—an argument which is hard to believe. As soon as you realize that there is a lot of original air mixed in with the rising air, the thermodynamic argument shows that there can be a descent of the cold air which was originally at some great height. This explains the picture of the active thunderstorm sketched in Fig. 9-9.

As the air comes down, rain begins to come out of the bottom of the thunderstorm. In addition, the relatively cold air spreads out after it arrives at the earth's surface. So just before the rain comes there is a certain little cold wind that gives

is a forewarning of the oncoming storm. In the storm itself there are rapid and irregular gusts of air, there is an enormous turbulence in the clouds, and so on. But basically we have an updraft, then a downdraft—in general, a very complicated process.

The moment at which precipitation starts is the same moment that the large downdraft begins and is the same moment, in fact, when the electrical phenomena arise. Before we describe lightning, however, we can finish the story by looking at what happens to the thunderstorm cell after about one-half an hour to an hour. The cell looks as shown in Fig. 9-10. The updraft stops because there is no longer enough warm air to maintain it. The downward precipitation continues for a while, the last little bits of water come out, and things get quieter and quieter—although there are small ice crystals left way up in the air. Because the winds at very great altitude are in different directions, the top of the cloud visually spreads into an anvil shape. The cell comes to the end of its life.

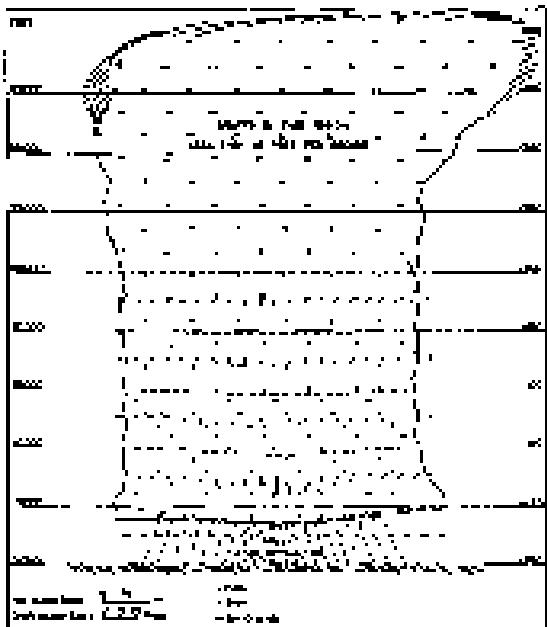


Fig. 9-10. The late phase of a thunderstorm cell. [From U.S. Department of Commerce Weather Bureau Report, June 1949.]

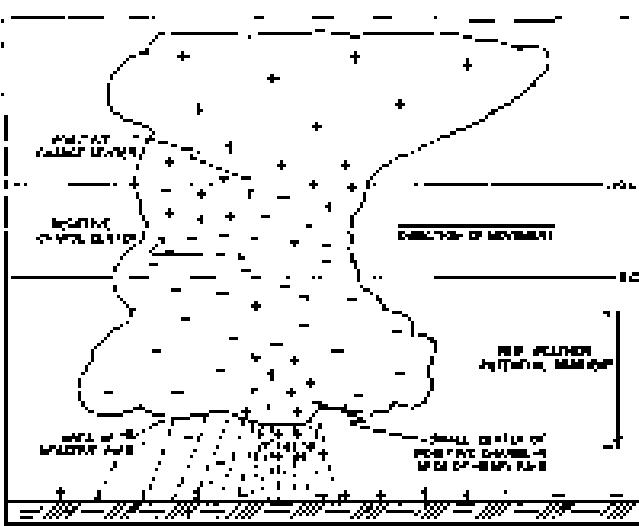


Fig. 9-11. The distribution of electrical charges in a mature thunderstorm cell. [From U.S. Department of Commerce Weather Bureau Report, June 1949.]

9-5 The mechanism of charge separation

We want now to discuss the most important aspect for our purposes—the development of the electrical charges. Experiments of various kinds—including flying airplanes through thunderstorms (the pilots who do this are brave ones!)—tell us that the charge distribution in a thunderstorm cell is something like that shown in Fig. 9-11. The top of the thunderstorm has a positive charge, and the bottom a negative one—except for a small local region of positive charge in the bottom of the cloud, which has caused everybody a lot of worry. No one seems to know why it is there, how important it is—whether it is a secondary effect of the positive rain coming down, or whether it is an essential part of the machinery. Things would be much simpler if it weren't there. Anyway, the predominantly negative charge at the bottom and the positive charge at the top have the correct sign for the battery needed to drive the earth negative. The positive charges are 6 or 7 kilometers up in the air, where the temperature is about -20°C , whereas the negative charges are 3 or 4 kilometers high, where the temperature is between zero and 10°C .

The charge at the bottom of the cloud is large enough to produce potential differences of 30, 000, 000 volts between the cloud and the earth—much bigger than the 0.4 million volts from the "sky" to the ground in a clear

atmosphere. These large voltages break down the air and create giant arc discharges. When the breakdown occurs the negative charges at the bottom of the thunderstorm are carried down to the earth in the lightning strokes.

Now we will describe in some detail the character of the lightning. First of all, there are large voltage differences around, so that the air breaks down. There are lightning strokes between one piece of a cloud and another piece of a cloud, or between one cloud and another cloud, or between a cloud and the earth. In each of the independent discharge flashes—the kind of lightning strokes you see—there are approximately 20 or 30 coulombs of charge brought down. One question is: How long does it take for the cloud to regenerate the 20 or 30 coulombs which are taken away by the lightning bolt? This can be seen by measuring, far from a cloud, the electric field produced by the cloud's dipole moment. In such measurements you see a sudden decrease in the field when the lightning strikes, and then an exponential return to the previous value with a time constant which is slightly different for different cases but which is in the neighborhood of 5 seconds. It takes a thundercloud only 5 seconds after each lightning stroke to build its charge up again. That doesn't necessarily mean that another strike is going to occur in exactly 5 seconds every time, because, otherwise, the geometry is changed, and so on. The strokes occur more or less irregularly, but the important point is that it takes about 5 seconds to recreate the original condition. Thus there are approximately 4 amperes of current in the generating machine of the thunderstorm. This means that any model made to explain how this storm generates its electricity must be one with plenty of juice—it must be a big, rapidly operating device.

Before we go further we shall consider something which is almost certainly completely irrelevant, but nevertheless interesting, because it does show the effect of an electric field on water drops. We say that it may be irrelevant because it relates to an experiment one can do in the laboratory with a stream of water to show the rather striking effects of the electric field on drops of water. In a thunderstorm there is no stream of water; there is a cloud of condensing ice and drops of water. So the question of the mechanisms at work in a thundercloud is probably not at all related to what you can see in the simple experiment we will describe. If you take a small nozzle connected to a water faucet and direct it upward at a steep angle, as in Fig. 9-12, the water will come out in a fine stream that eventually breaks up into a spray of fine drops. If you now put an electric field across the stream at the nozzle (by bringing up a charged rod, for example), the form of the stream will change. With a weak electric field you will find that the stream breaks up into a smaller number of larger-sized drops. But if you apply a stronger field, the stream breaks up into many, many fine drops—smaller than before.* With a weak electric field there is a tendency to inhibit the breaking of the stream into drops. With a stronger field, however, there is an increase in the tendency to separate into drops.

The explanation of these effects is probably the following. If we have the stream of water coming out of the nozzle and we put a small electric field across it one side of the water gets slightly positive and the other side gets slightly negative. Then, when the stream breaks, the drops on one side may be positive, and those on the other side may be negative. They will attract each other and will have a tendency to stick together more than they would have before—the stream doesn't break up as much. On the other hand, if the field is stronger, the charge in each one of the drops gets much larger, and there is a tendency for the charge itself to help break up the drops through their own repulsion. Each drop will break into many smaller ones, each carrying a charge, so that they are all repelled, and spread out so rapidly. So as we increase the field, the stream becomes more finely separated. The only point we wish to make is that in certain circumstances electric fields can have considerable influence on the drops. The exact machinery by which something happens in a thunderstorm is not at all known, and is not at all necessarily related to what we have just described. We have included it just so that

* A handy way to observe the sizes of the drops is to let the spray fall on a large thin metal plate. The larger drops make a bumpy noise.

you will appreciate the complexities that could come into play. In fact, nobody has a theory applicable to clouds based on that idea.

We would like to describe two theories which have been invented to account for the separation of the charges in a thunderstorm. All the theories involve the idea that there should be some charge on the precipitation particles and a different charge in the air. Then by the movement of the precipitation particles—the water or the ice—through the air there is a separation of electric charge. The only question is: How does the charging of the drop begin? One of the older theories is called the "breaking-drop" theory. Somebody discovered that if you have a drop of water that breaks up into two pieces in a windstream, there is positive charge on the water and negative charge in the air. This breaking-drop theory has several disadvantages, among which the most serious is that the sign is wrong. Second, in the large number of temperate-zone thunderstorms which do exhibit lightning, the precipitation effects at high altitudes are in ice, not in water.

From what we have just said, we note that if we could imagine some way for the charge to be different at the top and bottom of a drop and if we could also see some reason why drops in a high-speed airstream would break up into unequal pieces—a large one in the front and a smaller one in the back because of the motion through the air or something—we would have a theory. (Different from any known theory!) Then the small drops would not fall through the air as fast as the big ones, because of the air resistance, and we would get a charge separation. You see, it is possible to conceive all kinds of possibilities.

One of the more ingenious theories, which is more satisfactory in many respects than the breaking-drop theory, is due to C. T. R. Wilson. We will describe it, as Wilson did, with reference to water drops, although the same phenomenon would also work with ice. Suppose we have a water drop that is falling in the electric field of about 100 volts per meter toward the negatively charged earth. The drop will have no induced dipole moment—with the bottom of the drop positive and the top of the drop negative, as drawn in Fig. 9-12. Now there are in the air the "quarks" that we mentioned earlier—the large slow-moving ions. (The fast ions do not have an important effect here.) Suppose that as a drop comes down, it approaches a large ion. If the ion is positive, it is repelled by the positive bottom of the drop and is pushed away. So it does not become attached to the drop. If the ion were to approach from the top, however, it might attach to the negative, top side. But since the drop is falling through the air, there is an air drift relative to it, going upwards, which carries the ion away. If that motion through the air is slow enough, thus the positive ion cannot catch up to the drop either. That would apply, you see, only to the large, slow-moving ions. The positive ions of this type will not attach themselves either to the front or the back of a falling drop. On the other hand, as the large, slow, negative ions are approached by a drop, they will be attracted and will be caught. The drop will acquire negative charge—the sign of the charge having been determined by the original potential difference on the entire earth—and we get the right sign. Negative charge will be brought down to the bottom part of the cloud by the drops, and the positively charged ions which are left behind will be blown to the top of the cloud by the various updraft currents. The theory looks pretty good, and it at least gives the right sign. Also it doesn't depend on having liquid drops. We will see, when we learn about polarized ice in dielectric, that pieces of ice will do the same thing. They also will develop positive and negative charges on their extremities when they are in an electric field.

There are, however, some problems even with this theory. First of all, the total charge involved in a thunderstorm is very high. After a short time, the supply of large ions would get used up. So Wilson and others have had to propose that there are additional sources of the large ions. Once the charge separation starts, very large electric fields are developed, and in these large fields there may be places where the air will become ionized. If there is a highly charged point, or any small object like a drop, it may concentrate the field enough to make a "brush discharge." When there is a strong enough electric field—let us say it is positive electrons will fall into the field and will pick up a lot of speed between collisions. Their speed will be such that in hitting another atom they will tear electrons off of that

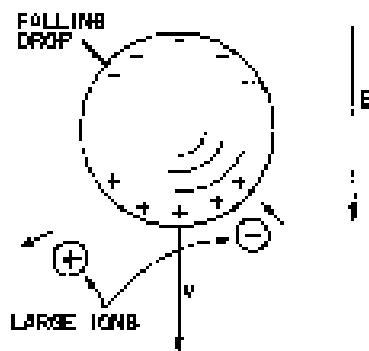


Fig. 9-13. C. T. R. Wilson's theory of charge separation in a thundercloud.

atom, leaving positive charges behind. These new electrons also pick up speed and collide with more electrons. So a kind of chain reaction or avalanche occurs, and there is a rapid accumulation of ions. The positive charges are left near their original positions, so the net effect is to distribute the positive charge so the point into a region around the point. Then, of course, there is no longer a strong field, and the process stops. This is the character of a brush discharge. It is possible that the fields may become strong enough in the cloud to produce a little bit of brush discharge; there may also be other mechanisms. Once the thing is started, to produce a large amount of ionization. But nobody knows exactly how it works. So the fundamental origin of lightning is really not thoroughly understood. We know it comes from the thunderstorms. (And we know, of course, that thunder comes from the lightning—from the thermal energy released by the bolt.)

At least we can understand, in part, the origin of atmospheric electricity. Due to the air currents, ions, and water droplets or particles in a thunderstorm, positive and negative charges are separated. The positive charges are carried upward to the top of the cloud (see Fig. 9-11), and the negative charges are dumped into the ground in lightning strikes. The positive charges leave the top of the cloud, enter the high-altitude layers of more highly conducting air, and spread throughout the earth. In regions of clear weather, the positive charges in this layer are slowly conducted to the earth by the ions in the air—ions formed by cosmic rays, by the sea, and by man's activities. The atmosphere is a busy electrical machine!

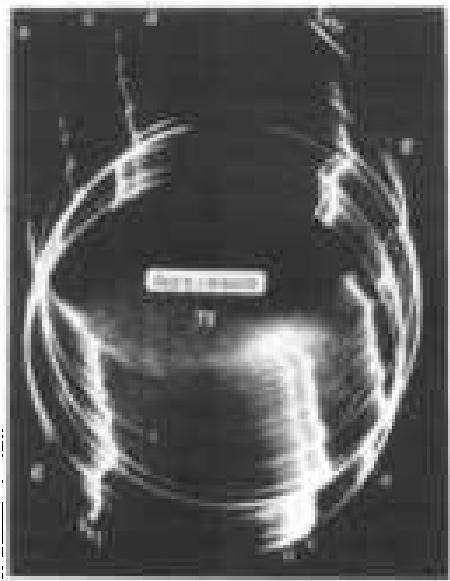
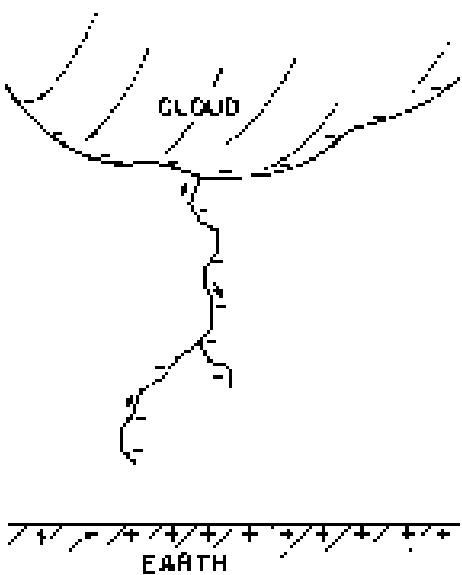


Fig. 9-14. Photograph of a lightning flash taken with a "Boys" camera. [From Schonland, Malan, and Collie, Proc. Roy. Soc. London, Vol. 152 (1935).]

9-6 Lightning

The first evidence of what happens in a lightning stroke was obtained in photographs taken with a camera held by hand and moved back and forth with the shutter open—while pointed toward a place where lightning was expected. The first photographs obtained this way showed clearly that lightning strokes are usually multiple discharges along the same path. Later, the "Boys" camera, which has two lenses mounted 180° apart on a rapidly rotating disc, was developed. The image made by each lens moves across the film—the picture is spread out in time. If, for instance, the stroke repeats, there will be two images side by side. By comparing the images of the two lenses, it is possible to work out the details of the time sequence of the flashes. Figure 9-14 shows a photograph taken with a "Boys" camera.

We will now describe the lightning. Again, we don't understand exactly how it works. We will give a qualitative description of what it looks like, but we won't go into any details of why it does what it appears to do. We will describe only the ordinary case of the cloud with a negative bottom over flat country. Its potential is much more negative than the earth underneath, so negative electrons will be accelerated toward the earth. What happens is the following. It all starts with a thing called a "step leader," which is just as bright as the stroke of lightning. On the photographs one can see a little bright spot at the beginning that starts from the cloud and moves downward very rapidly—at a sixth of the speed of light! It goes only about 50 meters and stops. It pauses for about 50 microseconds, and then takes another step. It pauses again and then goes another step, and so on. It moves in a series of steps toward the ground, along a path like that shown in Fig. 9-15. In the leader there are negative charges from the cloud; the whole column is full of negative charge. Also, the air is becoming ionized by the rapidly moving charges that produce the leader, so the air becomes a conductor along the path traced out. The moment the leader touches the ground, we have a conducting "wye" that runs all the way up to the cloud and is full of negative charge. Now, at last, the negative charge of the cloud can simply escape and run out. The electrons at the bottom of the leader are the first ones to realize this; they dump out, leaving positive charge behind that attracts more negative charge from higher up in the leader, which in its turn pours out, etc. So finally all the negative charge in a part of the cloud runs out along the return in a rapid and energetic way. So the lightning stroke you see runs upwards from the ground, as indicated in Fig. 9-16. In fact, this main stroke—the far the brightest part—is called the return



stroke. It is what produces the very bright light, and the heat, which by causing a rapid expansion of the air makes the thunder clap.

The current in a lightning stroke is about 10,000 amperes at its peak, and it carries down about 20 coulombs.

But we are still not finished. After a time of, perhaps, a few hundredths of a second, when the return stroke has disappeared, another leader comes down. But this time there are no patches. It is called a "dark leader" this time, and it goes all the way down—from top to bottom in one sweep. It goes full stroke on exactly the old track, because there is enough debris there to make it the easiest route. The new leader is again full of negative charge. The moment it touches the ground—bang!—there is a return stroke going straight up along the path. So you see the lightning strike again, and again, and again. Sometimes it strikes only once or twice, sometimes five or ten times—but as many as 42 times on the same track have been—but always in rapid succession.

Sometimes things get even more complicated. For instance, after one of its pauses the leader may develop a branch by sending out sub-steps—which regard the ground but in somewhat different directions, as shown in Fig. 9-15. What happens then depends on whether one branch reaches the ground definitely before the other. If that does happen, the bright return stroke (of negative charge jumping into the ground) works its way up along the branch that touches the ground, and when it reaches and passes the branching point on its way up to the cloud, a bright stroke appears to go down the other branch. Why? Because negative charge is dumping out and that is what lights up the bolt. This charge begins to move at the top of the secondary branch, carrying successive, longer pieces of the branch, so the bright lightning bolt appears to work its way down that branch, at the same time as it works up toward the cloud. If, however, one of these extra leader branches happens to have reached the ground almost simultaneously with the original leader, it can sometimes happen that the dark leader of the second stroke will take the second branch. Then you will see the first main flash in one place and the second flash in another place. It is a variant of the original idea.

Also, our description is oversimplified for the region very near the ground. When the step leader gets to within a hundred meters or so from the ground, there is evidence that a discharge rises from the ground to meet it. Presumably, the field gets big enough for a streamer-type discharge to occur. If, for instance, there is a sharp object, like a building with a point at the top, then as the leader comes down nearby the fields are so large that a discharge starts from the sharp point and reaches up to the leader. The lightning tends to strike such a point.

It has apparently been known for a long time that high objects are struck by lightning. There is a quotation of Artabazis, the advisor to Xerxes, giving his master advice on a contemplated attack on the Greeks during Xerxes' campaign to bring the entire Persian world under the control of the Persians. Artabazis said, "See how God with his lightning always smites the bigger animals and will not suffer them to live long, while these of a lesser bulk escape him not. How likewise his bolts fall ever on the greatest houses and tallest trees." And then he explains the reason: "So, plainly, doth he love to bring down everything that exalt itself."

Do you think—now that you know a true account of lightning striking tall trees—that you have a greater wisdom in advising kings on military matters than did Artabazis 2300 years ago? Do not exalt yourself. You could only do it less poetically.

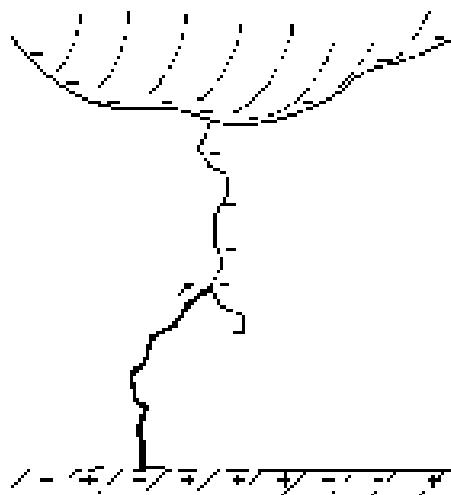


Fig. 9-16. The return lightning stroke runs back up the path made by the leader.

Dielectrics

10-1 The dielectric constant

Here we begin to discuss another of the peculiar properties of matter under the influence of the electric field. In an earlier chapter we considered the behavior of conductors, in which the charges move freely in response to all electric fields to such points that there is no field left inside a conductor. Now we will discuss insulators, materials which do not conduct electricity. One might at first believe that there should be no effect whatsoever. However, using a simple electrostatics and a parallel-plate capacitor, Faraday discovered that this was not so. His experiments showed that the capacitance of such a capacitor is increased when an insulator is put between the plates. If the insulator completely fills the space between the plates, the capacitance is increased by a factor κ which depends only on the nature of the insulating material. Insulating materials are also called dielectrics; the factor κ is then a property of the dielectric, and is called the dielectric constant. The dielectric constant of a vacuum is, of course, unity.

Our problem now is to explain why there is any electrical effect if the insulators are indeed insulators and do not conduct electricity. We begin with the experimental fact that the capacitance is increased and try to reason out what might be going on. Consider a parallel-plate capacitor with some charges on the surfaces of the conductors, let us say negative charge on the top plate and positive charge on the bottom plate. Suppose that the spacing between the plates is d and the area of each plate is A . As we have proved earlier, the capacitance is

$$C = \frac{\epsilon_0 A}{d}, \quad (10.1)$$

and the charge and voltage on the capacitor are related by

$$Q = C V. \quad (10.2)$$

Now the experimental fact is that if we put a piece of insulating material like lucite or glass between the plates, we find that the capacitance is larger. That means, of course, that the voltage is lower for the same charge. But the voltage difference is the integral of the electric field across the capacitor; so we must conclude that inside the capacitor, the electric field is reduced even though the charges on the plates remain unchanged.

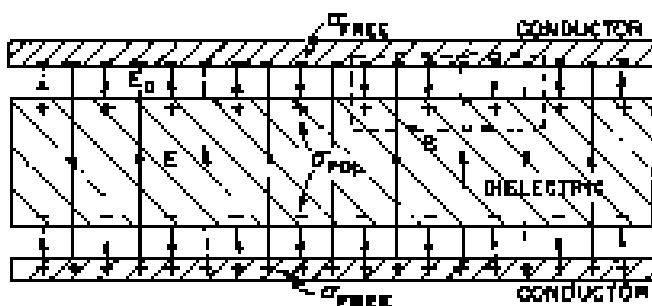


Fig. 10-1. A parallel-plate capacitor with a dielectric. The lines of E are shown.

Now how can that be? We have a law due to Gauss that tells us that the flux of the electric field is directly related to the enclosed charge. Consider the gaussian surface S shown by broken lines in Fig. 10-1. Since the electric field is reduced with the dielectric present, we conclude that the net charge inside the surface must

10-1 The dielectric constant

10-2 The polarization vector P

10-3 Polarization charges

10-4 The electrostatic equations with dielectrics

10-5 Fields and forces with dielectrics

be lower than it would be without the material. There is only one possible conclusion, and that is that there must be positive charge on the surface of the dielectric. Since the field is reduced but is not zero, we would expect this positive charge to be smaller than the negative charge on the conductor. So the phenomena can be explained if we could understand in some way that when a dielectric material is placed in an electric field there is positive charge induced on one surface and negative charge induced on the other.

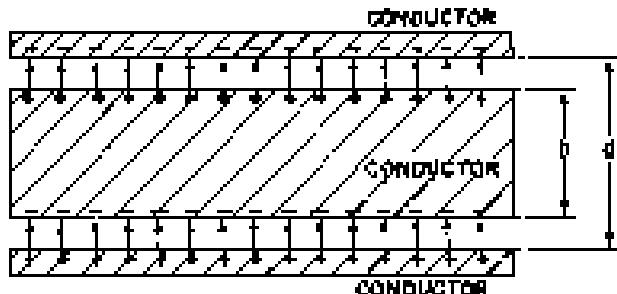


Fig. 10-2. If we put a conducting plate in the gap of a parallel-plate capacitor, the induced charges reduce the field in the conductor to zero.

We would expect that to happen for a conductor. For example, suppose that we had a capacitor with a plate spacing d , and we put between the plates a neutral conductor whose thickness is b , as in Fig. 10-2. The electric field induces a positive charge on the upper surface and a negative charge on the lower surface, so there is no field inside the conductor. The field in the rest of the space is the same as it was without the conductor, because it is the surface density of charge divided by ϵ_0 ; but the distance over which we have to integrate to get the voltage (the potential difference) is reduced. The voltage is

$$V = \frac{\sigma}{\epsilon_0} (d - b).$$

The resulting equation for the capacitance is like Eq. (10.2), with $(d - b)$ substituted for d :

$$C = \frac{\epsilon_0 d}{d - (b/d)} \quad (10.3)$$

The capacitance is increased by a factor which depends upon (b/d) , the proportion of the volume which is occupied by the conductor.

This gives us an obvious model for what happens with dielectrics—that inside the material there are many little sheets of conducting material. The trouble with such a model is that it has a specific axis, the normal to the sheets, whereas most dielectrics have no such axis. However, this difficulty can be eliminated if we assume that all insulating materials contain small conducting spheres separated from each other by insulation, as shown in Fig. 10-3. The phenomenon of the dielectric constant is explained by the effect of the charges which would be induced on each sphere. This is one of the earliest physical models of dielectrics used to explain the phenomenon that Faraday observed. More specifically, it was assumed that each of the atoms of a material was a perfect conductor, but isolated from the others. The dielectric constant κ would depend on the proportion of space which was occupied by the conducting spheres. This is not, however, the model that is used today.



Fig. 10-3. A model of a dielectric as small conducting spheres embedded in an idealized insulator.

10-2 The polarization vector P

If we follow the above analysis further, we discover that the idea of regions of perfect conductivity and insulation is not useful. Each of the small spheres acts like a dipole, the moment of which is induced by the external field. The only thing that is essential to the understanding of dielectrics is that there are many little dipoles induced in the material. Whether the dipoles are induced because there are tiny conducting spheres or for any other reason is irrelevant.

Why should a field induce a dipole moment in an atom if the atom is not a conducting sphere? This subject will be discussed in much greater detail in the next chapter, which will be about the inner workings of dielectric materials. However, we give here one example to illustrate a possible mechanism. An atom has a positive charge on the nucleus, which is surrounded by negative electrons. In an electric field, the nucleus will be attracted in one direction and the electrons in the other. The orbits or wave patterns of the electrons (or whatever picture is used in quantum mechanics) will be distorted to some extent, as shown in Fig. 10-4; the center of gravity of the negative charge will be displaced and will no longer coincide with the positive charge of the nucleus. We have already discussed such distributions of charge. If we look from a distance, such a neutral configuration is equivalent, to a first approximation, to a little dipole.

It seems reasonable that if the field is not too enormous, the amount of induced dipole moment will be proportional to the field. That is, a small field will displace the charges a little bit and a larger field will displace them further-- and in proportion to the field-- unless the displacement gets too large. For the remainder of this chapter, it will be supposed that the dipole moment is exactly proportional to the field.

We will now assume that in each atom there are charges q separated by a distance b , so that qb is the dipole moment per atom. (We use b because we are already using d for the plate separation.) If there are N atoms per unit volume, there will be a dipole moment per unit volume equal to Nqb . This dipole moment per unit volume will be represented by a vector, P . Needless to say, it is in the direction of the *undisturbed* dipole moment, i.e., in the direction of the charge separation b :

$$P = Nqb. \quad (10-4)$$

In general, P will vary from place to place in the dielectric. However, at any point in the material, P is proportional to the electric field E . The constant of proportionality, which depends on the ease with which the electron are displaced, will depend on the kinds of atoms in the material.

What actually determines how this constant of proportionality behaves, how accurately it is constant for very large fields, and what is going on inside different materials, we will discuss at a later time. For the present, we will simply suppose that there exists a mechanism by which a dipole moment is induced which is proportional to the electric field.

10-3 Polarization charges

Now let us see what this model gives for the theory of a condenser with a dielectric. First consider a sheet of material in which there is a certain dipole moment per unit volume. Will there be on the average any charge density produced by this? Not if P is uniform. If the positive and negative charges being displaced relative to each other have the same average density, the fact that they are displaced does not produce any net charge inside the volume. On the other hand, if P were larger at one place and smaller at another, that would mean that more charge would be moved into some region than away from it; we would then expect to get a volume density of charge. For the parallel-plate condenser, we suppose that P is uniform, so we need to look only at what happens at the surfaces. At one surface the negative charges, the electrons, have effectively moved out a distance b ; at the other surface they have moved in, leaving some positive charge effectively out a distance b . As shown in Fig. 10-5, we will have a surface density of charge, which will be called the *surface polarization charge*.

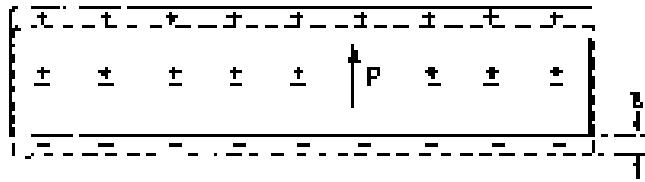


Fig. 10-5. A dielectric slab in a uniform field. The positive charges displaced the distance b with respect to the negatives.

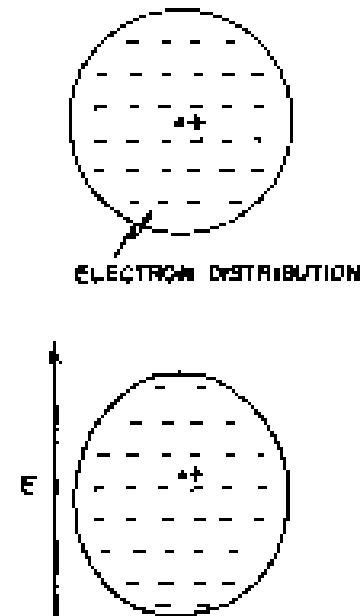


Fig. 10-4. An atom in an electric field has its distribution of electrons displaced with respect to the nucleus.

This charge can be calculated as follows. If A is the area of the plate, the number of electrons that appear at the surface is the product of A and N , the number per unit volume, and the displacement δ , which we assume here is perpendicular to the surface. The total charge is obtained by multiplying by the electronic charge q_e . To get the surface density of the polarization charge induced on the surface, we divide by A . The magnitude of the surface charge density is

$$\sigma_{\text{pol}} = N q_e \delta.$$

But this is just equal to the magnitude P of the polarization vector \mathbf{P} . Eq. (10.4):

$$\sigma_{\text{pol}} = P. \quad (10.5)$$

The surface density of charge is equal to the polarization inside the material. The surface charge is, of course, positive on one surface and negative on the other.

Now let us assume that our slab is the dielectric of a parallel-plate capacitor. The plates of the capacitor also have a surface charge, which we will call σ_{ext} because they can move "freely" anywhere on the conductor. This is, of course, the charge that we put on when we charge the capacitor. It should be emphasized that σ_{ext} exists only because of σ_{int} . If σ_{int} is removed by discharging the capacitor, then σ_{ext} will disappear, not by going out on the discharging wire, but by moving back into the material, by the relaxation of the polarization inside the material.

We can now apply Gauss' law to the Gaussian surface S in Fig. 10-1. The electric field E in the dielectric is equal to the total surface charge density divided by ϵ_0 . It is clear that σ_{ext} and σ_{int} have opposite signs, so

$$E = \frac{\sigma_{\text{ext}} + \sigma_{\text{int}}}{\epsilon_0} \quad (10.6)$$

Note that the field E_0 between the metal plate and the surface of the dielectric is higher than the field E ; it corresponds to σ_{ext} alone. But here we are concerned with the field inside the dielectric which, if the dielectric nearly fills the gap, is the field over nearly the whole volume. Using Eq. (10.5), we can write

$$E = \frac{\sigma_{\text{ext}} - P}{\epsilon_0}. \quad (10.7)$$

This equation doesn't tell us what the electric field is unless we know what P is. Here, however, we are assuming that P depends on E —in fact, that it is proportional to E . This proportionality is usually written as

$$P = \chi \epsilon_0 E. \quad (10.8)$$

The constant χ (Greek "khi") is called the electric susceptibility of the dielectric. Then Eq. (10.7) becomes

$$E = \frac{\sigma_{\text{ext}}}{\epsilon_0} \frac{1}{(1 + \chi)}, \quad (10.9)$$

which gives us the factor $1/(1 + \chi)$ by which the field is reduced.

The voltage between the plates is the integral of the electric field. Since the field is uniform, the integral is just the product of E and the plate separation d . We have that

$$V = Ed = \frac{\sigma_{\text{ext}} d}{\epsilon_0(1 + \chi)}.$$

The total charge on the capacitor is stored, so that the capacitance defined by (10.2) becomes

$$C = \frac{\epsilon_0 d(1 + \chi)}{d} = \frac{\chi \epsilon_0 d}{d}. \quad (10.10)$$

We have explained the observed facts. When a parallel-plate capacitor is filled with a dielectric, the capacitance is increased by the factor

$$\kappa = 1 + \chi. \quad (10.11)$$

which is a property of the material. Our explanation, of course, is not complete until we have explained—as we will do later—how the atomic polarization occurs about.

Let's now consider something a little bit more complicated—the situation in which the polarization \mathbf{P} is not everywhere the same. As mentioned earlier, if the polarization is not constant, we would expect it to look like a charge density in the volume, because more charge might move into one side of a small volume element than leaves it on the other. How can we find out how much charge is gained or lost from a small volume?

First let's compute how much charge moves across any imaginary surface when the material is polarized. The amount of charge that goes across a surface is just \mathbf{P} times the surface area if the polarization is normal to the surface. Of course, if the polarization is tangential to the surface, no charge moves across it.

Following the same arguments we have already used, it is easy to see that the charge moved across any surface element is proportional to the component of \mathbf{P} perpendicular to the surface. Compare Fig. 10-6 with Fig. 10-5. We see that Eq. (10.5) should, in the general case, be written

$$\sigma_{\text{pol}} = \mathbf{P} \cdot \mathbf{n} \quad (10.12)$$

If we are thinking of an imagined surface element inside the dielectric, Eq. (10.12) gives the charge moved across the surface but doesn't result in a net surface charge, because there are equal and opposite contributions from the dielectric on the two sides of the surface.

The displacements of the charges can, however, result in a net charge density. The total charge displaced out of any volume V by the polarization is the integral of the outward normal component of \mathbf{P} over the surface S that bounds the volume (see Fig. 10-7). An equal excess charge of the opposite sign is left behind. Denoting the net charge inside V by ΔQ_{pol} , we write

$$\Delta Q_{\text{pol}} = - \int_S \mathbf{P} \cdot \mathbf{n} d\mathbf{a}. \quad (10.13)$$

We call attribute ΔQ_{pol} to a volume distribution of charge with the density ρ_{pol} , and so

$$\Delta Q_{\text{pol}} = \int_V \rho_{\text{pol}} dV. \quad (10.14)$$

Combining the two equations yields

$$\int_V \rho_{\text{pol}} dV = - \int_S \mathbf{P} \cdot \mathbf{n} d\mathbf{a}. \quad (10.15)$$

We have a kind of Gauss' theorem that relates the charge density from polarized materials to the polarization vector \mathbf{P} . We can see that it agrees with the result we got for the surface polarization charge on the dielectric in a parallel-plate capacitor. Using Eq. (10.15) with the gaussian surface of Fig. 10-1, the surface integral gives \mathbf{P} and, and the charge inside is $\sigma_{\perp\perp}$, so we get again that $\sigma = \mathbf{P}$.

Just as we did for Gauss' law of electrostatics, we can convert Eq. (10.15) to a differential form—using Gauss' mathematical theorem:

$$\int_S \mathbf{P} \cdot \mathbf{n} d\mathbf{a} = \int_V \nabla \cdot \mathbf{P} dV,$$

We get

$$\rho_{\perp\perp} = - \nabla \cdot \mathbf{P}. \quad (10.16)$$

If there is a nonuniform polarization, its divergence gives the net density of charge appearing in the material. We emphasize that this is a perfectly real charge density; we call it "polarization charge" only to remind ourselves how it got there.

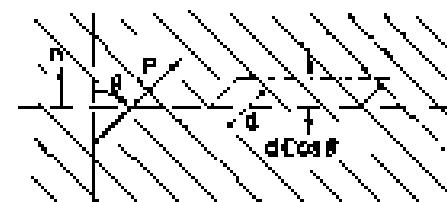


Fig. 10-6. The charge moved across an element of an imaginary surface in a dielectric is proportional to the component of \mathbf{P} normal to the surface.

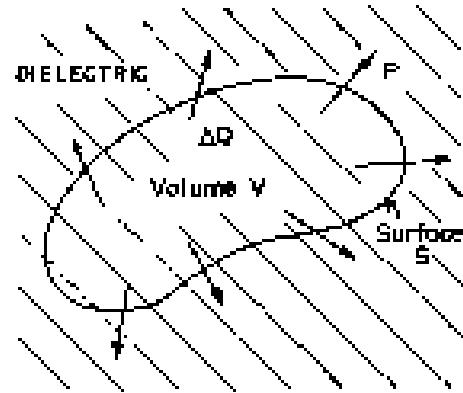


Fig. 10-7. A nonuniform polarization \mathbf{P} can result in a net charge in the body of a dielectric.

10-6 The electrostatic equations with dielectrics

Now let's combine the above result with our theory of electrostatics. The fundamental equation is

$$\nabla \cdot E = \frac{\rho}{\epsilon_0}. \quad (10.17)$$

The ρ here is the density of all electric charges. Since it is not easy to keep track of the polarization charges, it is convenient to separate ρ into two parts. Again we call ρ_{ext} the charge due to nonuniform polarization, and call ρ_{int} all the rest. Usually ρ_{ext} is the charge we put on conductors, or at known places in space. Equation (10.17) then becomes

$$\nabla \cdot E = \frac{\rho_{ext}}{\epsilon_0} + \frac{\rho_{int}}{\epsilon_0} = \frac{\rho_{ext}}{\epsilon_0} - \nabla \cdot P,$$

or

$$\nabla \cdot \left(E + \frac{P}{\epsilon_0} \right) = \frac{\rho_{ext}}{\epsilon_0}. \quad (10.18)$$

If course, the equation for the curl of E is unchanged:

$$\nabla \times E = 0. \quad (10.19)$$

Taking P from Eq. (10.8), we get the simpler equation

$$\nabla \cdot [(1 - \kappa)E] = \nabla \cdot (\kappa E) = \frac{\rho_{ext}}{\epsilon_0}. \quad (10.20)$$

These are the equations of electromatics when there are dielectrics. They don't, of course, say anything new, but they are in a form which is more convenient for computation in cases where ρ_{ext} is known and the polarization P is proportional to E .

Notice that we have not taken the dielectric "constant," κ , out of the divergence. That is because it may not be the same everywhere. If it has everywhere the same value, it can be factored out and the equations are just those of electromatics with the charge density ρ_{ext} divided by κ . In the form we have given, the equations apply to the general case where different dielectrics may be in different places in the field. Then the equations may be quite difficult to solve.

There is a history of some historical importance which should be mentioned here. In the early days of electricity, the atomic mechanism of polarization was not known and the existence of ρ_{int} was not appreciated. The charge ρ_{ext} was considered to be the entire charge density. In order to write Maxwell's equations in a simple form, a new vector D was defined to be equal to a linear combination of E and P :

$$D = \epsilon_0 E + \kappa P. \quad (10.21)$$

As a result, Eqs. (10.18) and (10.19) were written in an apparently very simple form:

$$\nabla \cdot D = \rho_{ext}, \quad \nabla \times E = 0. \quad (10.22)$$

Can one solve these? Only if a third equation is given for the relationship between D and E . When Eq. (10.8) holds, this relationship is

$$D = \epsilon_0 E + \kappa \epsilon_0 E = \kappa \epsilon_0 E. \quad (10.23)$$

This equation was usually written

$$D = \epsilon E, \quad (10.24)$$

where ϵ is still another constant for describing the dielectric property of materials. It is called the "permittivity." (Now you see why we have ϵ_0 in our equations. It is the "permittivity of empty space.") Evidently,

$$\epsilon = \kappa \epsilon_0 = (1 + \kappa) \epsilon_0 \quad (10.25)$$

Today we took up these matters from another point of view, namely, that we have simpler equations in a vacuum, and if we exhibit in every case all the charges, whatever their origin, the equations are always correct. If we separate source of the charges away for convenience, or because we do not want to discuss what is going on in detail, then we can, if we wish, write our equations in any other form that may be convenient.

One more point should be emphasized. An equation like $D \sim \epsilon E$ is an attempt to describe a property of matter. But matter is extremely complicated, and such an equation is at best not correct. For instance, if E gets too large, then D is no longer proportional to E . For small substances, the proportionality breaks down even with relatively small fields. Also, the "constant" of proportionality may depend on how fast E changes with time. Therefore this kind of equation is a kind of approximation, like Hooke's law. It cannot be a deep and fundamental equation. On the other hand, our fundamental equations for E , (10.17) and (10.19), represent our deepest and most complete understanding of electrostatics.

10-5 Fields and forces with dielectrics

We will now prove some rather general theorems for electrostatics in situations where dielectrics are present. We have seen that the capacitance of a parallel-plate capacitor is increased by a definite factor if it is filled with a dielectric. We can show that this is true for a capacitor of any shape, provided the entire region in the neighborhood of the two conductors is filled with a uniform linear dielectric. Without the dielectric, the equations to be solved are

$$\nabla \cdot E_0 = \frac{\rho_{free}}{\epsilon_0} \quad \text{and} \quad \nabla \times E_0 = 0.$$

With the dielectric present, the first of these equations is modified; we have instead the equation:

$$\nabla \cdot (\kappa E) = \frac{\rho_{free}}{\epsilon_0} \quad \text{and} \quad \nabla \times E = 0 \quad (10.26)$$

Now since we are taking κ to be everywhere the same, the last two equations can be written as

$$\nabla \cdot (\kappa E) = \frac{\rho_{free}}{\epsilon_0} \quad \text{and} \quad \nabla \times (\kappa E) = 0. \quad (10.27)$$

We therefore have the same equations for κE as for E_0 , so they have the solution $\kappa E = E_0$. In other words, the field is everywhere reduced, by the factor $1/\kappa$, than in the case without the dielectric. Since the voltage difference is a line integral of the field, the voltage is reduced by this same factor. Since the charge on the electrodes of the capacitor has been taken the same in both cases, Fig. (10.2) tells us that the capacitance, in the case of an everywhere uniform dielectric, is increased by the factor κ .

Let us now ask what the force would be between two charged conductors in a dielectric. We consider a liquid dielectric that is homogeneous everywhere. We have seen earlier that one way to obtain the force is to differentiate the energy with respect to the appropriate distance. If the conductors have equal and opposite charges, the energy $U = Q^2/2C$, where C is their capacitance. Using the principle of virtual work, any component is given by a differentiation; for example,

$$F_x = -\frac{\partial U}{\partial x} = -\frac{Q^2}{2} \frac{1}{x} \left(\frac{1}{C} \right). \quad (10.28)$$

Since the dielectric increases the capacity by a factor κ , all forces will be reduced by this same factor.

One point should be emphasized. What we have said is true only if the dielectric is a liquid. Any mention of conductors that are embedded in solid dielectric changes the mechanical stress conditions of the dielectric and alters its electrical

properties, as well as carrying some mechanical energy change in the dielectric. Moving the conductors in a liquid does not change the liquid. The liquid moves to a new place but its electrical characteristics are not changed.

Much older books on electricity start with the "fundamental" law that the force between two charges is

$$F = \frac{q_1 q_2}{4\pi\epsilon_0 r^2}. \quad (10-29)$$

a point of view which is thoroughly unsatisfactory. For one thing, it is not true in general; it is true only for a world filled with a liquid. Secondly, it depends on the fact that ϵ_0 is a constant, which is only approximately true for most dielectrics. It is much better to start with Coulomb's law for charges in a vacuum, which is always right (for stationary charges).

What does happen in a solid? This is a very difficult problem which has not been solved, because it is, in a strict, indecomposable. If you put charges inside a dielectric solid, there are many kinds of pressures and strains. You cannot deal with virtual work without including also the mechanical energy required to compress the solid, and it is a difficult matter, generally speaking, to make a unique distinction between the electrical forces and the mechanical forces due to the solid material itself. Fortunately, no one ever really needs to know the answer to the question proposed. He may sometimes want to know how much strain there is going to be in a solid, and that can be worked out. But it is much more complicated than the simple result we get for liquids.

A surprisingly complicated problem in the theory of dielectrics is the following: Why does a charged object pick up little pieces of dielectric? If you comb your hair on a dry day, the comb readily picks up small scraps of paper. If you thought carefully about it, you probably assumed the comb had one charge on it and the paper had the opposite charge on it. But the paper is initially electrically neutral. It hasn't any net charge, but it is attracted anyway. It is true that sometimes the paper will come up to the comb and then fly away, repelled immediately after it touches the comb. The reason is, of course, that when the paper touches the comb, it picks up some negative charges and thus the like charges repel. But that doesn't answer the original question. Why did the paper come toward the comb in the first place?

The answer has to do with the polarization of a dielectric when it is placed in an electric field. There are polarization charges of both signs, which are attracted and repelled by the comb. There is a net attraction, however, because the field nearer the comb is stronger than the field farther away—the comb is not an infinite sheet. Its charge is localized. A neutral piece of paper will not be attracted to either plate inside the parallel plates of a capacitor. The variation of the field is an essential part of the attraction mechanism.

As illustrated in Fig. 10-8, a dielectric is always drawn from a region of weak field toward a region of stronger field. In fact, one can prove that for small objects the force is proportional to the gradient of the square of the electric field. Why does it depend on the square of the field? Because the induced polarization charges are proportional to the field, and for given charges the forces are proportional to the field. However, as we have just indicated, there will be a net force only if the square of the field is changing from point to point. So the force is proportional to the gradient of the square of the field. The constant of proportionality involves, among other things, the dielectric constant of the object, and it also depends upon the size and shape of the object.

There is a related problem in which the force on a dielectric can be worked out quite accurately. If we have a parallel-plate capacitor with a dielectric slab only partially inserted, as shown in Fig. 10-9, there will be a force driving the sheet in. A detailed examination of the force is quite complicated; it is related to nonuniformities in the field near the edges of the dielectric and the plates. However, if we do not look at the details, but merely use the principle of conservation of energy, we can easily calculate the force. We can find the force from the formula we de-

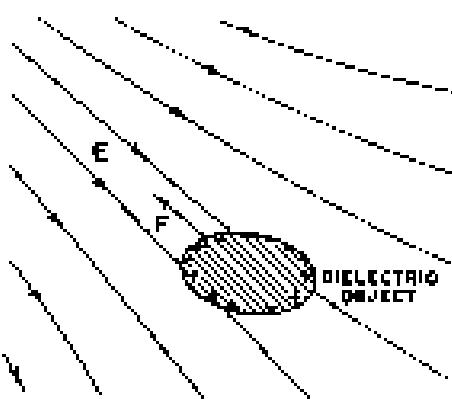


Fig. 10-8. A dielectric object in a nonuniform field feels a force toward regions of higher field strength.

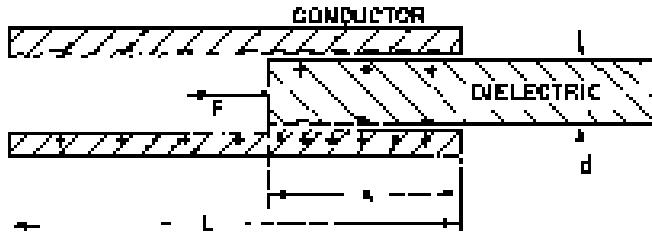


Fig. 10-9. The force on a dielectric sheet in a parallel-plate capacitor can be computed by applying the principle of energy conservation.

rived earlier. Equation (10.28) is equivalent to

$$F_x = -\frac{\partial U}{\partial x} = -\frac{P^2}{2} \frac{\partial C}{\partial x}. \quad (10.30)$$

We need only find out how the capacitance varies with the position of the dielectric slab.

Let's suppose that the total length of the plates is L , that the width of the plates is δY , that the plate separation and dielectric thickness are d , and that the distance x which the dielectric has been inserted is x . The capacitance is the ratio of the total free charge on the plates to the voltage between the plates. We have seen above that for a given voltage V the surface charge density of free charge is σ_0/ϵ_0 . So the total charge on the plates is

$$Q = \frac{\sigma_0 \delta Y}{\epsilon_0} x \Omega + \frac{\sigma_0 \delta Y}{\epsilon_0} (L - x) \Omega,$$

from which we get the capacitance:

$$C = \frac{\epsilon_0 \delta Y}{d} (x + L - x). \quad (10.31)$$

Using (10.30), we have

$$F_x = \frac{P^2}{2} \frac{\epsilon_0 \delta Y}{d} (x - 1). \quad (10.32)$$

Now this equation is not particularly useful for anything unless you happen to need to know the force in such circumstances. We only wished to show that the theory of energy can often be used to avoid enormous complications in determining the forces on dielectric materials—as force would be in the present case.

The discussion of the theory of dielectrics has dealt only with electrical phenomena, accepting the fact that the material has a polarization which is important to the electric field. Why there is such a proportionality is perhaps of greater interest to physics. Once we understand the origin of the dielectric constants from an atomic point of view, we can use electrical measurements of the dielectric constants in varying environments to obtain detailed information about atomic or molecular structures. This aspect will be treated in part in the next chapter.

Inside Dielectrics

11-1 Molecular dipoles

In this chapter we are going to discuss why it is that materials are dielectric. We said in the last chapter that we could understand the properties of electrical systems with dielectrics once we appreciated that when an electric field is applied to a dielectric it induces a dipole moment in the atoms. Specifically, if the electric field E induces an average dipole moment per unit volume P , then κ , the dielectric constant, is given by

$$\kappa = 1 + \frac{P}{\epsilon_0 E} \quad (11.1)$$

We have already discussed how this equation is applied; now we have to discuss the mechanism by which polarization arises when there is an electric field inside a molecule. We begin with the simplest possible example—the polarized air of gases. But even gases already have complications: there are two types. The molecules of atomic gases, like oxygen, which has a symmetrical pair of atoms in each molecule, have no internal dipole moment. But the molecules of others, like water vapor (which has a nonplanar-like arrangement of hydrogen and oxygen atoms) carry a permanent electric dipole moment. As we pointed out in Chapters 6 and 7, there is in the water vapor molecule an average plus charge on the hydrogen atoms and a negative charge on the oxygen. Since the center of gravity of the negative charge and the center of gravity of the positive charge do not coincide, the total charge distribution of the molecule has a dipole moment. Such a molecule is called a polar molecule. In oxygen, because of the symmetry of the molecule, the centers of gravity of the positive and negative charges are the same, so it is a nonpolar molecule. It does, however, become a dipole when placed in an electric field. The factors of the two types of molecule are sketched in Fig. 11-1.

11-2 Electronic polarization

We will first discuss the polarization of nonpolar molecules. We can start with the simplest case of a monoatomic gas (for instance, helium). When an atom of such a gas is in an electric field, the electrons are pulled one way by the field while the nucleus is pulled the other way, as shown in Fig. 10-4. Although the atoms are very stiff with respect to the electrical forces we can apply experimentally, there is a slight net displacement of the centers of charge, and a dipole moment is induced. For small fields, the amount of displacement, and so also the dipole moment, is proportional to the electric field. The displacement of the electron distribution which produces this kind of induced dipole moment is called electronic polarization.

We have already discussed the influence of an electric field on an atom in Chapter 31 of Vol. I, when we were dealing with the theory of the index of refraction. If you think about it for a moment, you will see that what we just discussed is exactly the same as we did there. But now we need worry only about fields that do not vary with time, while the index of refraction depended on time-varying fields.

In Chapter 31 of Vol. I we supposed that when an atom is placed in an oscillating electric field the center of charge of the electrons obey the equation

$$m \frac{d^2x}{dt^2} - m\omega_0^2 x = qE \quad (11.2)$$

11-3 Molecular dipoles

11-4 Electronic polarization

11-5 Polar molecules; instantaneous polarization

11-6 Dielectric constant of liquids; the Debye-Hückel equation

11-7 Solid dielectrics

11-8 Ferroelectricity; BaTiO₃

Review: Chapter 11, Vol. I, *The Origin of the Refractive Index*.
Chapter 30, Vol. I, *The Principles of Statistical Mechanics*.

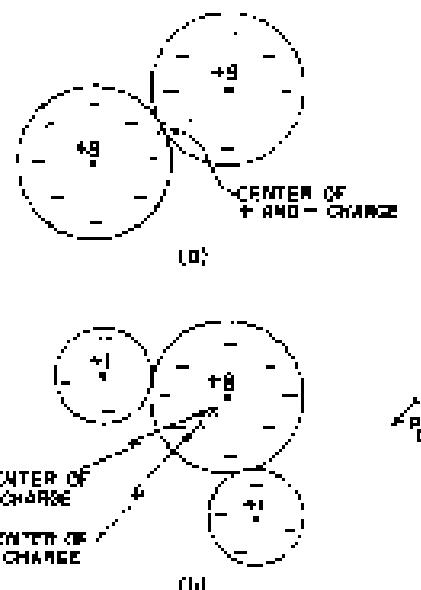


Fig. 11-1. [a] An oxygen molecule with zero dipole moment. [b] The water molecule has a permanent dipole moment P_0 .

The first term is the electron mass times its acceleration and the second is a restoring force, while the right-hand side is the force from the outside electric field. If the electric field varies with the frequency ω , Eq. (11.2) has the solution

$$\epsilon = \frac{ieE}{i\omega(\omega_0^2 - \omega^2)}, \quad (11.3)$$

which has a resonance at $\omega = \omega_0$. When we previously found this solution, we interpreted it as saying that ω_0 was the frequency at which light (in the optical region or in the ultraviolet, depending on the atom) was absorbed. For our purposes, however, we are interested only in the case of constant fields, i.e., for $\omega = 0$, so we can disregard the acceleration term in (11.2), and we find that the displacement is

$$x = \frac{q_0 E}{m\omega_0^2}. \quad (11.4)$$

From this we see that the dipole moment p of a single atom is

$$p = q_0 x = \frac{q_0^2 E}{m\omega_0^2}. \quad (11.5)$$

In this theory the dipole moment p is indeed proportional to the electric field.

People usually write

$$p = \alpha \epsilon_0 E. \quad (11.6)$$

(Again the ϵ_0 is put in for historical reasons.) The constant α is called the polarizability of the atom, and has the dimensions A^3 . It is a measure of how easy it is to induce a moment in an atom with an electric field. Comparing (11.5) with (11.6), one simple theory says that

$$\alpha = \frac{q_0^2}{m\omega_0^2} = \frac{4\pi r^3}{m\omega_0^2}. \quad (11.7)$$

If there are N atoms in a unit volume, the polarization P —the dipole moment per unit volume—is given by

$$P = Np = N\alpha \epsilon_0 E. \quad (11.8)$$

Putting (11.1) and (11.8) together, we get

$$\epsilon - 1 = \frac{P}{\epsilon_0 E} = N\alpha \quad (11.9)$$

or, using (11.7),

$$\epsilon - 1 = \frac{4\pi N r^3}{m\omega_0^2}. \quad (11.10)$$

From Eq. (11.9) we would predict that the dielectric constant ϵ of different gases should depend on the density of the gas and on the frequency ω_0 of its optical absorption.

Our formula is, of course, only a very rough approximation, because in Eq. (11.2) we have taken a model which ignores the complications of quantum mechanics. For example, we have assumed that an atom has only one resonant frequency, when it really has many. To calculate properly the polarizability α of atoms we must use the complete quantum-mechanical theory, but the classical ideas above give us a reasonable estimate.

Let's see if we can get the right order of magnitude for the dielectric constant of some substance. Suppose we try hydrogen. We have once estimated (Chapter 18, Vol. I) that the energy needed to ionize the hydrogen atom should be approximately

$$E \approx \frac{1}{2} \frac{me^4}{h^2 c^2} \quad (11.11)$$

For an estimate of the natural frequency ω_0 , we can set this energy equal to $\hbar\omega_0$, the energy of an atomic oscillator whose natural frequency is ω_0 . We get

$$\omega \approx \frac{1 \text{ cm}^{-1}}{2 \pi N}$$

If we now use this value of ω_0 in Eq. (11.7), we find for the electric susceptibility

$$\chi = \frac{\mu^2 / \epsilon_0}{16 \pi^2 N c^2} \quad (11.12)$$

The quantity $(\mu^2/\epsilon_0 c^2)$ is the radius of the ground-state orbit of a Bohr atom (see Chapter 18, Vol. I) and equals 0.528 nanostrobes. In a gas at standard pressure and temperature (1 atmosphere, 0°C) there are 1.69×10^{27} atoms/cm³, so Eq. (11.9) gives us

$$\chi = 1 + (2.69 \times 10^{19} \text{ Hg} \pi / 0.528 \times 10^{-4})^2 = 1.00020. \quad (11.13)$$

The dielectric constant for hydrogen gas is measured to be

$$\kappa_{\text{exp}} \approx 1.00026.$$

We see that our theory is almost right. We should not expect any better, because the measurements were, of course, made with ordinary hydrogen gas, which has diatomic molecules, not single atoms. We should not be surprised if the polarization of the atoms in a molecule is not quite the same as that of the separate atoms. The molecular effect, however, is not really that large. An exact quantum-mechanical calculation of χ for hydrogen atoms gives a result about 12% higher than (11.12) (the factor changes in (2.69)), and therefore predicts a dielectric constant somewhat closer to the observed one. In any case, it is clear that our model of a dielectric is fairly good.

Another check on our theory is to try Eq. (11.12) on atoms which have a higher frequency of excitation. For instance, it takes about 24.5 volts to pull the electron off helium, compared with the 13.6 volts required to ionize hydrogen. We would, therefore, expect that the absorption frequency ω_0 for helium would be about twice as big as for hydrogen and that χ would be one-quarter as large. We expect that

$$\chi_{\text{He}} \approx 1.000050.$$

Experimentally,

$$\chi_{\text{He}, \text{exp}} = 1.000055,$$

so you see that our simple estimates are coming out on the right track. So we have understood the dielectric constant of nonpolar gases, but only qualitatively, because we have never used a correct quantum theory of the motion of the atomic electrons.

11-3 Polar molecules; orientational polarization

Next we will consider a molecule which carries a permanent dipole moment μ_0 —such as a water molecule. With no electric field, the individual dipoles point in random directions, so the net moment per unit volume is zero. But when an electric field is applied, two things happen: First, there is a surface dipole moment induced because of the forces on the electrons; this part gives just the same kind of electric polarization we found for a nonpolar molecule. For very accurate work, this effect should, of course, be included, but we will neglect it for the moment. (It can always be added in at the end.) Second, the electric field tends to line up the individual dipoles to produce a net induced per unit volume. If all the dipoles in a gas were to line up, there would be a very large polarization, but that does not happen. As ordinary temperature and electric fields break the rotations of the molecules in their thermal motion, keep them from lining up very much. But there is some net alignment, and so some polarization (see Fig. 11-2). The polarization that does occur can be computed by the methods of statistical mechanics we described in Chapter 40 of Vol. I.

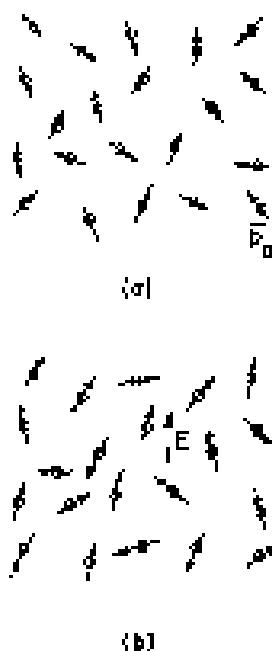


Fig. 11-2. (a) In a gas of polar molecules, the individual moments are oriented at random; the average moment in a small volume is zero. (b) When there is an electric field, there is some average alignment of the molecules.

To use this method we need to know the energy of a dipole in an electric field. Consider a dipole of moment μ_0 in an electric field, as shown in Fig. 11-1. The energy of the positive charge is $q\phi(1)$, and the energy of the negative charge is $-q\phi(2)$. Thus the energy of the dipole is

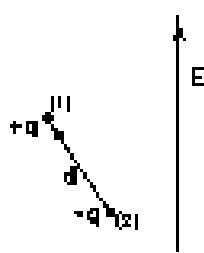


Fig. 11-3. The energy of a dipole μ_0 in the field $E = \rho_0 \cdot \vec{E}$.

$$U = q\phi(1) - q\phi(2) = qd \cdot \nabla\phi,$$

or

$$U = -\rho_0 \cdot E = -\rho_0 E \cos \theta, \quad (11.14)$$

where θ is the angle between μ_0 and E . As we would expect, the energy is lower when the dipoles are lined up with the field.

We now find out how much lining up occurs by using the methods of statistical mechanics. We found in Chapter 40 of Vol. 1 that in a state of thermal equilibrium, the relative number of molecules with the potential energy U is proportional to

$$e^{-U/kT}, \quad (11.15)$$

where $\phi(x, y, z)$ is the potential energy as a function of position. The same arguments would say that, using Eq. (11.14) for the potential energy as a function of angle, the number of molecules at θ per unit solid angle is proportional to $e^{-U/kT}$.

Letting $n(\theta)$ be the number of molecules per unit solid angle at θ , we have

$$n(\theta) = n_0 e^{-U_0/kT}, \quad (11.16)$$

For normal temperatures and fields, the exponent is small, so we can approximate by expanding the exponential:

$$n(\theta) = n_0 \left(1 + \frac{\rho_0 E \cos \theta}{kT} \right). \quad (11.17)$$

We can find n_0 if we integrate (11.17) over all angles; the result should be just N , the total number of molecules per unit volume. The average value of $\cos \theta$ over all angles is zero, so the integral is just n_0 times the total solid angle 4π . We get

$$n_0 = \frac{N}{4\pi}. \quad (11.18)$$

We see from (11.17) that there will be more molecules oriented along the field ($\cos \theta = 1$) than against the field ($\cos \theta = -1$). So in any small volume containing many molecules there will be a net dipole moment per unit volume—that is, a polarization P . To calculate P , we want the vector sum of all the molecular moments in a unit volume. Since we know that the result is going to be in the direction of E , we will just sum the components in that direction (the components at right angles to E will sum to zero);

$$P = \sum_{\text{unit volume}} \rho_0 \cos \theta.$$

We can evaluate the sum by integrating over the angular distribution. The solid angle at θ is $2\pi \sin \theta d\theta$, so

$$P = \int_0^\pi n(\theta) \rho_0 \cos \theta 2\pi \sin \theta d\theta. \quad (11.19)$$

Substituting for $n(\theta)$ from (11.17), we have

$$P = -\frac{N}{2} \int_{-1}^1 \left(1 + \frac{\rho_0 E}{kT} \cos \theta \right) \rho_0 \cos \theta d(\cos \theta),$$

which is easily integrated to give

$$P = \frac{N \rho_0^2 E}{3kT}. \quad (11.20)$$

The polarization is proportional to the field E , so there will be normal dielectric behavior. Also, as we expect, the polarization depends inversely on the temperature, because at higher temperatures there is more disalignment by collisions. This $1/T$ dependence is called Curie's law. The permanent moment p_0 appears squared for the following reason: In a given electric field, the aligning force depends upon σ , and the mean moment lost is produced by the lining up is again proportional to p_0^2 . The average induced moment is proportional to p_0^2 .

We should now try to see how well Eq. (11.20) agrees with experiment. Let's look at the case of steam. Since we don't know what p_0 is, we cannot compute κ directly, but Eq. (11.20) does predict that $\kappa - 1$ should vary inversely as the temperature, and that we should check.

From (11.20) we get

$$\kappa - 1 \approx \frac{F}{\epsilon_0 E} \approx \frac{N p_0^2}{3 \pi k T}. \quad (11.21)$$

So $\kappa - 1$ should vary in direct proportion to the density N , and inversely as the absolute temperature. The dielectric constant has been measured at several different pressures and temperatures, chosen such that the number of molecules in a unit volume remained fixed.⁴ [Notice that if the measurements had all been taken at constant pressure, the number of molecules per unit volume would decrease linearly with increasing temperature and $\kappa - 1$ would vary as T^{-1} instead of as T^{-4} .] In Fig. 11-4 we plot the experimental observations for $\kappa - 1$ as a function of $1/T$. The dependence predicted by (11.21) is followed quite well.

There is another characteristic of the dielectric constant of polar molecules—the variation with the frequency of the applied field. Due to the moment of inertia of the molecules, it takes a certain amount of time for the heavy molecules to turn toward the direction of the field. So if we apply frequencies in the high microwave region or above, the polar contribution to the dielectric constant begins to fall away because the molecules cannot follow. In contrast to this, the electronic polarizability still remains the same up to optical frequencies, because of the smaller inertia is the electrons.

11-4 Electric fields in cavities of a dielectric

We now turn to an interesting (but unapplied) question—the problem of the dielectric constant in dense materials. Suppose that we take liquid helium or liquid argon or some other nonpolar material. We still expect electronic polarization. But in a dense material, P can be large, so the field on an individual atom will be influenced by the polarization of the atoms in its close neighborhood. The question is, what electric field acts on the individual atom?

Imagine that the liquid is put between the plates of a condenser. If the plates are charged, they will produce an electric field in the liquid. But there are also charges in the individual atoms, and the total field E is the sum of both of these effects. This true electric field varies very, very rapidly from point to point in the liquid. It is very high inside the atoms—particularly right next to the nucleus—and relatively small between the atoms. The potential difference between the plates is the line integral of this total field. If we ignore all the fine-grained variations, we can think of an average electric field E , which is just V/d . (This is the field we were using in the last chapter.) We should think of this field as the average over a space containing many atoms.

Now you might think that an "average" atom in an "average" location would feel this average field. But it is not that simple, as we can show by considering what happens if we imagine different-shaped holes in a dielectric. For instance, suppose that we cut a slot in a polarized dielectric, with the slot oriented parallel to the field, as shown in part (a) of Fig. 11-5. Since we know that $\nabla \times E = 0$, the line integral of E around the curve, C , which goes as shown in (b) of the figure, should

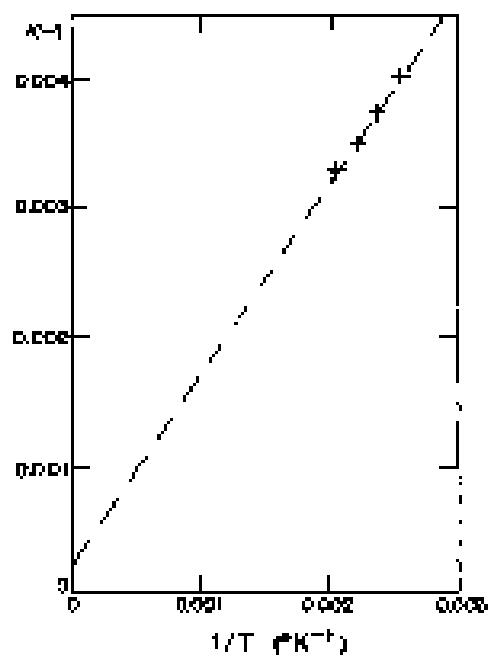


Fig. 11-4. Experimental measurements of the dielectric constant of water vapor at various temperatures.

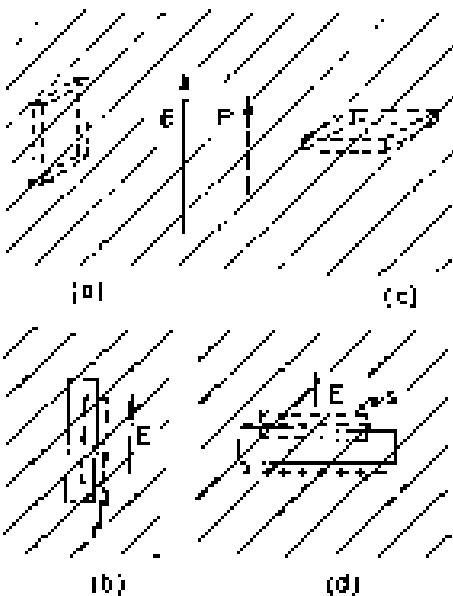


Fig. 11-5. The field in a slot cut in a dielectric depends on the shape and orientation of the slot.

⁴ Singwi, Stenflo, and Göttsche, *Svensk Fysik Revy* 5, 201 (1952).

be zero. The field inside the slot just give a contribution which just cancels the part from the field outside. Therefore the field E_0 actually found in the center of a long thin slot is equal to E , the average electric field found in the dielectric.

Now consider another slot whose boundaries are perpendicular to E , as shown in part (c) of Fig. 11-5. In this case, the field E_0 in the slot is not the same as E because polarization charges appear on the surfaces. If we apply Gauss' law to a surface S drawn as in (d) of the figure, we find that the field E_0 in the slot is given by

$$E_0 = E + \frac{P}{\epsilon_0} \quad (11.22)$$

where E is again the electric field in the dielectric. (The gaussian surface contains the surface polarization charge $\sigma_{\text{surf}} = P$.) We mentioned in Chapter 10 that $\epsilon_0 E + P$ is often called D , so $\epsilon_0 E_0 = D_0$ is equal to D in the dielectric.

Earlier in the history of physics, when it was supposed to be very important to define every quantity by direct experiment, people were delighted to discover that they could define what they meant by E and D in a dielectric without having to crawl around between the atoms. The average field E is numerically equal to the field E_0 that would be measured in a slot cut parallel to the field. And the field D could be measured by finding E_0 in a slot cut normal to the field. But nobody ever measures them that way anyway, so it was just one of those philosophical things.

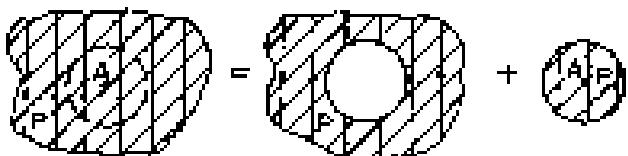


Fig. 11-6. The field at any point A in a dielectric can be considered as the sum of the field in a spherical hole plus the field due to a spherical plug.

For most liquids which are not too complicated in structure, we could expect that an atom finds itself, on the average, surrounded by the other atoms in what would be a good approximation to a spherical hole. And so we should ask: "What would be the field in a spherical hole?" We can find out by noticing that if we imagine carving out a spherical hole in a uniformly polarized material, we are just removing a sphere of polarized material. (We must imagine that the polarization is "frozen in" before we cut out the hole.) By superposition, however, the fields inside the dielectric, before the sphere was removed, is the sum of the fields from all charges outside the spherical volume plus the fields from the charges within the polarized sphere. That is, if we call E the field in the uniform dielectric, we can write

$$E = E_{\text{hole}} + E_{\text{polar}} \quad (11.23)$$

where E_{hole} is the field in the hole and E_{polar} is the field inside a sphere which is uniformly polarized (see Fig. 11-6). The fields due to a uniformly polarized sphere are shown in Fig. 11-7. The electric field inside the sphere is uniform, and its value is

$$E_{\text{polar}} = -\frac{P}{3\epsilon_0} \quad (11.24)$$

Using (11.23), we get

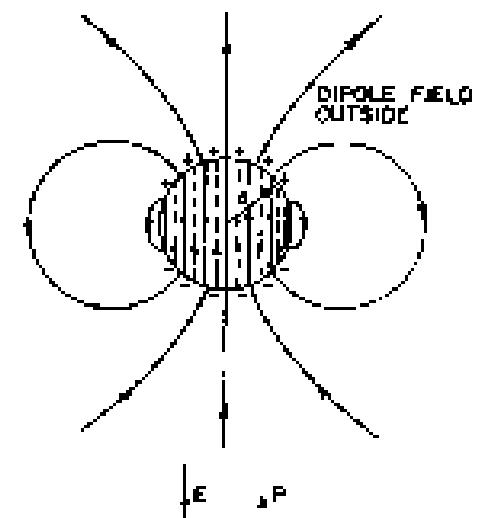
$$E_{\text{hole}} = E + \frac{P}{3\epsilon_0} \quad (11.25)$$

The field in a spherical cavity is greater than the average field by the amount $P/3\epsilon_0$. (The spherical hole gives a field $1/3$ of the way between a slot parallel to the field and a slot perpendicular to the field.)

11-5 The dielectric constant of liquids; the Clausius-Mossotti equation

In a liquid we expect that the field which will polarize an individual atom is more like E_{hole} than just E . If we use the result of (11.25) for the polarizing field in

Fig. 11-7. The electric field of a uniformly polarized sphere.



Eq. (11.6), then Eq. (11.8) becomes

$$P = Nae \left(E + \frac{E}{3\epsilon_0} \right), \quad (11.26)$$

or

$$P = \frac{Nae}{1 - (\epsilon_0/3)} \epsilon_0 E. \quad (11.27)$$

Remembering that $\epsilon - 1$ is just $P/\epsilon_0 E$, we have

$$\epsilon - 1 = \frac{Nae}{1 - (\epsilon_0/3)}, \quad (11.28)$$

which gives us the dielectric constant of a liquid in terms of a , the atomic polarizability. This is called the Gouy-Chapman equation.

Because Nae is very small, as it is for a gas (because the density N is small), then the term $Nae/3$ can be neglected compared with 1, and we get our old result, Eq. (11.9), that

$$\epsilon - 1 \approx Nae. \quad (11.29)$$

Let's compare Eq. (11.28) with some experimental results. It is best necessary to look at gases for which, using the measurement of ϵ , we can find a from Eq. (11.29). For instance, for carbon disulfide at zero degrees centigrade the dielectric constant is 1.0029, so Nae is 0.0029. Now the density of the gas is easily worked out and the density of the liquid can be found in handbooks. At 20°C, the density of liquid CS₂ is 381 times higher than the density of the gas at 0°C. This means that N is 381 times higher in the liquid than it is in the gas so, that—if we make the approximation that the basic atomic polarizability of the carbon disulfide doesn't change when it is condensed into a liquid— Nae in the liquid is equal to 381 times 0.0029, or 1.11. Notice that the $Nae/3$ term amounts to almost 0.4, so it is quite significant. With these numbers we predict a dielectric constant of 2.76, which agrees reasonably well with the observed value of 2.64.

In Table 11-2 we give some experimental data on various materials (taken from the *Handbook of Chemistry and Physics*), together with the dielectric constants calculated from Eq. (11.28) in the way just described. The agreement between observation and theory is even better for argon and oxygen than for CS₂—and not so good for carbon tetrachloride. On the whole, the results show that Eq. (11.28) works very well.

Table 11-1
Computation of the dielectric constants of liquids
from the dielectric constant of the gas.

Substance	Gas			Liquid				
	ϵ (exp)	Nae	Density	Density	Ratio*	Nae	ϵ (predicted)	ϵ (exp)
CS ₂	1.0029	0.0029	0.00039	1.290	381	1.11	2.76	2.64
O ₂	1.000223	0.000223	0.000145	1.19	832	0.435	1.2119	1.2117
CCl ₄	1.00011	0.00011	0.000489	1.59	325	0.907	2.45	2.24
A	1.000545	0.000545	0.00178	1.17	810	0.411	1.517	1.51

* Ratio = Density of liquid/density of gas.

Our derivation of Eq. (11.28) is valid only for *nonpolar* polarization in liquids. It is not right for a polar molecule like H₂O. If we go through the same calculations for water, we get 13.2 for Nae , which means that the dielectric constant for the liquid is negative, while the observed value of ϵ is 80. The problem lies to do with the correct treatment of the permanent dipoles, and Onsager has pointed out the right way to go. We do not have the time to treat the case now, but if you are interested it is discussed in Kittel's book, *Introduction to Solid State Physics*.

11-6 Solid dielectrics

Now we turn to the solids. The first interesting fact about solids is that there can be a permanent polarization built in which exists even without applying an electric field. An example occurs with a material like wax, which contains long molecules having a permanent dipole moment. If you melt some wax and put a strong electric field on it when it is a liquid, so that the dipole moments get partly lined up, they will stay that way when the liquid freezes. The solid material will have a permanent polarization which remains when the field is removed. Such a solid is called an *electret*.

An electret has permanent polarization charges on its surface. It is the electrical analog of a magnet. It is not as useful, though, because free charges from the air are attracted to its surfaces, eventually cancelling the polarization charges. The electret is "discharged" and there are no visible external fields.

A permanent internal polarization P is also found occurring naturally in some crystalline substances. In such crystals, each unit cell of the lattice has an identical permanent dipole moment, as drawn in Fig. 11-8. All the dipoles point in the same direction, even with no applied electric field. Many complicated crystals have, in fact, such a polarization; we do not normally notice it because the external fields are discharged, just as for the electrets.

If these internal dipole moments of a crystal are changed, however, external fields appear because there is not time for stray charges to gather and cancel the polarization charges. If the dielectric is in a condenser, free charges will be induced on the electrodes. For example, the moments can change when a dielectric is heated, because of thermal expansion. The effect is called *pyroelectricity*. Similarly, if we change the stresses in a crystal - for instance, if we bend it - again the moment may change a little bit, and a small electrical effect, called *piezoelectricity*, can be detected.

For crystals that do not have a permanent moment, one can work out a theory of the dielectric constant that involves the electronic polarizability of the atoms. It goes much the same as for liquids. Some crystals also have rotatable dipoles inside, and the rotation of these dipoles will then contribute to ϵ . In ionic crystals such as NaCl there is also ion polarizability. The crystal consists of a checkerboard of positive and negative ions, and in an electric field the positive ions are pulled one way and the negatives the other; there is a net relative motion of the plus and minus charges, and so a volume polarization. We could estimate the magnitude of the ionic polarizability from our knowledge of the stiffness of salt crystals, but we will not go into that subject here.

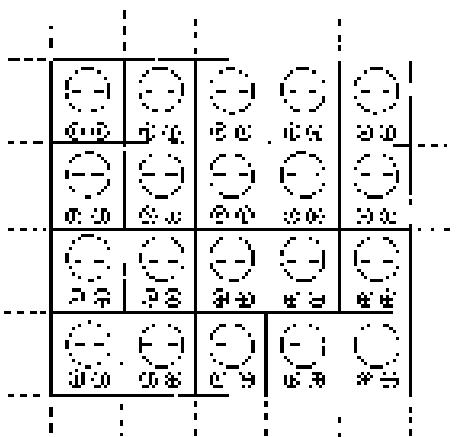


Fig. 11-8. A complex crystal lattice can have a permanent intrinsic polarization P .

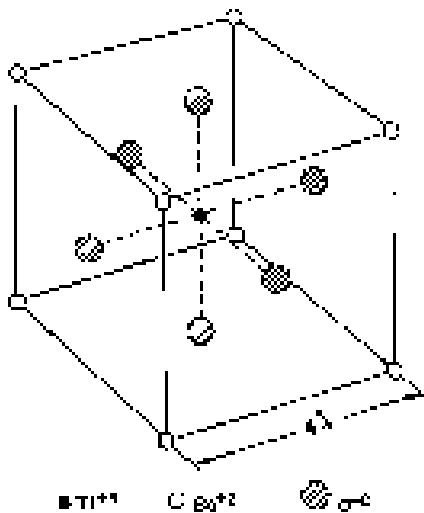


Fig. 11-9. The unit cell of BaTiO_3 . The atomic nuclei fill up most of the space; for clarity, only the positions of their centers are shown.

11-7 Ferroelectricity; BaTiO_3

We want to describe now one special class of crystals which have, just by accident almost, a built-in permanent moment. The situation is so marginal that if we increase the temperature a little bit they lose the permanent moment completely. On the other hand, if they are nearly cubic crystals, so that their moments can be turned in different directions, we can detect a large change in the moment when an applied electric field is changed. All the moments flip over and we get a large effect. Substances which have this kind of permanent moment are called *ferroelectric*, after the corresponding ferromagnetic effects which were first discovered in iron.

We would like to explain how ferroelectricity works by describing a particular example of a ferroelectric material. There are several ways in which the ferroelectric property can originate; but we will take up only one mysterious case - that of barium titanate, BaTiO_3 . This material has a crystal lattice whose basic cell is sketched in Fig. 11-9. It turns out that above a certain temperature, specifically 118°C, barium titanate is an ordinary dielectric with no enormous dielectric constant. Below this temperature, however, it suddenly takes on a permanent polarization.

In working out the polarization of solid material, we must first figure out the local fields in each unit cell. We must include the fields from the polarization

level, just as we did for the case of a liquid. But a crystal is not a homogeneous liquid, so we cannot use for the local field value we would get in a spherical shell. If you work it out for a crystal, you find that the factor $1/3$ in Eq. (11.28) becomes slightly different, but not far from $1/3$. (For a simple cubic crystal, it is just $2/3$.) We will, therefore, assume for our preliminary discussion that the factor is $1/3$ for BaTiO_3 .

Now when we wrote Eq. (11.28) you may have wondered what would happen if N_x became greater than 3 . It appears as though x would become negative. But that surely cannot be right. Let's see what should happen if we were gradually to increase σ in a ferroelectric crystal. As σ gets larger, the polarization gets bigger, making a bigger local field. But a bigger local field will polarize each atom more, raising the local field still more. If the "drive" of the atoms is enough, the process keeps going; there is a kind of feedback that causes the polarization to increase without limit, assuming that the polarization of each atom increases in proportion to the field. The "runaway" condition occurs when $N_x = 3$. The polarization does not become infinite, of course, because the proportionality between the induced moment and the electric field breaks down at high fields, so that our formulas are no longer correct. What happens is that the lattice gets "locked in" with a high, self-generated, internal polarization.

In the case of BaTiO_3 , there is, in addition to an electronic polarization, also a rather large ionic polarization, presumed to be due to titanium ions which will move a little within the cubic lattices. The lattice resists large motions, so after the titanium has gone a little way, it jams up and stops. But the crystal cell is then left with a permanent d-pole moment.

In most crystals, this is really the situation for all temperatures that can be reached. The very interesting thing about barium titanate is that there is such a delicate condition that if N_x is decreased just a little bit it comes back. Since N decreases with increasing temperature "because of thermal expansion" we can vary N_x by varying the temperature. Below the critical temperature it is just barely stuck, so it is easy -by applying an external field- to shift the polarization and have it back in a different direction.

Let's see if we can analyze what happens in more detail. We call T_c the critical temperature at which N_x is exactly 3 . As the temperature increases, N goes down a little bit because of the expansion of the lattice. Since the expansion is small, we can say that near the critical temperature

$$N_x = 3 - \beta(T - T_c), \quad (11.30)$$

where β is a small constant, of the same order of magnitude as the thermal expansion coefficient, or about 20^{-4} to 20^{-5} per degree C. Now if we substitute this relation into Eq. (11.28), we get that

$$x = 1 - \frac{3 - \beta(T - T_c)}{\beta(T - T_c)/3}.$$

Since we have assumed that $\beta(T - T_c)$ is small compared with one, we can approximate this formula by

$$x \approx 1 - \frac{9}{\beta(T - T_c)}. \quad (11.31)$$

This relation is right, of course, only for $T > T_c$. We see that just above the critical temperature x is enormous. Because N_x is so close to 3 , there is a tremendous magnetization effect, and the dielectric constant can easily be as high as 50,000 to 100,000. It is also very sensitive to temperature. For insulators, in temperature, the dielectric constant goes down inversely as the temperature, but, unlike the case of a dipolar gas, for which $x - 1$ goes inversely as the absolute temperature, for ferroelectrics it varies inversely as the difference between the absolute temperature and the critical temperature (this law is called the Curie-Weiss law).

When we lower the temperature to the critical temperature, what happens? If we imagine a lattice of unit cells like that in Fig. 11-9, we see that it is possible

to pick out chains of ions along vertical lines. One of the η 's consists of alternating oxygen and titanium ions. There are other lines made up of either barium or oxygen ions, but the spacing along these lines is greater. We make a simple model to simulate this situation by imagining, as shown in Fig. 11-10(a), a series of chains of ions. Along what we call the main chain, the separation of the ions is a , which is half the lattice constant; the lateral distance between identical chains is $2a$. There are less-dense chains in between which we will ignore for the moment. To make the analysis a little easier, we will also suppose that all the ions on the main chain are identical. (It's not a very simple simplification because all the important effects will still appear. This is one of the tricks of theoretical physics. One does a different problem because it's easier to figure out the first time, then when one understands how the thing works, it's time to put in all the complications.)

Now let's try to find out what would happen with our model. We suppose that the dipole moment of each atom is p and we wish to calculate the field at one of the atoms of the chain. We ignore all the fields from all the other atoms. We will first calculate the field from the dipoles in only one vertical chain; we will talk about the other chains later. The field at the distance r from a dipole in a direction along its axis is given by

$$E = \frac{1}{4\pi\epsilon_0} \frac{2p}{r^3}. \quad (11.32)$$

(a)

At any given atom, the dipoles at equal distances above and below it give fields in the same direction, so for the whole chain we get

$$E_{\text{chain}} = \frac{p}{4\pi\epsilon_0} \frac{2}{a} \cdot \left(1 + \frac{2}{3} + \frac{2}{27} + \frac{2}{64} + \dots \right) = \frac{p}{\epsilon_0} \frac{0.383}{a^2}. \quad (11.33)$$

It is not too hard to show that if our model were like a completely cubic crystal—that is, if the next identical lines were only the distance a away—the number 0.383 would be changed to 1/3. In other words, if the next lines were at the distance a they would contribute only -0.033 unit to our sum. (However, the next main chain we are considering is at the distance $2a$ and, as you remember from Chapter 7, the field from a periodic structure dies off exponentially with distance. Therefore these lines contribute much less than -0.033 and we can just ignore all the other chains.)

It is necessary now to find out what polarizability α is needed to make the runaway process work. Suppose that the induced moment ρ of each atom of the chain is proportional to the field on it, as in Eq. (11.6). We get the polarizing field on the atom from E_{chain} using Eq. (11.32). So we have the two equations

$$\rho = \alpha E_{\text{chain}}$$

and

$$E_{\text{chain}} = \frac{0.383}{a^2} \frac{\rho}{\epsilon_0}.$$

There are two solutions: E and ρ both zero, or

$$\alpha = \frac{a^2}{0.383},$$

with E and ρ both finite. Thus if α is as large as $a^2/0.383$, a permanent polarization established by its own field will set in. This critical equality must be reached for barium titanate at just the temperature T_c . (Notice that if α were larger than the critical value for small fields, it would decrease at larger fields and at equilibrium the same equality we have found would hold.)

For BaTiO_3 , the spacing a is 3×10^{-8} cm, so we never expect that $\alpha = 21.8 \times 10^{-41} \text{ cm}^2$. We can compare this with the known polarizabilities of the individual atoms. The oxygen, $\alpha = 30.2 \times 10^{-41} \text{ cm}^2$; we're on the right track! But barium, $\alpha = 2.4 \times 10^{-41} \text{ cm}^2$; rather small. To use our model we should probably take the average. (We could work out the chain again for alternating

atoms, but the result would be about the same. So $\alpha_{\text{average}} = 16.2 \times 10^{-21}$, which is not high enough to give a permanent polarization.

But wait a moment! We have so far only added up the electronic polarizabilities. There is also some ionic polarization due to the motion of the titanium ion. All we need is an ionic polarizability of $9.1 \times 10^{-24} \text{ cm}^3$. (A more precise computation using alternating atoms shows that actually 11.9×10^{-24} is needed.) To understand the properties of BaTiO₃, we have to assume that such an ionic polarizability exists.

Why the titanium ion in barium titanate should have that much ionic polarizability is not known. Furthermore, why, at a lower temperature, it polarizes along the cube diagonal and the face diagonal equally well is not clear. If we figure out the actual size of the spheres in Fig. 11-9, and ask whether the titanium is a little bit loose in the box formed by its neighboring oxygen atoms—which is what you would hope, so that it could be easily shifted—you find quite the contrary. It fits very tightly. The barium atoms are slightly loose, but if you let them be the ones that move, it doesn't work at all. So you see that the subject is really not one-hundred percent clear; there are still mysteries we would like to understand.

Returning to our simple model of Fig. 11-10(a), we see that the field from one chain would tend to polarize the neighboring chain in the opposite direction, which means that although each chain would be locked, there would be no net permanent moment per unit volume. (Although there would be no external electric effect, there are still certain thermodynamic effects one could observe.) Such systems exist, and are called antiferroelectric. Barium titanate, however, is really like the arrangement in Fig. 11-10(b). The oxygen-titanium chains are all polarized in the same direction because there are intermediate chains of atoms in between. Although the atoms in these chains are not very polarizable, or very dense, they will be somewhat polarized, in the direction parallel to the oxygen-titanium chains. The small dipoles produced at the real oxygen-titanium chains will get it started parallel to the first. So BaTiO₃ is really ferroelectric, and it is because of the atoms in between. You may be wondering: "But what about the direct effect between the two O-Ti chains?" Remember, though, the direct effect dies off exponentially with the separation; the effect of the chain of strong dipoles at $2a$ can be less than the effect of a chain of weak ones at the distance a .

This completes our rather sketchy report on our present understanding of the dielectric constants of gases, of liquids, and of solids.

Electrostatic Analogs

12-1 The same equations have the same solutions

The total amount of information which has been acquired about the physical world since the beginning of scientific progress is enormous, and it seems almost impossible that any one person could know a reasonable fraction of it. But it is actually quite possible for a physicist to retain a broad knowledge of the physical world rather than to become a specialist in some narrow area. The reasons for this are threefold. First, there are great principles which apply to all the different kinds of phenomena—such as the principles of the conservation of energy and of angular momentum. A thorough understanding of such principles gives an understanding of a great deal all at once. Second, there is the fact that many complicated phenomena, such as the behavior of solids under compression, only partially depend on electrical and quantum-mechanical forces, so that if one understands the fundamental laws of electricity and quantum mechanics, there is at least some possibility of understanding many of the phenomena that occur in complex situations. Finally, there is a most remarkable coincidence: The equations for many different physical situations have exactly the same appearance. Of course, the symbols may be different—one letter is substituted for another—but the mathematical form of the equations is the same. This means that having studied one subject, we immediately have a great deal of direct and precise knowledge about the solutions of the equations of another.

We are now finished with the subject of electrostatics, and will soon go on to study magnetism and electrodynamics. But before doing so, we would like to show that while learning electrostatics we have simultaneously learned about a large number of other subjects. We will find that the equations of electrostatics appear in several other places in physics. By a direct translation of the solutions (of course the same mathematical equations must have the same solutions), it is possible to solve problems in other fields with the same ease—or with the same difficulty—as in electrostatics.

The equations of electrostatics, we know, are

$$\nabla \cdot (\epsilon \mathbf{E}) = \frac{\rho_{\text{free}}}{\epsilon_0} . \quad (12.1)$$

$$\nabla \times \mathbf{E} = 0 . \quad (12.2)$$

(We take the equations of electrostatics with dielectrics as to have the most general situation.) The same physics can be expressed in another mathematical form:

$$\mathbf{E} = -\nabla \phi , \quad (12.3)$$

$$\nabla \cdot (\epsilon \nabla \phi) = -\frac{\rho_{\text{free}}}{\epsilon_0} . \quad (12.4)$$

Now the point is that there are many physics problems where mathematical equations have the same form. There is a potential (ϕ) whose gradient multiplied by a scalar function (ϵ) has a divergence equal to another scalar function ($-\rho/\epsilon_0$).

Whatever we know about electrostatics can immediately be carried over into that other subject, and vice versa. (It works both ways, of course—if the other subject has some particular characteristics that are known, then we can apply that knowledge to the corresponding electrostatic problem.) We want to consider a series of examples from different subjects that produce equations of this form.

12-2 The same equations have the same solutions

12-3 The stretched membrane

12-4 The diffusion of neutrinos; a uniform spherical source in a homogeneous medium

12-5 Rotating fluid flow; the flow past a sphere

12-6 Illumination; the uniform lighting of a plane

12-7 The "underlying unity" of nature

12-1 The flow of heat; a point source near an infinite plane boundary

We have discussed one example earlier (Section 5-4)—the flow of heat. Imagine a block of material, which does not be homogeneous but may consist of different materials at different places, in which the temperature varies from point to point. As a consequence of these temperature variations there is a flow of heat, which can be represented by the vector \mathbf{h} . It represents the amount of heat energy which flows per unit time through a unit area perpendicular to the flow. The divergence of \mathbf{h} represents the rate per unit volume at which heat is leaving a region:

$$\nabla \cdot \mathbf{h} = \text{rate of heat out per unit volume.}$$

(We could, of course, write the equation in integral form, just as we did in electrodynamics with Gauss' law—which would say that the flux through a surface is equal to the rate of change of heat energy inside the material. We will not bother to translate the equations back and forth between the differential and the integral forms, because it goes exactly the same as in electrodynamics.)

The rate at which heat is generated or absorbed at various places depends, of course, on the origin. Suppose, for example, that there is a source of heat inside the material (perhaps a hot water source, or a resistor heated by an electric current). Let us call s the heat energy produced per unit volume per second by this source. There may also be losses (or gains) of heat energy in other internal energies in the volume. If ω is the internal energy per unit volume, $d\omega/dt$ will also be a “source” of heat energy. We have, then,

$$\nabla \cdot \mathbf{h} = s - \frac{d\omega}{dt}. \quad (12.5)$$

We are not going to discuss just now the complete equation in which things change with time, because we are making an analogy to electrodynamics, where nothing depends on the time. We will consider only steady heat-flow problems, in which constant sources have produced an equilibrium state. In these cases,

$$\nabla \cdot \mathbf{h} = s. \quad (12.6)$$

It is, of course, necessary to have another equation, which describes how the heat flows at various places. In many materials the heat current is approximately proportional to the rate of change of the temperature with position: the larger the temperature difference, the more the heat current. As we have seen, the vector heat current is proportional to the temperature gradient. The constant of proportionality K , a property of the material, is called the thermal conductivity.

$$h = -K \nabla T. \quad (12.7)$$

If the properties of the material vary from place to place, then $K = K(x, y, z)$, a function of position. [Equation (12.7) is not as fundamental as (12.5), which expresses the conservation of heat energy, since the former depends upon a special property of the substance.] If now we substitute Eq. (12.7) into Eq. (12.6) we have

$$\nabla \cdot (K \nabla T) = -s. \quad (12.8)$$

which has exactly the same form as (12.4). Steady heat-flow problems and electromagnetic problems are the same. The heat flow vector \mathbf{h} corresponds to \mathbf{E} , and the temperature T corresponds to ϕ . We have already decided that a point heat source produces a temperature field which varies as $1/r$ and a heat flow which varies as $1/r^2$. This is nothing more than a translation of the statement from electrodynamics that a point charge generates a potential which varies as $1/r$ and an electric field which varies as $1/r^2$. We can, in general, solve static heat problems as easily as we can solve electrostatic problems.

Consider a simple example. Suppose that we have a cylinder of radius a at the temperature T_1 , maintained by the generation of heat in the cylinder. (It could be, for example, a wire carrying a current, or a pipe with steam condensing inside.)

The cylinder is covered with a concentric sheath of insulating material which has a conductivity K . Say the outside radius of the insulation is b and the outside is kept at temperature T_2 (Fig. 12-1a). We want to find out at what rate heat will be lost by the pipe, or steam pipe, or whatever it is in the ocean. Let the total amount of heat lost from a length L of the pipe be called Q —which is what we are trying to find.

How can we solve this problem? We have two differential equations, but since these are the same as those of electrodynamics, we have really already solved the mathematical problem. The analogous problem is that of a conductor of radius a at the potential ϕ_1 , separated from another conductor of radius b at the potential ϕ_2 , with a conductive layer of dielectric material in between, as drawn in Fig. 12-1(b). Now since the heat flow A corresponds to the electric field E , the quantity Q that we want to find corresponds to the flux of the electric field from a unit length. In other words, to the electric charge per unit length over L . We have solved the electrostatic problem by using Gauss' law. We follow the same procedure for our heat-flow problem.

From the symmetry of the situation, we know that A depends only on the distance from the center. So we enclose the pipe in a gaussian cylinder of length L and radius r . From Gauss' law, we know that the heat flow A multiplied by the area $2\pi rL$ of the surface must be equal to the total amount of heat generated inside, which is what we are calling Q .

$$2\pi r L k = Q \quad \text{or} \quad A = \frac{Q}{2\pi r L}. \quad (12.9)$$

The heat flow is proportional to the temperature gradient:

$$A = -K \nabla T,$$

or, in this case, the magnitude of A is

$$A = -K \frac{dT}{dr}.$$

This, together with (12.9), gives

$$\frac{dT}{dr} = -\frac{Q}{2\pi K L r}. \quad (12.10)$$

Differentiating from $r = a$ to $r = b$, we get

$$T_2 - T_1 = -\frac{Q}{2\pi K L} \ln \frac{b}{a}. \quad (12.11)$$

Solving for Q , we find

$$Q = \frac{2\pi K L (T_2 - T_1)}{\ln(b/a)}. \quad (12.12)$$

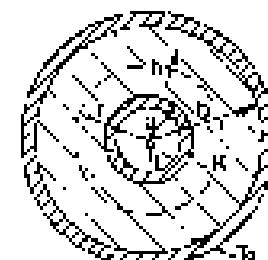
This result corresponds exactly to the result for the charge on a cylindrical condenser:

$$Q = \frac{2\pi e_0 L (\phi_2 - \phi_1)}{\ln(b/a)}.$$

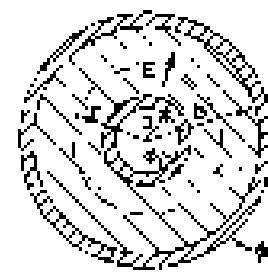
The problems are the same, and they have the same solutions. From our knowledge of electrodynamics, we also know how much heat is lost by an insulated pipe.

Let's consider another example of heat flow. Suppose we wish to know the heat flow in the neighborhood of a point source of heat located a little way beneath the surface of the earth, or near the surface of a large metal block. The idealized heat source might be an atomic bomb that was set off underground, leaving an intense source of heat, or it might correspond to a small radioactive source inside a block of iron—there are numerous possibilities.

We will treat the idealized problem of a point heat source of strength q at the distance a beneath the surface of an infinite block of uniform material whose thermal conductivity is K . And we will neglect the thermal conductivity of the



(a)



(b)

Fig. 12-1. (a) Heat flow in a cylindrical geometry. (b) The corresponding electrical problem.

air outside the material. We want to determine the distribution of the temperature on the surface of the block. How far is it right above the source and at various places on the surface of the block?

How shall we solve it? It is like an electrostatic problem with two materials with different dielectric coefficients ϵ on opposite sides of a plane boundary. And? Perhaps it is the analog of a point charge near the boundary between a dielectric and a conductor, or something similar. Let's see what the situation is near the surface. The physical condition is that the normal component of \mathbf{E} on the surface is zero, since we have assumed there is no heat flow out of the block. We should ask: In what electrostatic problem do we have the condition that the normal component of the electric field \mathbf{E} (which is the analog of \mathbf{h}) is zero at a surface? There is none!

That is one of the things that we have to watch out for. For physical reasons, there may be certain restrictions in the kinds of mathematical conditions which arise in any one subject. So if we have analyzed the differential equation only for certain limited cases, we may have missed some kinds of solutions that can occur in other physical situations. For example, there is no material with a dielectric constant of zero, whereas a vacuum does have zero thermal conductivity. So there is no electrostatic analogy for a perfect heat insulator. We can, however, still use the same method. We can try to imagine what would happen if the dielectric constant were zero. (Of course, the dielectric constant is never zero in any real situation. But we might have a case in which there is a material with a very high dielectric constant, so that we could neglect the dielectric constant of the air outside.)

How shall we find an electric field that has an component perpendicular to the surface? That is, one which is always zero at the surface? You will notice that our problem is opposite to the one of a point charge near a plane conductor. There we wanted the field to be perpendicular to the surface, because the conductor was all at the same potential. In the electrical problem, we invented a solution by imagining a point charge behind the conducting plate. We can use the same idea again. We try to pick an "image source" that will automatically make the normal component of the field zero at the surface. The solution is shown in Fig. 12-2. An image source of the same sign and the same strength placed at the distance a above the surface will cause the field to be always tangential at the surface. The normal components of the two sources cancel out.

Thus our first heat problem is solved. The temperature everywhere is the same, by direct analogy, as the potential due to two equal point charges. The temperature T at the distance r from a single point source G to a infinite medium is

$$T = \frac{G}{4\pi K r} \quad (12.13)$$

(This, of course, is just the analog of $\phi = q/4\pi\epsilon_0 r$.) The temperature for a point source, together with its image source, is

$$T = \frac{G}{4\pi K r_1} + \frac{G}{4\pi K r_2} \quad (12.14)$$

This formula gives us the temperature everywhere in the block. Several isothermal surfaces are shown in Fig. 12-2. Also shown are lines of \mathbf{h} , which can be obtained from $\mathbf{h} = -K \nabla T$.

We originally asked for the temperature distribution on the surface. For a point on the surface at the distance r from the source, $r_1 = r_2 = \sqrt{r^2 - a^2}$, so

$$T(\text{surface}) = \frac{1}{4\pi K} \frac{2G}{\sqrt{r^2 - a^2}} \quad (12.15)$$

This function is also shown in the figure. The temperature is, naturally, higher right above the source than it is farther away. This is the kind of problem that geophysicists often need to solve. We now see that it is the same kind of thing we have already been solving for electricity.

12-3 The stretched membrane

Now let us consider a completely different physical situation which, nevertheless, gives the same equations again. Consider a thin rubber sheet—a membrane—which has been stretched over a large horizontal frame (like a drum-head). Suppose now that the membrane is pushed up in one place and down in another; as shown in Fig. 12-3. Can we describe the shape of the surface? We will show how the problem can be solved when the deflections of the membrane are not too large.

There are forces in the sheet because it is stretched. If we were to make a small cut anywhere, the two sides of the cut would pull apart (see Fig. 12-2). So there is a surface tension in the sheet, analogous to the one-dimensional tension in a stretched string. We define the magnitude of the surface tension τ as the force per unit length which will just hold together the two sides of a cut such as one of those shown in Fig. 12-4.

Suppose now that we look at a vertical cross section of the membrane. It will appear as a curve, like the one in Fig. 12-4. Let u be the vertical displacement of the membrane from its normal position, and x and y the coordinates in the horizontal plane. (The cross section shows a parallel to the x -axis.)

Consider a little piece of the surface of length Δx and width Δy . There will be forces on the piece from the surface tension along each edge. The force along edge 1 (at the figure will be $\tau_1 \Delta y$, directed tangent to the surface—that is, at the angle θ_1 from the horizontal). Along edge 2, the force will be $\tau_2 \Delta y$ at the angle θ_2 . (There will be similar forces on the other two edges of the piece, but we will forget them for the moment.) The net upward force on the piece from edges 1 and 2 is

$$\Delta F = \tau_2 \Delta y \sin \theta_2 - \tau_1 \Delta y \sin \theta_1.$$

We will limit our considerations to small distortions of the membrane, i.e., to small slopes; we can then replace $\sin \theta$ by $\tan \theta$, which can be written as $\partial u / \partial x$. The force is then

$$\Delta F = \left[\tau_2 \left(\frac{\partial u}{\partial x} \right)_2 - \tau_1 \left(\frac{\partial u}{\partial x} \right)_1 \right] \Delta x \Delta y.$$

The quantity in brackets can be equally well written (for small Δx) as

$$\frac{\partial}{\partial x} \left(\tau \frac{\partial u}{\partial x} \right) \Delta x \Delta y,$$

then

$$\Delta F = \frac{\partial}{\partial x} \left(\tau \frac{\partial u}{\partial x} \right) \Delta x \Delta y.$$

There will be another contribution to ΔF from the forces at the other two edges; the total is evidently

$$\Delta F = \left[\frac{\partial}{\partial x} \left(\tau \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left(\tau \frac{\partial u}{\partial y} \right) \right] \Delta x \Delta y. \quad (12.16)$$

The distortions of the diaphragm are caused by external forces. Let's let f represent the upward force per unit area on the sheet (a kind of "pressure") from the external forces. When the membrane is in equilibrium (the static case), this force must be balanced by the internal force we have just computed, Eq. (12.16). That is

$$f = -\frac{\Delta F}{\Delta x \Delta y}.$$

Equation (12.16) can also be written

$$f = -\nabla \cdot (\gamma \nabla u), \quad (12.17)$$

where by ∇ we now mean, of course, the two-dimensional gradient operator ($\partial / \partial x, \partial / \partial y$). We have the differential equation that relates $\partial u / \partial x, y$ to the applied

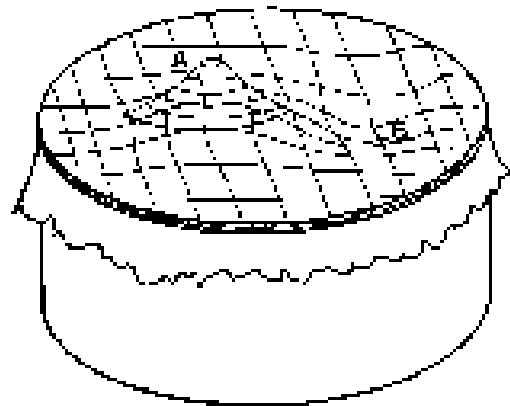


Fig. 12-3. A thin rubber sheet stretched over a cylindrical frame like a drumhead. If the sheet is pulled up at A and down at B, what is the shape of the surface?

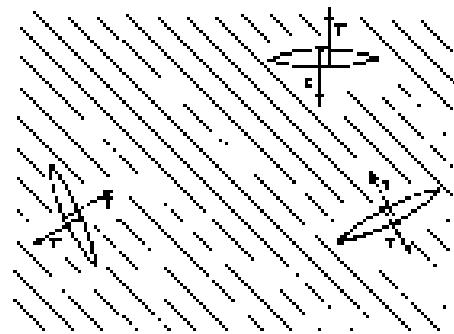


Fig. 12-4. The surface tension τ of a stretched rubber sheet is the force per unit length across a line.

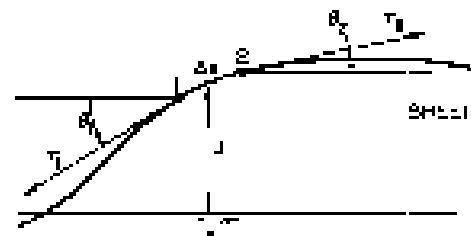


Fig. 12-5. Cross section of the deflected sheet.

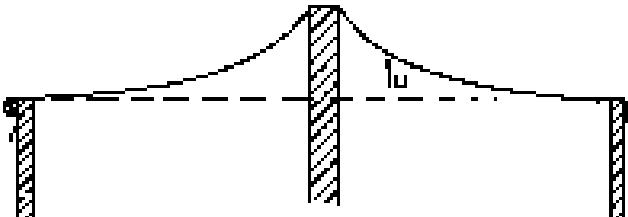
force $f(x, y)$ and the surface tension $\tau(x, y)$, which may, in general, vary from place to place in the sheet. (The distortions of a three-dimensional elastic body are also governed by similar equations, but we will stick to two-dimensions.) We will worry only about the case in which the tension τ is constant throughout the sheet. We can then write for Eq. (12.17a)

$$\nabla^2 u = -\frac{f}{\tau}. \quad (12.18)$$

We have another equation that is the same as for electrostatics!—only this time, limited to two-dimensions. The displacement u corresponds to ϕ , and f/τ corresponds to a/ϵ_0 . So all the work we have done for infinite plane charged sheets, or long parallel wires, or charged cylinders is directly applicable to the stretched membrane.

Suppose we push the membrane at some points up to a definite height—that is, we fix the value of u at some places. That is the analog of having a definite potential at the corresponding places in an electrical situation. So, for instance, we may make a positive "potential" by pushing up on the membrane with an object having the cross-sectional shape of the corresponding cylindrical conductor. For example, if we push the sheet up with a round rod, the surface will take on the shape shown in Fig. 12-6. The height u is the same as the electrostatic potential ϕ of a charged cylindrical rod. It falls off as in (12.1). (The slope, which corresponds to the electric field E , drops off as $1/r$.)

Fig. 12-6. Cross section of a stretched rubber sheet pushed up by a round rod. The function $u(x, y)$ is the same as the electric potential $\phi(x, y)$ near a very long charged rod.



The stretching method has often been used as a way of solving complicated electrical problems experimentally. The analogy is used backwards: various rods and bars are pushed against the sheet so heights that correspond to the potentials of a set of electrodes. Measurements of the height then give the electrical potential for the electrical situation. The analogy has been carried even further. If little balls are placed on the membrane, their motion corresponds approximately to the motion of electrons in the corresponding electric field. One can actually watch the "electrons" move on their trajectories. This method was used to design the complicated geometry of many photomultiplier tubes (such as the ones used for scintillation counters, and the one used for controlling the headlight beams on Cadillacs). The method is still used, but the accuracy is lower. For the most accurate work, it is better to determine the fields by numerical methods, using the large electronic computing machines.

12-4 The diffusion of neutrons; a uniform spherical source in a homogeneous medium

We take another example that gives the same kind of equation, this time having to do with diffusion. In Chapter 49 of Vol. I we considered the diffusion of ions in a single gas, and of one gas through another. This time, let's take a different example—the diffusion of neutrons in a material like graphite. We choose to speak of graphite (a pure form of carbon) because carbon doesn't absorb slow neutrons. In it, the neutrons are free to wander around. They travel in a straight line for several centimeters, on the average, before being scattered by a nucleus and deflected into a new direction. So if we have a large block—many meters on a side—the neutrons initially at one place will diffuse to other places. We want to find a description of their average behavior—that is, their average flux.

Let $N(x, y, z)$ & ΔV be the number of neutrons in the element of volume ΔV at the point (x, y, z) . Because of their motion, some neutrons will be leaving ΔV , and others will be coming in. If there are more neutrons in one region than in a nearby region, more neutrons will go from the first region to the second than come back; there will be a net flow. Following the arguments of Chapter 40 in Vol. I, we describe the flow by a flow vector J . The x -component J_x is the net number of neutrons that pass in unit time a unit area perpendicular to the x -direction. We find that

$$J_x = -D \frac{\partial N}{\partial x}, \quad (12.19)$$

where the diffusion constant D is given in terms of the mean velocity v , and the mean-free-path λ between scatterings is given by

$$D = \frac{1}{3} v \lambda.$$

The vector equation for J is

$$\mathbf{J} = -D \nabla N. \quad (12.20)$$

The rate at which neutrons flow across any surface element $d\mathbf{a}$ is $J \cdot d\mathbf{a}$ (where, as usual, \mathbf{n} is the unit normal). The net flow out of a volume element is then (following the usual gaussian argument) $\mathbf{n} \cdot J dV$. This flow would result in a decrease with time of the number N of ΔV unless neutrons are being created in ΔV (by some nuclear process). If there are sources in the volume that generate S neutrons per unit time in a unit volume, then the net flow out of ΔV will be equal to $(S - \delta N/\delta t) \Delta V$. We have then that

$$\nabla \cdot \mathbf{J} = S - \frac{\delta N}{\delta t}. \quad (12.21)$$

Combining (12.21) with (12.20), we get the neutron diffusion equation

$$\nabla \cdot (-D \nabla N) = S - \frac{\delta N}{\delta t}. \quad (12.22)$$

In the static case—where $\delta N/\delta t = 0$ —we have Eq. (12.4) all over again! We can use our knowledge of electrostatics to solve problems about the diffusion of neutrons. So let's solve a problem. (You may wonder: Why do a problem if we have already done its two problems in electrostatics? We can do it faster this time because we have done the electrostatic problems!)

Suppose we have a block of graphite in which neutrons are being generated, say, by uranium fission—uniformly throughout a spherical region of radius a (Fig. 12-7). We would like to know: What is the density of neutrons everywhere? How uniform is the density of neutrons in the region where they are being generated? What is the ratio of the neutron density at the center to the neutron density at the surface of the source region? Finding the answers is easy. The source density S replaces the charge density ρ , so our problem is the same as the problem of a sphere of uniform charge density. Finding N is just like finding the potential ϕ . We have already worked out the fields inside and outside of a uniformly charged sphere; we can integrate them to get the potential. Outside, the potential is $Q/4\pi r^2$, with the total charge Q given by $4\pi a^3 \rho/3$. So

$$\phi_{\text{outside}} = \frac{\rho a^3}{3 \epsilon_0}, \quad (12.23)$$

For points inside, the field is due only to the charge $Q(r)$ inside the sphere of radius r , $Q(r) = 4\pi r^3 \rho/3$, so

$$\mathbf{E} = \frac{\rho r^2}{3 \epsilon_0}, \quad (12.24)$$

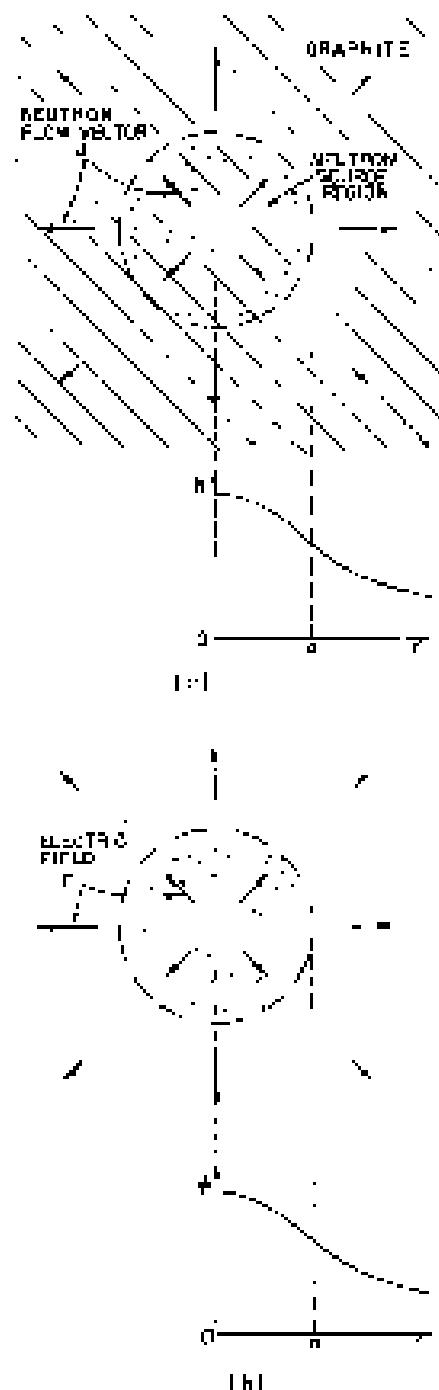


Fig. 12-7. (a) Neutrons are produced uniformly throughout a sphere of radius a in a large graphite block and diffuse outward. The neutron density N is found as a function of r , the distance from the center of the source. (b) The analogous electrostatic situation: a uniform sphere of charge, where N corresponds to ϕ and J corresponds to E .

The field increases linearly with r . Integrating E to get ϕ , we have

$$\phi_{\text{center}} = -\frac{\rho r^2}{6\epsilon_0} + \text{a constant.}$$

At the radius a , ϕ_{center} must be the same as ϕ_{surface} , so the constant must be $\rho a^2/6\epsilon_0$. (We are assuming that ϕ is zero at large distances from the source, which will correspond to N being zero for the densities.) Therefore,

$$\phi_{\text{surface}} = \frac{\rho}{6\epsilon_0} \left(\frac{3a^2}{2} - \frac{r^2}{2} \right). \quad (12.25)$$

We know immediately the neutron density in our other problem. The answer is

$$N_{\text{center}} = \frac{3a^3}{32\pi}, \quad (12.26)$$

and

$$N_{\text{surface}} = \frac{3}{16} \left(\frac{3a^3}{2} - \frac{r^3}{2} \right). \quad (12.27)$$

N is shown as a function of r in Fig. 12-7.

Now what is the ratio of density at the center to that at the edge? At the center ($r = 0$), it is proportional to $3a^3/2$. At the edge ($r = a$) it is proportional to $2a^3/2$, so the ratio of densities is $3/2$. A uniform source doesn't produce a uniform density of neutrons. You see, our knowledge of electrostatics gives us a good start in the physics of nuclear reactors.

There are many physical circumstances in which diffusion plays a big part. The motion of ions through a liquid, or of electrons through a semiconductor, follows the same equation. We find again and again the same equations.

12-5 Irrotational fluid flow; the flow past a sphere

Let's now consider an example which is not really a very good one, because the equations we will use will not really represent the subject with complete generality but only in an artificial idealized situation. We take up the problem of water flow. In the case of the stretched sheet, our arguments were an approximation which we could only for small deflections. For our consideration of water flow, we will not make that kind of an approximation; we must make restrictions that do not apply at all to real water. We treat only the case of the steady flow of an incompressible, nonviscous, circulatory-free liquid. Then we represent the flow by giving the velocity $v(r)$ as a function of position r . If the motion is steady (the only case for which there is an electrostatic analog) v is independent of time. If ρ is the density of the fluid, then ρv is the amount of mass which passes per unit time through a unit area. By the conservation of matter, the divergence of ρv will be, in general, the time rate of change of the mass of the material per unit volume. We will assume that there are no processes for the continuous creation or destruction of matter. The conservation of matter then requires that $\nabla \cdot \rho v = 0$. (It should, in general, be equal to $-\partial \rho / \partial t$, but since our fluid is incompressible, ρ cannot change.) Since v is everywhere the same, we can factor it out, and our equation is simply

$$\nabla \cdot v = 0.$$

Good! We have electrostatics again (with no charges); it's just like $\nabla \cdot E = 0$. Not so! Electrostatics is not simply $\nabla \cdot E = 0$. It is a pair of equations. One equation does not tell us enough; we need still an additional equation. To match electrostatics, we should have also that the curl of v is zero. But that is not generally true for real liquids. Most liquids will ordinarily develop some circulation. So let's not restrict to the situation in which there is no circulation of the fluid. Such flow is often called *irrotational*. Anyway, if we make all our assumptions, we can

imagine a case of fluid flow that is analogous to electrostatics. So we take

$$\nabla \cdot v = 0 \quad (12.28)$$

and

$$\nabla \times v = 0. \quad (12.29)$$

We want to emphasize that the number of circumstances in which liquid flow follows these equations is far from the great majority, but there are a few. They must be cases in which we can neglect surface tension, compressibility, and viscosity, and in which we can assume that the flow is irrotational. These assumptions are valid so rarely for real water that the mathematician Jean von Neumann said that people who analyze Eqs. (12.28) and (12.29) are studying "dry water"! (We take up the problem of fluid flow in more detail in Chapters 19 and 40.)

Because $\nabla \times v = 0$, the velocity of "dry water" can be written as the gradient of some potential:

$$v = -\nabla \phi. \quad (12.30)$$

What is the physical meaning of ϕ ? There isn't any very useful meaning. The velocity can be written as the gradient of a potential simply because the flow is irrotational. And by analogy with electrostatics, ϕ is called the velocity potential, but it is not related to a potential energy in the way that ϕ is. Since the divergence of v is zero, we have

$$\nabla \cdot (\nabla \phi) = \nabla^2 \phi = 0. \quad (12.31)$$

The velocity potential ϕ obeys the same differential equation as the electrostatic potential in free space ($\rho = 0$).

Let's pose a problem in irrotational flow and see whether we can solve it by the methods we have learned. Consider the problem of a spherical ball falling through a liquid. If it is going too slowly, the viscous forces, which we are disregarding, will be important. If it is going too fast, little whirlpools (turbulence) will appear in its wake and there will be some retardation of the water. But if the ball is going neither too fast nor too slow, it is more or less true that the water flow will fit our assumptions, and we can describe the motion of the water by our simple equations.

It is convenient to describe what happens in a frame of reference fixed in the sphere. In this frame we are asking the question: How does water flow past a sphere at rest when the flow at large distances is uniform? That is, when far from the sphere, the flow is everywhere law-like. The flow near the sphere will be as shown by the streamlines drawn in Fig. 12-8. These lines, always parallel to v , correspond to lines of electric field. We want to get a quantitative description for the velocity field, i.e., an expression for the velocity at any point P .

We can find the velocity from the gradient of ϕ , so we first work out the potential. We want a potential that satisfies Eq. (12.31) everywhere, and which also satisfies two restrictions: (1) There is no flow in the spherical region inside the surface of the ball, and (2) the flow is constant at large distances. To satisfy (1), the component of v normal to the surface of the sphere must be zero. That means that $\partial \phi / \partial r$ is zero at $r = a$. To satisfy (2), we must have $\partial \phi / \partial z = v_0$ at all points where $r \gg a$. Strictly speaking, there is no electrostatic case which corresponds exactly to our problem. It really corresponds to putting a sphere of dielectric constant zero in a uniform electric field. If we had worked out the solution to the problem of a sphere of a dielectric constant κ in a uniform field, then by putting $\kappa = 0$ we would immediately have the solution to this problem.

We have not actually worked out this particular electrostatic problem in detail, but let's do it now. (We could work directly on the fluid problem with v and ϕ , but we will use E and ϕ because we are so used to them.)

The problem is: Find a solution of $\nabla^2 \phi = 0$ such that $E = -\nabla \phi$ is a constant, say E_0 for large r , and such that the radial component of E is equal to zero at $r = a$. That is,

$$\frac{\partial}{\partial r} \left(\frac{1}{r^2} \frac{\partial \phi}{\partial r} \right) = 0. \quad (12.32)$$

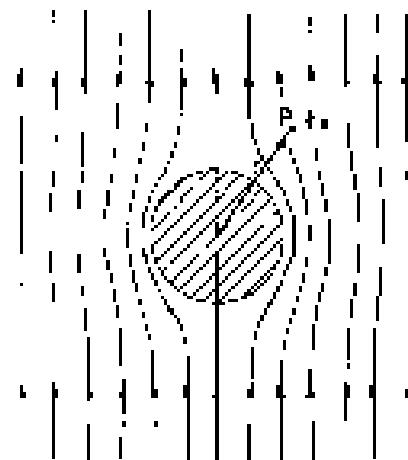


Fig. 12-8. The velocity field of irrotational fluid flow past a sphere.

Our problem involves a new kind of boundary condition, not one for which ϕ is a constant on a surface, but for which $\partial\phi/\partial r$ is a constant. That is a little different. It is not easy to get the answer immediately. First of all, without the sphere, ϕ would be $-E_0 z$. Then E would be in the z -direction and have the constant magnitude E_0 , everywhere. Now we have analyzed the case of a dielectric sphere which has a uniform polarization inside it, and we found that the field inside such a polarized sphere is a uniform field, and that outside it is the same as the field of a point dipole located at the center. So let's guess that the solution we want is a superposition of a uniform field plus the field of a dipole. The potential of a dipole (Chapter 6) is $p_z/4\pi\epsilon_0 r^3$. Thus we assume that

$$\phi = -E_0 z + \frac{p_z}{4\pi\epsilon_0 r^3}. \quad (12.33)$$

Since the dipole field falls off as $1/r^3$, at large distances we have just the field E_0 . Our guess will automatically satisfy condition (2) above. But what do we take for the dipole strength p_z ? To find out, we may use the other boundary condition, Eq. (12.32). We must differentiate ϕ with respect to r , but of course we never do see a constant angle θ , so it is more convenient if we first express ϕ in terms of r and θ , rather than r and θ . Since $z = r \cos \theta$, we get

$$\phi = -E_0 r \cos \theta + \frac{p_z \cos \theta}{4\pi\epsilon_0 r^2}. \quad (12.34)$$

The radial component of E is

$$-\frac{\partial \phi}{\partial r} = E_0 \cos \theta + \frac{p_z \cos \theta}{2\pi\epsilon_0 r^3}. \quad (12.35)$$

This must be zero at $r = a$ for all θ . This will be true if

$$p_z = -2\pi\epsilon_0 a^3 E_0. \quad (12.36)$$

Note carefully that if both terms in Eq. (12.35) had not had the same θ -dependence, it would not have been possible to choose p_z so that (12.35) turned out to be zero at $r = a$ for all angles. The fact that it works out means that we have guessed wisely in writing Eq. (12.33). Of course, when we made the guess we were looking ahead; we knew that we would need another term that (a) satisfied $\nabla^2 \phi = 0$ (any real field would do that), (b) dependent on $\cos \theta$, and (c) fall to zero at large r . The dipole field is the only one that does all three.

Using (12.36), our potential is

$$\phi = E_0 r \cos \theta \left(r + \frac{a^3}{2r^2} \right). \quad (12.37)$$

The solution of the fluid flow problem can be written simply as

$$\psi = -r \cos \theta \left(r - \frac{a^3}{2r^2} \right). \quad (12.38)$$

It is straightforward to find \mathbf{v} from this potential. We will not pursue the matter further.

12-6 Illumination; the uniform lighting of a plane

In this section we turn to a completely different physical problem—we want to illustrate the great variety of possibilities. This time we will do something that leads to the same kind of integral that we found in electrostatics. (If we have a mathematical problem which gives us a certain integral, then we know something about the properties of that integral if it is the same integral that we had to do for another problem.) We take our example from illumination engineering. Suppose there is a light source at the distance a above a plane surface. What is the illumination of the surface? That is, what is the radiant energy per unit time arriving at a unit area of the surface? (See Fig. 12-9.) We suppose that the source is spherically

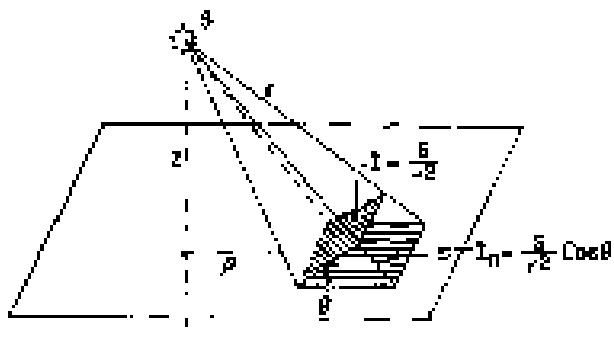


Fig. 12-9. The illumination I_n of a surface is the radiant energy per unit time arriving at a unit area of the surface.

symmetric, so that light is radiated equally in all directions. Then the amount of radiant energy which passes through a unit area at right angles to a light ray varies inversely as the square of the distance. It is evident that the intensity of the light in the direction normal to the flow is given by the same kind of formula as for the electric field from a point source. If the light rays meet the surface at an angle θ to the normal, then I , the energy arriving per unit area of the surface, is only $\cos \theta$ as great, because the same energy goes onto an area larger by $1/\cos \theta$. If we call the strength of our light source S , then I_n , the illumination of a surface, is

$$I_n = \frac{S}{r^2} \cos \theta, \quad (12.39)$$

where r is the unit vector from the source and n is the unit normal to the surface. The illumination I_n corresponds to the normal component of the electric field from a point charge of strength $4\pi r^2 S$. Knowing this, we see that for any distribution of light sources, we can find the answer by solving the corresponding electrostatic problem. We calculate the vertical component of electric field on the plane due to a distribution of charge in the same way as for that of the light sources.*

Consider the following example. We wish for some special experimental situation to arrange that the top surface of a table will have a very uniform illumination. We have available long tubular fluorescent lights which radiate uniformly along their lengths. We can illuminate the table by placing the fluorescent tubes in a regular array on the ceiling, which is at the height a above the table. What is the "ideal" spacing b from tube to tube (but we should use if we want the surface illumination to be uniform to, say, within one part in a thousand)? Answer: (1) Find the electric field from a grid of wires with the spacing b , each charged uniformly; (2) compute the vertical component of the electric field; (3) find out what b must be so that the ripples of the field are not more than one part in a thousand.

In Chapter 7 we saw that the electric field of a grid of charged wires could be represented as a sum of terms, each one of which gave a sinusoidal variation of the field with a period of b/n , where n is an integer. The amplitude of any one of these terms is given by Eq. (7.44):

$$E_1 = 4\pi e^{-2\pi n b/a}.$$

We need consider only $n = 1$, so long as we only want the field at points not too close to the grid. For a complete solution, we would still need to determine the coefficients A_n , which we have not yet done (although it is a straightforward calculation). Since we need only E_z , we can estimate that its magnitude is roughly the same as that of the average field. The exponential factor would then give us directly the relative amplitude of the variations. If we want this factor to be 10^{-3} , we find that b must be $0.91a$. If we make the spacing of the fluorescent tubes $3/4$

* Since we are talking about incoherent sources whose intensities always add linearly, the analogous electric charges will always have the same sign. Also, our analogy applies only to the light energy striking a the top of an opaque surface, so we must include in our integral only the sources which situate on the surface (and, naturally, not sources located below the surface).

of the distance to the ceiling, the exponential factor is then $1/4000$, and we have a safety factor of 4, so we are fairly sure that we will have the illumination constant to one part in a thousand. (An exact calculation shows that A_1 is really twice the average field, so the exact answer is $b = 0.82$.) It is somewhat surprising that for such a uniform illumination the allowed separation of the tubes comes out so large.

12-7 The "underlying unity" of nature

In this chapter, we wished to show that in learning electrodynamics you have learned at the same time how to handle many subjects in physics, and that by keeping this in mind, it is possible to learn almost all of physics in a limited number of years.

However, a question surely suggests itself at the end of such a discussion: Why are the equations from different phenomena so similar? We might say: "It is the underlying unity of nature." But what does that mean? What could such a statement mean? It could mean simply that the equations are similar for different phenomena; but then, of course, we have given no explanation. The "underlying unity" might mean that everything is made out of the same stuff, and therefore obeys the same equations. That sounds like a good explanation, but let us think. The electrostatic potential, the diffusion of neutrons, heat flow—are we really dealing with the same stuff? Can we really imagine that the electrostatic potential is physically identical to the temperature, or to the density of particles? Certainly ϕ is not exactly the same as the thermal energy of particles. The displacement of a membrane is certainly not like a temperature. Why, then, is there "an underlying unity"?

A closer look at the physics of the various subjects shows, in fact, that the equations are not really identical. The equation we found for neutron diffusion is only an approximation that is good when the distance over which we are looking is large compared with the mean free path. If we look more closely, we would see the individual neutrons running around. Certainly the motion of an individual neutron is a completely different thing from the smooth variation we get from solving the differential equation. The differential equation is an approximation, because we assume that the neutrons are smoothly distributed in space.

Is it possible that this is the clue? Thus the thing which is common to all the phenomena is the space, the framework into which the physics is put? As long as things are reasonably smooth in space, then the important things that will be involved will be the rates of change of quantities with position in space. That is why we always get an equation with a gradient. The derivatives must appear in the form of a gradient or a divergence; because the laws of physics are independent of direction, they must be expressible in vector form. The equations of electrodynamics are the simplest vector equations that one can get which involve only the spatial derivatives of quantities. Any other complex problem—or simplification of a complicated problem—must look like electrodynamics. What is common to all our problems is that they involve space and that we have *buried* what is actually a complicated phenomenon by a simple differential equation.

That leads us to another interesting question. Is the same statement perhaps also true for the electrostatic equations? Are they also correct only as a smooth-but-imitation of a really much more complicated microscopic world? Could it be that the real world consists of little X-ons which can be seen only at very tiny distances? And that in our measurements we are always observing on such a large scale that we can't see these little X-ons, and that is why we get the differential equations?

Our currently most complete theory of electrodynamics does indeed have its difficulties at very short distances. So it is possible, in principle, that these equations are smoothed-out versions of something. They appear to be correct at distances down to about 10^{-14} cm, but then they begin to lose weight. It is possible that there is some as yet undiscovered underlying "messiness," and that the details of an underlying complexity are hidden in the smooth-looking equations—as is so

in the "standard" discussion of neutrinos. But no one has yet formulated a successful theory that works that way.

Strangely enough, it turns out (for reasons that we do not at all understand) that the combination of relativity and quantum mechanics as we know them seems to forbid the execution of an operation that is fundamentally different from Eq. (12.4), and which does not at the same time lead to some kind of contradiction. Not simply a disagreement with experiment, but an *internal contradiction*. As, for example, the prediction that the sum of the probabilities of all possible occurrences is not equal to unity, or that energies may sometimes come out as complex numbers, or worse still, negativity. No one has yet made up a theory of electricity for which $\nabla^2\phi = -\rho/e_n$ is understood as a smoothed-out approximation to a mechanism underneath, and which does not lead ultimately to some kind of an absurdity. But, it must be added, it is also true that the assumption that $\nabla^2\phi = -\rho/e_n$ is valid for all distances, no matter how small, leads to absurdities of its own (the electrical energy of an electron is infinite)—absurdities from which we can yet escape.

Magnetostatics

13-1 The magnetic field

The force on an electric charge depends not only on where it is, but also on how fast it is moving. Every point in space is characterized by two vector quantities which determine the force on any charge. First, there is the electric force, which gives a force component independent of the motion of the charge. We describe it by the electric field, E . Second, there is an additional force component, called the magnetic force, which depends on the velocity of the charge. This magnetic force has a strange directional character: At any particular point in space, both the direction of the force and its magnitude depend on the direction of motion of the particle; at every instant the force is always at right angles to the velocity vector; also, at any particular point, the force is always at right angles to a fixed direction in space (see Fig. 13-1); and finally, the magnitude of the force is proportional to the component of the velocity at right angles to this unique direction. It is possible to describe all of this behavior by defining the magnetic field vector B , which specifies both the unique direction in space and the constant of proportionality with the velocity, and to write the magnetic force as $\mathbf{q}v \times \mathbf{B}$. The total electrodynamic force on a charge can, then, be written as

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \quad (13.1)$$

This is called the *Lorentz force*.

The magnetic force is easily demonstrated by bringing a bar magnet close to a cathode-ray tube. The deflection of the electron beam shows that the presence of the magnet results in forces on the electrons transverse to their direction of motion, as we described in Chapter 12 of Vol. I.

The unit of magnetic field B is evidently one newton-second per coulomb-meter. The same unit is also one volt-second per meter². It is also called one weber per square meter.

13-2 Electric current; the conservation of charge

We consider first how we can understand the magnetic forces on wires carrying electric currents. In order to do this, we define what is meant by the current density. Electric currents are electric or ionic charges in motion with a net drift or flow. We can represent the charge flow by a vector which gives the amount of charge passing per unit area and per unit time through a surface element at right angles to the flow (just as we did for the case of heat flow). We call this the current density and represent it by the vector j . It is directed along the motion of the charges. If we take a small area ΔS at a given place in the material, the amount of charge flowing across that area in a unit time is

$$j \cdot \mathbf{n} \Delta S, \quad (13.2)$$

where \mathbf{n} is the unit vector normal to ΔS .

The current density is related to the average flow velocity of the charges. Suppose that we have a distribution of charges whose average motion is a drift with the velocity v . As this distribution passes over a surface element ΔS , the charge Δq passing through the surface element in a time Δt is equal to the charge contained in a parallelepiped whose base is ΔS and whose height is $v \Delta t$, as shown in Fig. 13-2. The volume of the parallelepiped is the projection of ΔS at right angles to v times

13-1 The magnetic field

13-2 Electric current; the conservation of charge

13-3 The magnetic force on a current

13-4 The magnetic field of steady currents; Ampere's law

13-5 The magnetic field of a straight wire and of a solenoid; atomic currents

13-6 The relativity of magnetic and electric fields

13-7 The transformation of charges and charges

13-8 Superposition; the right-hand rule

Review: Chapter 13, Vol. I: The Special Theory of Relativity

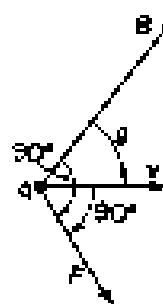


Fig. 13-1. The velocity-dependent component of the force on a moving charge is at right angles to v and to the direction of B . It is also proportional to the component of v at right angles to B , that is, to $v \sin \theta$.

Δt , which when multiplied by the charge density ρ will give Δq . Thus

$$\Delta q = \rho v \cdot n \Delta S \Delta t.$$

The charge per unit time is then $\rho v \cdot n \Delta S$, from which we get

$$j = \rho v \quad (13.3)$$

If the charge distribution consists of individual charges, say electrons, each with the charge q and moving with the mean velocity v , then the current density is

$$j = Nqv, \quad (13.4)$$

where N is the number of charges per unit volume.

The total charge passing per unit time through any surface S is called the electric current, I . It is equal to the integral of the normal component of the flow through all of the elements of the surface:

$$I = \int_S j \cdot n dS \quad (13.5)$$

(see Fig. 13-3).

The current I out of a closed surface S represents the rate at which charge leaves the volume V enclosed by S . One of the basic laws of physics is that electric charge is indestructible; it is never lost or created. Electric charges can move from place to place but never appear from nowhere. We say that charge is conserved. If there is a net current out of a closed surface, the amount of charge inside must decrease by the corresponding amount (Fig. 13-4). We can, therefore, write the law of the conservation of charge as

$$\int_{\text{closed surface}} j \cdot n dS = - \frac{d}{dt} (Q_{\text{inside}}). \quad (13.6)$$

The charge inside can be written as a volume integral of the charge density:

$$Q_{\text{inside}} = \int_V \rho dV. \quad (13.7)$$

If we apply (13.6) to a small volume ΔV , we know that the left-hand integral is $\nabla \cdot j \Delta V$. The charge inside is $\rho \Delta V$, so the conservation of charge can also be written as

$$\nabla \cdot j = - \frac{\partial \rho}{\partial t} \quad (13.8)$$

(Gauss' mathematics once again!).

13-3 The magnetic force on a current

Now we are ready to find the force on a current-carrying wire in a magnetic field. The current consists of charged particles moving with the velocity v along the wire. Each charge feels a transverse force

$$\mathbf{F} = qv \times \mathbf{B}$$

(Fig. 13-5a). If there are N such charges per unit volume, the number in a small volume ΔV of the wire is $N \Delta V$. The total magnetic force ΔF on the volume ΔV is the sum of the forces on the individual charges, that is,

$$\Delta F = (N \Delta V)(qv \times \mathbf{B}).$$

But Nqv is just j , so

$$\Delta F = j \times \mathbf{B} \Delta V \quad (13.9)$$

(Fig. 13-5b). The force per unit volume is $j \times \mathbf{B}$.

If the current is uniform across a wire whose cross-sectional area is A , we may take as the volume element a cylinder with the base area A and the length ΔL . Then

$$\Delta F = j \times B A \Delta L. \quad (13.10)$$

Now we can call jA the vector current I in the wire. (Its magnitude is the electric current in the wire, and its direction is along the wire.) Then

$$\Delta F = I \times B A \Delta L. \quad (13.11)$$

The force per unit length on a wire is $I \times B$.

This equation gives the important result that the magnetic force on a wire, due to the movement of charge in it, depends only on the total current, and not on the amount of charge carried by each particle—or even its sign! The magnetic force on a wire near a magnet is easily shown by observing its deflection when a current is turned on, as was described in Chapter 1 (see Fig. 1-6).

13-4 The magnetic field of steady currents; Ampere's law

We have seen that there is a force on a wire in the presence of a magnetic field, produced, say, by a magnet. From the principle that action equals reaction we might expect that there should be a force on the source of the magnetic field, i.e., on the magnet, when there is a current through the wire.* Thus we indeed find forces, as is seen by the deflection of a compass needle near a current-carrying wire. Now we know that magnets feel forces from other magnets, so that proves that when there is a current in a wire, the wire itself generates a magnetic field. Moving charges, then, produce a magnetic field. We would like now to try to discover the laws that determine how such magnetic fields are created. The question is: Given a current, what magnetic field does it make? The answer to this question was determined experimentally by three critical experiments and a brilliant abooretical argument given by Ampere. We will pass over this interesting historical development and simply say that a large number of experiments have demonstrated the validity of Maxwell's equations. We take them as our starting point. If we drop the terms involving time derivatives in these equations we get the equations of magnetostatics:

$$\nabla \cdot H = 0 \quad (13.12)$$

and

$$\nabla \times B = \frac{j}{\epsilon_0}. \quad (13.13)$$

These equations are valid only if all electric charge densities are constant and all currents are steady, so that the electric and magnetic fields are not changing with time—all of the fields are “static.”

We may remark that it is rather dangerous to think that there is such a thing as a static magnetic situation because there must be currents in order to get a magnetic field at all—and currents can come only from moving charges. “Magnetostatics” is, therefore, an approximation. It refers to a special kind of dynamic situation with large numbers of charges in motion, which we can approximate by a steady flow of charge. Only then can we speak of a current density j which does not change with time. The subject should more accurately be called the study of steady currents. Assuming that all fields are steady, we drop all terms in $\partial E/\partial t$ and $\partial B/\partial t$ from the complete Maxwell equations, Eqs. (2.11), and obtain the two equations (13.12) and (13.13) above. Also notice that since the divergence of the curl of any vector is necessarily zero, Eq. (13.13) requires that $\nabla \cdot j = 0$. This is true, by Eq. (13.8), only if $\partial j/\partial t$ is zero. But that must be so if E is not changing with time, so our assumptions are consistent.

* We will see later, however, that such assumptions are not generally correct for electromagnetic forces!

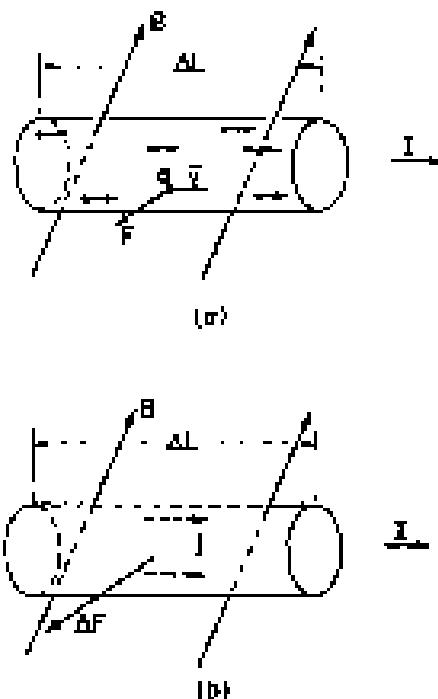


Fig. 13-5. The magnetic force on a current-carrying wire is the sum of the forces on the individual moving charges.

The requirement that $\nabla \cdot \mathbf{B} = 0$ means that we may only have charges which flow in paths that close back on themselves. Only they, for instance, flow in wires that form complete loops—called circuits. The circuits may, of course, contain generators or batteries that keep the charges flowing. But they may not include conductors which are charging or discharging. (We will, of course, extend the theory later to include dynamic fields, but we will first take the simpler case of steady currents.)

Now let us look at Eqs. (13.13) and (13.13) to see what they mean. The first says that the divergence of \mathbf{B} is zero. Comparing it to the analogous equation in electromagnetism, which says that $\nabla \cdot \mathbf{E} = \rho/\epsilon_0$, we can conclude that there is no magnetic analog of an electric charge. There are no magnetic charges from which lines of \mathbf{B} can emerge. If we think in terms of “lines” of the vector field \mathbf{B} , they can never start and they never stop. Then where do they come from? Magnetic fields “appear” in the presence of currents; they have a curl proportional to the current density. Wherever there are currents, there are lines of magnetic field linking loops around the currents. Since lines of \mathbf{B} do not begin or end, they will often close back on themselves, making closed loops. But there can also be complicated situations in which the lines are not simple closed loops. But whatever they do, they never diverge from points. No magnetic charges have ever been discovered, so $\nabla \cdot \mathbf{B} = 0$. This much is true not only for magnetostatics, it is always true even for dynamic fields.

The connection between the \mathbf{B} field and currents is contained in Eq. (13.13). Here we have a new kind of situation which is quite different from electromagnetism, where we had $\nabla \times \mathbf{E} = 0$. That equation meant that the line integral of \mathbf{E} around any closed path is zero:

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = 0.$$

We get this result from Stokes' theorem, which says that the integral around any closed path of any vector field \mathbf{A} is equal to the surface integral of the curl of \mathbf{A} projected at the end of the vector (taken over any surface which has the closed loop as its periphery). Applying the same theorem to the magnetic field vector and using the symbols shown in Fig. 13-6, we get

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \int_S (\nabla \times \mathbf{B}) \cdot \mathbf{n} dS. \quad (13.14)$$

Taking the curl of \mathbf{B} from Eq. (13.13), we have

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \frac{1}{\epsilon_0 c^2} \int_S j \cdot \mathbf{n} dS. \quad (13.15)$$

The integral over j , according to (13.5), is the total current I through the surface S . Since for steady currents the current through S is independent of the shape of S , so long as it is bounded by the curve C , one usually speaks of “the current through the ‘loop’.” We have, then, a general law: the circulation of \mathbf{B} around any closed curve is equal to the current I through the loop, divided by $\epsilon_0 c^2$:

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \frac{I_{\text{through } S}}{\epsilon_0 c^2}. \quad (13.16)$$

This law—called Ampere's law—plays the same role in magnetostatics that Gauss' law played in electromagnetism. Ampere's law alone does not determine \mathbf{B} from currents; we must, in general, also use $\nabla \cdot \mathbf{B} = 0$. But, as we will see in the next section, it can be used to find the field in special circumstances which have certain simple symmetries.

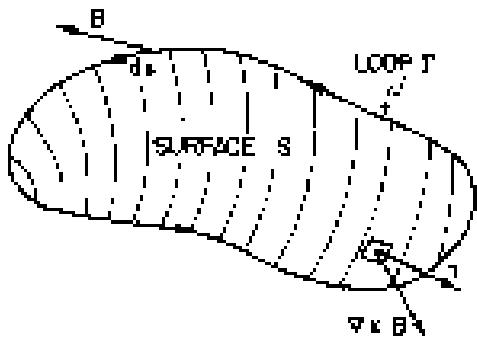


Fig. 13-4. The line integral of the tangential component of \mathbf{B} is equal to the surface integral of the normal component of $\nabla \times \mathbf{B}$.

13-5 The magnetic field of a straight wire and of a solenoid; summing currents

We can illustrate the use of Ampere's law by finding the magnetic field near a wire. We ask: What is the field outside a long, straight wire with a cylindrical cross section? We will assume something which may not be at all evident, but which is nevertheless true: that the field lines of B go around the wire in closed circles. If we make this assumption, then Ampere's law, Eq. (13.16), tells us how strong the field is. From the symmetry of the problem, B has the same magnitude at all points on a circle concentric with the wire (see Fig. 13-7). We can then do the line integral of B at quite easily; it is just the magnitude of B times the circumference. If r is the radius of the circle, then

$$\oint \mathbf{B} \cdot d\mathbf{s} = B \cdot 2\pi r.$$

The total current through the loop is merely the current I in the wire, so

$$B \cdot 2\pi r = \frac{I}{\mu_0 r^2},$$

or

$$B = \frac{1}{4\pi\mu_0 r^2} \frac{2I}{r}. \quad (13.17)$$

The strength of the magnetic field drops off inversely as r , the distance from the axis of the wire. We can, if we wish, write Eq. (13.17) in vector form. Remembering that \mathbf{B} is at right angles both to \mathbf{I} and to \mathbf{r} , we have

$$\mathbf{B} = \frac{1}{4\pi\mu_0 r^2} \frac{2I \times \mathbf{e}_r}{r}. \quad (13.18)$$

We have separated out the factor $1/4\pi\mu_0 r^2$, because it appears often. It is worth remembering that it is exactly 10^{-7} (in the mks system), since an equation like (13.17) is used to define the unit of current, the ampere. At one meter from a current of one ampere the magnetic field is 2×10^{-7} webers per square meter.

Since a current produces a magnetic field, it will exert a force on a nearby wire which is also carrying a current. In Chapter 1 we described a simple demonstration of the forces between two current-carrying wires. If the wires are parallel, each is at right angles to the B field of the other; the wires should then be pushed either toward or away from each other. When currents are in the same direction, the wires attract; when the currents are moving in opposite directions, the wires repel.

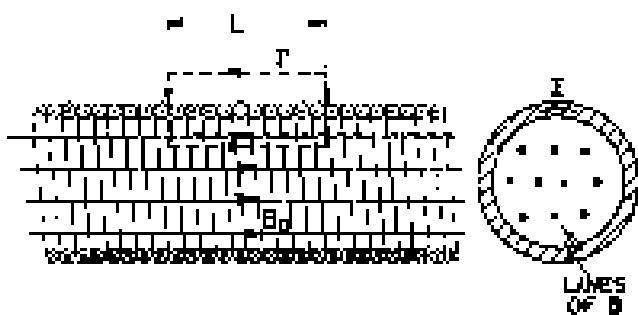


Fig. 13-8. The magnetic field of a long solenoid.

Let's take another example that can be analyzed by Ampere's law if we add some knowledge about the field. Suppose we have a long coil of wire wound in a tight spiral, as shown by the cross sections in Fig. 13-8. Such a coil is called a solenoid. We observe experimentally that when a solenoid is very long compared with its diameter, the field outside is very small compared with the field inside. Using just that fact, together with Ampere's law, we can find the size of the field inside.

Since the field inside must have zero divergence, its lines must go along parallel to the axis, as shown in Fig. 13-8. That being the case, we can use Ampere's law with the rectangular "curve" L shown in the figure. This loop goes the distance

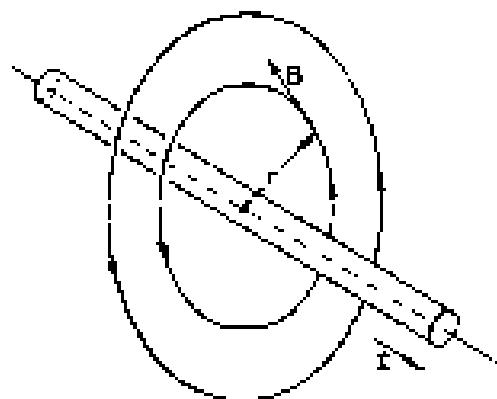


Fig. 13-7. The magnetic field outside of a long wire carrying the current I .

7. inside the solenoid, where the field is, say, B_0 , then goes at right angles to the field, and returns along the outside, where the field is negligible. The line integral of \mathbf{B} for this curve is just B_0L , and it must be $1/\epsilon_0c^2$ times the total current through C , which is NI if there are N turns of the solenoid in the length L . We have

$$B_0L = \frac{NI}{\epsilon_0c^2}.$$

Or, letting n be the number of turns per unit length of the solenoid (that is, $n = N/L$), we get

$$B_0 = \frac{nI}{\epsilon_0c^2}. \quad (13.19)$$

What happens to the lines of \mathbf{B} when they get to the end of the solenoid? Presumably, they spread out in some way and return to enter the solenoid at the other end, as sketched in Fig. 13-9. Such a field is just what is observed outside of a bar magnet. But what is a magnet anyway? Our equations say that \mathbf{B} comes from the presence of currents. Yet we know that ordinary bars of iron (no batteries or generators) also produce magnetic fields. You might expect that there should be some other terms on the right-hand side of (13.12) or (13.13) to represent "the density of magnetization" or some such quantity. But there is no such term. Our theory says that the magnetic effects of iron come from some internal currents which are already taken care of by the j term.

Matter is very complex when looked at from a fundamental point of view—but we saw when we tried to understand Electricity. In order not to interrupt our present discussion, we will wait until later to deal in detail with the interior mechanisms of magnetic materials like iron. You will have to accept, for the moment, that all magnetism is produced from currents, and that in a permanent magnet there are permanent induced currents. In the case of iron, these currents consist of electrons spinning around their own axes. Every electron has such a spin, which corresponds to a tiny circulating current. Of course, one electron doesn't produce much magnetic field, but in an ordinary piece of matter there are billions and billions of electrons. Normally their spin and point every which way, so that there is no net effect. The秘密 is that in a few low substances, like iron & some forms of the electrons spin with their axes in the same direction—for iron, two electrons from each atom take part in this cooperative motion. In a bar magnet there are large numbers of electrons all spinning in the same direction and, as we will see, their total effect is equivalent to a current circulating on the surface of the bar. (This is quite analogous to what we found for dielectrics—that a uniformly polarized dielectric is equivalent to a distribution of charges on its surface.) It is, therefore, no accident that a bar magnet is equivalent to a solenoid.

13-6 The relativity of magnetic and electric fields

When we said that the magnetic force on a charge was proportional to its velocity, you may have wondered: "What velocity? With respect to which reference frame?" It is, in fact, clear from the definition of \mathbf{B} given at the beginning of this chapter that what this vector is will depend on what we choose as a reference frame for specifying the velocity of charges. But we have said nothing about which is the proper frame for specifying the magnetic field.

It turns out that our inertial frame will do. We will also see that magnetism and electricity are not independent things—that they should always be taken together as one complete electromagnetic field. Although in the static case Maxwell's equations separate into two distinct pairs, one pair for electricity and one pair for magnetism, with no apparent connection between the two fields, nevertheless, in nature itself there is a very intimate relationship between them that arises from the principle of relativity. Historically, the principle of relativity was discovered after Maxwell's equations. It was, in fact, the study of electricity and magnetism which led ultimately to Einstein's discovery of his principle of relativity. But let's see

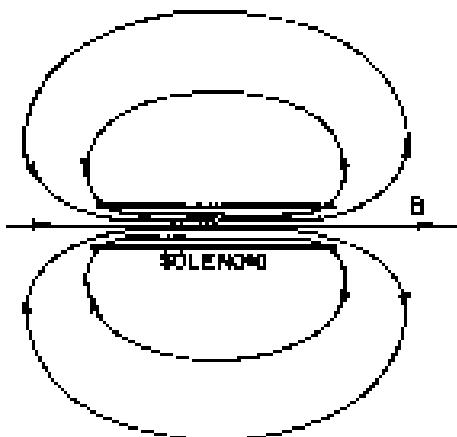


Fig. 13-9. The magnetic field outside of a solenoid.

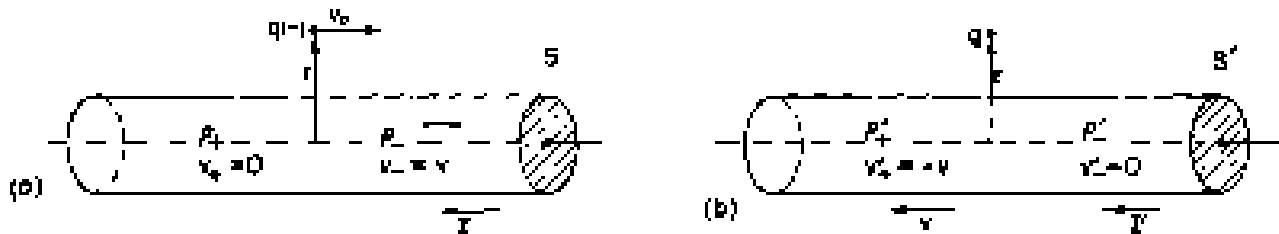


Fig. 13-10. The interaction of a current-carrying wire and a particle with the charge q as seen in two frames. In frame S [part (a)] the wire is at rest; in frame S' [part (b)], the charge is at rest.

what our knowledge of relativity would tell us about magnetic forces if we assume that the relativity principle is applicable—as it is—to electromagnetism.

Suppose we think about what happens when a negative charge moves with velocity v_0 parallel to a current-carrying wire, as in Fig. 13-10. We will try to understand what goes on in two reference frames: one fixed with respect to the wire, as in part (a) of the figure, and one fixed with respect to the particle, as in part (b). We will call the first frame S and the second S' .

In the S -frame, there is clearly a magnetic force on the particle. The force is directed toward the wire, so if the charge is moving freely we would see it curve in toward the wire. But in the S' -frame there will be no magnetic forces on the particle, because its velocity is zero. Does it, therefore, stay where it is? Would we see different things happening in the two systems? The principle of relativity would say that in S' we should also see the particle move closer to the wire. We must try to understand why that would happen.

We return to our atomic description of a wire carrying a current. In a normal conductor, like copper, the electric currents come from the motion of some of the negative electrons—called the conduction electrons—while the positive nuclear charges and the remainder of the electrons stay fixed in the body of the material. We let the density of the conduction electrons be ρ , and their velocity in S be v . The density of the charges at rest in S is ρ_0 , which must be equal to the negative of ρ , since we are considering an uncharged wire. There is thus no electric field outside the wire, and the force on the moving particle is just

$$\mathbf{F} = q\mathbf{v}_0 \times \mathbf{B}.$$

Using the result we found in Eq. (13.18) for the magnetic field at the distance r from the axis of a wire, we conclude that the force on the particle is directed toward the wire and has the magnitude

$$F = \frac{1}{4\pi\epsilon_0 r^2} \cdot \frac{2\mu_0 I^2}{r}.$$

Using Eqs. (13.1) and (13.5), the current I can be written as $\rho \cdot A$, where A is the area of a cross section of the wire. Then,

$$F = \frac{1}{4\pi\epsilon_0 r^2} \cdot \frac{2\mu_0 A \rho v_0}{r}. \quad (13.20)$$

We could continue to treat the general case of arbitrary velocities v_0 and v , but it will be just as good to look at the special case in which the velocity v_0 of the particle is the same as the velocity v of the conduction electrons. So we write $v_0 = v$, and Eq. (13.20) becomes

$$F = \frac{\sigma}{2\pi\epsilon_0} \frac{\mu_0 A \pi^2}{r^2}. \quad (13.21)$$

Now we turn our attention to what happens in S' , in which the particle is at rest and the wire is moving past (toward the left in the figure) with the speed v . The positive charges moving with the wire will make some magnetic field B' at the particle. But the particle is now at rest, so there is no magnetic force on it. If there is any force on the particle, it must come from an electric field. It must

be that the moving wire has produced an electric field. But it can do that only if it appears charged—it must be that a neutral wire with a current appears to be charged while set in motion.

We must look into this. We must try to compute the charge density in the wire in S' from what we know about it in S . One might, at first, think they are the same, but we know that lengths are changed between S and S' (see Chapter 15, Vol. I), so volumes will change also. Since the charge densities depend on the volume occupied by charges, the densities will change, too.

Before we can decide about the charge densities in S' , we must know what happens to the electric charge of a bunch of electrons when the charges are moving. We know that the apparent mass of a particle changes by $(\sqrt{1 - v^2/c^2})$. Does its charge do something similar? No. Charges are always the same, moving or not. Otherwise we would not always observe that the total charge is conserved.

Suppose that we take a block of material, say a conductor, which is initially uncharged. Now we heat it up. Because the electrons have a different mass than the protons, the velocities of the electrons and of the protons will change by different amounts. If the charge of a particle depended on the speed of the particle carrying it, in the heated block the charge of the electrons and protons would no longer balance. A block would become charged when heated. As we have seen earlier, a very small fractional change in the charge of all the electrons in a block would give rise to enormous electric fields. No such effect has ever been observed.

Also, we can point out that the mean speed of the electrons in reality depends on its chemical composition. If the charge on an electron changed with speed, the net charge in a piece of material would be changed in a chemical reaction. Again, a straightforward calculation shows that even a very small dependence of charge on speed would give enormous fields from the simplest chemical reactions. No such effect is observed, and we conclude that the electric charge of a single particle is independent of its state of motion.

So the charge q on a particle is an invariant scalar quantity, independent of the frame of reference. That means that in any frame the charge density of a distribution of electrons is just proportional to the number of electrons per unit volume. We need only worry about the fact that the volume can change because of the relativistic contraction of distances.

We now apply these ideas to our moving wire. If we take a length L_0 of the wire, in which there is a charge density ρ_0 of stationary charges, it will contain the total charge $Q = \rho_0 L_0 A_0$. If the same charges are observed in a different frame to be moving with velocity v , they will all be found in a piece of the material with the shorter length

$$L = L_0 \sqrt{1 - v^2/c^2}. \quad (13.22)$$

but with the same area A_0 (sizes dimensions transverse to the motion are unchanged). See Fig. 13-11.

If we call ρ the density of charges in the frame in which they are moving, the total charge Q will be $\rho L A_0$. This must also be equal to $\rho_0 L_0 A_0$ because charge is the same in any system so that $\rho L = \rho_0 L_0$, or, from (13.22),

$$\rho = \frac{\rho_0}{\sqrt{1 - v^2/c^2}}. \quad (13.23)$$

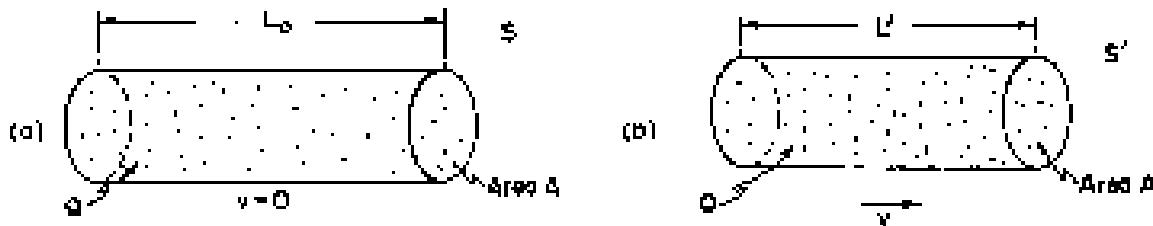


Fig. 13-11. If a distribution of charged particles at rest has the charge density ρ_0 , the same charges will have the density $\rho = \rho_0 / \sqrt{1 - v^2/c^2}$ when seen from a frame with the relative velocity v .

The charge density of a moving distribution of charges varies in the same way as the relativistic mass of a particle.

We now use this general result for the positive charge density ρ_+ of our wire. These charges are at rest in frame S . In S' , however, where the wire moves with the speed v , the positive charge density becomes

$$\rho'_+ = \frac{\rho_+}{\sqrt{1 - v^2/c^2}}. \quad (13.24)$$

The negative charges are at rest in S' . So they have their "rest density" ρ_- in this frame. In Eq. (13.23) $\rho_3 = \rho'_-$, because they have the density ρ'_- when the wire is at rest, i.e., in frame S , where the speed v of the negative charges is 0 . For the conduction electrons, we then have that

$$\rho'_- = \frac{\rho'_-}{\sqrt{1 - v^2/c^2}}. \quad (13.25)$$

or

$$\rho'_- = \rho_- \sqrt{1 - v^2/c^2}. \quad (13.26)$$

Now we can see why there are electric fields in S' —because in this frame the wire has the net charge density ρ' given by

$$\rho' = \rho'_+ + \rho'_-.$$

Using (13.24) and (13.26), we have

$$\rho' = \frac{\rho_+}{\sqrt{1 - v^2/c^2}} + \rho_- \sqrt{1 - v^2/c^2}.$$

Since the stationary wire is neutral, $\rho_+ = -\rho_-$, and we have

$$\rho' = \rho \frac{v^2/c^2}{\sqrt{1 - v^2/c^2}}. \quad (13.27)$$

Our moving wire is positively charged and will produce an electric field E' at the external stationary particle. We have already solved the electrostatic problem of a uniformly charged cylinder. The electric field at the distance r from the axis of the cylinder is

$$E' = \frac{\rho A}{3\pi r^2} = \frac{\rho_- v^2/c^2}{2\pi r^2 c \sqrt{1 - v^2/c^2}}. \quad (13.28)$$

The force on the negatively charged particle is toward the wire. We have, at least, a force in the same direction from the two points of view: the electric force in S' has the same direction as the magnetic force in S .

The magnitude of the force in S' is

$$F' = \frac{q}{2\pi\epsilon_0} \frac{\rho_- A}{r} \frac{v^2/c^2}{\sqrt{1 - v^2/c^2}}. \quad (13.29)$$

Comparing this result for F' with our result for F in Eq. (13.22), we see that the magnitudes of the forces are almost identical from the two points of view. In fact,

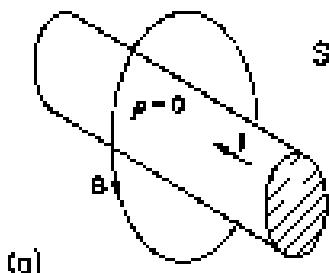
$$F' = \frac{F}{\sqrt{1 - v^2/c^2}}, \quad (13.30)$$

so for the small velocities we have been considering, the two forces are equal. We can say that for low velocities, at least, we understand that magnetism and electricity are just "two ways of looking at the same thing."

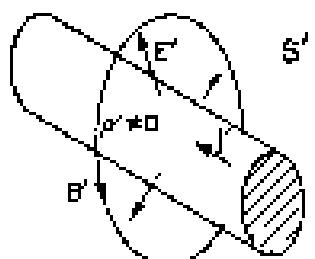
But things are even better than that. If we take into account the fact that forces also transform when we go from one system to the other, we find that the two ways of looking at what happens in motion give the same physical result for any velocity.

One way of seeing this is to ask a question like: What transverse momentum will the particle have after the force has acted for a little while? We know from Chapter 16 of Vol. I that the transverse momentum of a particle should be the same in both the S - and S' -frames. Calling the transverse coordinate p , we want to compare Δp_x and $\Delta p'_x$. Using the relativistically correct equation of motion, $F = dp/dt$, we expect that after the time Δt our particle will have a transverse momentum Δp_x in the S -system given by

$$\Delta p_x = F \Delta t. \quad (13.21)$$



(a)



(b)

Fig. 13-12. In frame S the charge density is zero and the current density is j . There is only a magnetic field. In S' , there is a charge density ρ' , and a different current density j' . The magnetic field B' is different and there is an electric field E' .

In the S' -system, the transverse momentum will be

$$\Delta p'_x = F' \Delta t', \quad (13.22)$$

We must, of course, compare Δp_x and $\Delta p'_x$ for corresponding time intervals Δt and $\Delta t'$. We have seen in Chapter 15 of Vol. I that these intervals referred to a moving particle appear to be longer than those in the rest system of the particle. Since our particle is initially at rest in S' , we expect, for small Δt , that

$$\Delta t = -\frac{\Delta t'}{\sqrt{1-v^2/c^2}}, \quad (13.23)$$

and everything comes out O.K. From (13.21) and (13.22),

$$\frac{\Delta p'_x}{\Delta p_x} = \frac{F' \Delta t'}{F \Delta t},$$

which is just $= 1$ if we combine (13.20) and (13.23).

We have found that we get the same physical result whether we analyze the motion of a particle moving along a wire in a coordinate system at rest with respect to the wire, or in a system at rest with respect to the particle. In the first instance, the force was purely "magnetic;" in the second, it was purely "electric." The two points of view are illustrated in Fig. 13-12 (although there is still a magnetic field B' in the second frame, it produces no forces on the stationary particle).

If we tried to use still another coordinate system, we would have found a different mixture of E and B fields. Electric and magnetic forces are part of one global phenomenon—the electromagnetic interactions of particles. The separation of this interaction into electric and magnetic parts depends very much on the reference frame chosen for the description. But a complete electromagnetic description is invariant; electricity and magnetism taken together are consistent with Lorentz's relativity.

Since electric and magnetic fields appear in different mixtures if we change our frame of reference, we must be careful about how we look at the fields E and B . For instance, if we think of "lines" of E or B , we might get too much reality to them. The lines may disappear if we try to observe them from a different coordinate system. For example, in system S' there are electric field lines, which we do not find "moving past us with velocity v " in system S . In system S there are no electric field lines at all! Therefore it makes no sense to say something like: When I move a magnet, it takes its field with it, so the lines of B are also moved. There is no way to make sense, in general, out of the idea of "the speed of a moving field line." The fields are our way of describing what goes on at a given in space. In particular, E and B tell us about the forces that will act on a moving particle. The question "What is the force on a charge from a moving magnetic field?" doesn't mean anything precise. The force is given by the values of E and B at the charge, and the formula (13.1) is not to be altered if the source of E or B is moving (j is the values of E and B that will be altered by the motion). Our mathematical description deals only with the fields as a function of x , y , z , and t with respect to some inertial frame.

We will take be speaking of "a wave of electric and magnetic fields travelling through space," as, for instance, a light wave. But that is like speaking of a wave travelling on a string. We don't then mean that some part of the string is moving

In the direction of the wave, we mean that the displacement of the string appears first at one point and later at another. Similarly, in an electromagnetic wave, the wave travels, but the magnitude of the fields changes. So in the future when we—or someone else—speak of a “moving” field, you should think of it as just a handy, short way of describing a changing field in some circumstances.

13-7 The transformation of currents and charges

You may have worried about the simplification we made above when we took the same velocity v for the particle and for the conduction electrons in the wire. We could go back and carry through the analysis again for two different velocities, but it is easier to simply notice that charge and current density are the components of a four-vector (see Chapter 17, Vol. I).

We have seen that if ρ_0 is the density of the charges in their rest frame, then in a frame in which they have the velocity v , the density is

$$\rho = \frac{\rho_0}{\sqrt{1 - v^2/c^2}}.$$

In that frame their current density is

$$j = \rho v = \frac{\rho_0 v}{\sqrt{1 - v^2/c^2}}. \quad (13.34)$$

Now we know that the energy U and momentum p of a particle moving with velocity v are given by

$$U = \frac{mc^2}{\sqrt{1 - v^2/c^2}}, \quad p = \frac{mc_0 v}{\sqrt{1 - v^2/c^2}},$$

where m_0 is its rest mass. We also know that U and p form a relativistic four-vector. Since ρ and j depend on the velocity v exactly as do U and p , we can conclude that ρ and j are also the components of a relativistic four-vector. This property is the key to a general analysis of the field of a wire moving with any velocity, which we would need if we want to do the problem again with the velocity v , of the particle different from the velocity of the conduction electrons.

If we wish to transform ρ and j to a coordinate system moving with a velocity w in the x -direction, we know that they transform just like x and (x, y, z) , so that we have (see Chapter 15, Vol. I)

$$\begin{aligned} x' &= \frac{x - wt}{\sqrt{1 - w^2/c^2}}, & j_x' &= \frac{j_x - wp}{\sqrt{1 - w^2/c^2}}, \\ y' &= y, & j_y' &= j_y, \\ z' &= z, & j_z' &= j_z, \\ t' &= \frac{t - px/c^2}{\sqrt{1 - w^2/c^2}}, & p' &= \frac{p - wt/c^2}{\sqrt{1 - w^2/c^2}}. \end{aligned} \quad (13.35)$$

With these equations we can relate charges and currents in one frame to those in another. Taking the charges and currents in either frame, we can solve the electromagnetic problem in that frame by using our Maxwell equations. The result we obtain for the *motion of particles* will be the same no matter which frame we choose. We will return at a later time to the relativistic transformations of the electromagnetic fields.

13-8 Superposition; the right-hand rule

We will conclude this chapter by making two further points regarding the subject of magnetostatics. First, our basic equations for the magnetic field.

$$\nabla \cdot \mathbf{B} = 0, \quad \nabla \times \mathbf{B} = \mu_0 i \epsilon_0 \mathbf{n}$$

are linear in \mathbf{B} and \mathbf{J} . This means that the principle of superposition also applies to magnetic fields. The field produced by two different steady currents is the sum of the individual fields from each current acting alone. Our second remark concerns the right-handed rules which we have encountered (such as the right-hand rule for the magnetic field produced by a current). We have also observed that the magnetization of an iron magnet is to be understood from the spin of the electrons in the material. The direction of the magnetic field of a spinning electron is related to its spin axis by the same right-hand rule. Because \mathbf{B} is determined by a "handed" rule involving either a cross product or a curl—it is called an axial vector. (Vectors whose direction in space does not depend on a reference to a right or left hand are called polar vectors. Displacement, velocity, force, and \mathbf{E} , for example, are polar vectors.)

Physically observable quantities in electromagnetism are not, however, right- (or left-) handed. Electromagnetic interactions are symmetrical under reflection (see Chapter 32, Vol. II). Whenever magnetic forces between two sets of currents are computed, the result is invariant with respect to a change in the hand convention. Our equations lead, independently of the right-hand convention, to the end result that parallel currents attract, and that currents in opposite directions repel. (Try working out the force using "left-hand rules.") An attracted or repulsive is a polar vector. This happens because in describing any excepted interaction, we use the right-hand rule twice—one to find \mathbf{B} from currents, again to find the force this \mathbf{B} produces on a second current. Using the right-hand rule twice is the same as using the left-hand rule twice. If we were to change our convention to a left-handed system all our \mathbf{B} fields would be reversed, but all forces—etc., what is perhaps more relevant, the observed accelerations of objects—would be unchanged.

Although physicists have usually found to their surprise that *not* the laws of nature are not always invariant for mirror reflections, the laws of electromagnetism do have such a basic symmetry.

The Magnetic Field in Various Situations

14-1 The vector potential

In this chapter we continue our discussion of magnetic fields associated with steady currents—the subject of magnetostatics. The magnetic field is related to electric currents by our basic equations:

$$\nabla \cdot \mathbf{B} = 0, \quad (14.1)$$

$$\epsilon_0^2 \nabla \times \mathbf{B} = \frac{\mathbf{j}}{\epsilon_0}. \quad (14.2)$$

We want now to solve these equations, mathematically in a general way, that is, without requiring any special symmetry or intuitive guessing. In electrostatics, we found that there was a straightforward procedure for finding the field when the positions of all electric charges are known: One simply works out the scalar potential ϕ by taking an integral over the charges—as in Eq. (4.23). Then if one wants the electric field, it is obtained from the derivatives of ϕ . We will now show that there is a corresponding procedure for finding the magnetic field \mathbf{B} if we know the current density \mathbf{j} of all moving charges.

In electrostatics we saw that (because the curl of \mathbf{A} was always zero) it was possible to represent \mathbf{E} as the gradient of a scalar field ϕ . Now the curl of \mathbf{B} is not always zero, so it is not possible, in general, to represent it as a gradient. However, the divergence of \mathbf{B} is always zero, and this means that we can always represent \mathbf{B} as the curl of another vector field. For, as we saw in Section 2.8, the divergence of a curl is always zero. Thus we can always relate \mathbf{B} to a field we will call \mathbf{A} by

$$\mathbf{B} = \nabla \times \mathbf{A}. \quad (14.3)$$

Or, by writing out the components,

$$\begin{aligned} B_x &= (\nabla \times \mathbf{A})_x = \frac{\partial A_y}{\partial z} - \frac{\partial A_z}{\partial y}, \\ B_y &= (\nabla \times \mathbf{A})_y = \frac{\partial A_z}{\partial x} - \frac{\partial A_x}{\partial z}, \\ B_z &= (\nabla \times \mathbf{A})_z = \frac{\partial A_x}{\partial y} - \frac{\partial A_y}{\partial x}. \end{aligned} \quad (14.4)$$

Writing $\mathbf{B} = \nabla \times \mathbf{A}$ guarantees that Eq. (14.2) is satisfied, since, necessarily,

$$\nabla \cdot \mathbf{B} = \nabla \cdot (\nabla \times \mathbf{A}) = 0.$$

The field \mathbf{A} is called the *vector potential*.

You will remember that the scalar potential ϕ was not completely specified by its definition. If we have found ϕ for some problem, we can always find another potential ϕ' that is equally good by adding a constant:

$$\phi' = \phi + C.$$

The new potential ϕ' gives the same electric fields, since the gradient ∇C is zero; ϕ' and ϕ represent the same physics.

Similarly, we can take different vector potentials \mathbf{A} , which give the same magnetic fields. Again, because \mathbf{B} is obtained from \mathbf{A} by differentiation, adding a

14-1 The vector potential

14-2 The vector potential of known currents

14-3 A straight wire

14-4 A long solenoid

14-5 The field of a small loop; the magnetic dipole

14-6 The vector potential of a circuit

14-7 The law of Biot and Savart

constant to A cannot change anything physical. But there is even more latitude for A . We can add to A any field which is the gradient of some scalar field, without changing the physics. We can show this as follows. Suppose we have on A that gives correctly the magnetic field B (as above) and suppose, and ask in what circumstances some other new vector potential A' will give the same field B if substituted into (14.3). Then A and A' must have the same curl:

$$\mathbf{B} = \nabla \times A' = \nabla \times A.$$

Therefore

$$\nabla \times A' - \nabla \times A = \nabla \times (A' - A) = 0.$$

But if the curl of a vector is zero it must be the gradient of some scalar field, say ψ , so $A' - A = \nabla\psi$. That means that if A is a satisfactory vector potential for a problem then, for any ψ at all,

$$A' = A + \nabla\psi \quad (14.5)$$

will be an equally satisfactory vector potential, leading to the same field B .

It is usually convenient to take some of the "inside" out of A by arbitrarily placing some other condition on it (in much the same way that we fixed it convenient after to choose to make the potential ϕ zero at large distances). We can, for instance, restrict A by choosing arbitrarily what the divergence of A must be. We can always do that without affecting B . This is because although A' and A have the same curl, and give the same B , they do not need to have the same divergence. In fact, $\nabla \cdot A' = \nabla \cdot A + \nabla^2\psi$, and by a suitable choice of ψ we can make $\nabla \cdot A'$ anything we wish.

What should we choose for $\nabla \cdot A$? The choice should be made to get the greatest mathematical convenience and will depend on the problem we are doing. For magnetostatics, we will make the simple choice

$$\nabla \cdot A = 0. \quad (14.6)$$

(Later, when we take up electrodynamics, we will change our choice.) Our complete definition* of A is then, for the moment, $\nabla \times A = B$ and $\nabla \cdot A = 0$.

To get some experience with the vector potential, let's look first at what it is for a uniform magnetic field B_0 . Taking our z -axis to the direction of B_0 , we must have

$$\begin{aligned} B_x &= \frac{\partial A_z}{\partial y} = \frac{\partial A_z}{\partial z} = 0, \\ B_y &= \frac{\partial A_z}{\partial x} = \frac{\partial A_z}{\partial y} = 0, \\ B_z &= \frac{\partial A_x}{\partial x} = \frac{\partial A_y}{\partial y} = B_0. \end{aligned} \quad (14.7)$$

By inspection, we see that one possible solution of these equations, is

$$A_x = zB_0, \quad A_y = 0, \quad A_z = 0.$$

Or we could equally well take

$$A_x = -yB_0, \quad A_y = 0, \quad A_z = 0.$$

Still another solution is a linear combination of the two:

$$A_x = -yzB_0, \quad A_y = yzB_0, \quad A_z = 0. \quad (14.8)$$

* Our definition still does not uniquely determine A . For a unique specification we would also have to say something about how the field A behaves on some boundary, or at large distances. It is sometimes convenient, for example, to choose a field which goes to zero at large distances.

It is clear that for any particular field B , the vector potential A is not unique; there are many possibilities.

The third solution, Eq. (14.8), has some interesting properties. Since the x -component is proportional to $-y$ and the y -component is proportional to $+x$, A must be at right angles to the vector from the z -axis, which we will call r' (the "prime" is to remind us that it is *not* the vector displacement from the origin). Also, the magnitude of A is proportional to $\sqrt{x^2 + y^2}$ and, hence, to r' . So A can be simply written (for our uniform field) as

$$A = \frac{1}{2}B \times r'. \quad (14.9)$$

The vector potential A has the magnitude $Br'/2$ and rotates about the z -axis as shown in Fig. 14-1. If, for example, the B field is the axial field inside a solenoid, then the vector potential circulates in the same sense as do the currents of the solenoid.

The vector potential for a uniform field can be obtained in another way. The circulation of A on any closed loop C can be related to the surface integral of $\nabla \times A$ by Stokes' theorem, Eq. (3.38):

$$\oint_C A \cdot dr = \int_{\text{Surface } S} (\nabla \times A) \cdot n \, da. \quad (14.10)$$

But the integral on the right is equal to the flux of B through the loop, so

$$\oint_C A \cdot dr = \int_{\text{Surface } S} B \cdot n \, da. \quad (14.11)$$

So the circulation of A around any loop is equal to the flux of B through the loop. If we take a circular loop, of radius r' in a plane perpendicular to a uniform field B , the flux is just

$$\pi r'^2 B.$$

If we choose our origin on an axis of symmetry, so that we can take A as circumferential and a function only of r' , the circulation will be

$$\oint A \cdot dr = 2\pi r' A = \pi r'^2 B.$$

We get, as before,

$$A = \frac{Br'}{2}.$$

In the example we have just given, we have calculated the vector potential from the magnetic field, which is opposite to what one normally does. In complicated problems it is usually easier to solve for the vector potential, and then determine the magnetic field from it. We will now show how this can be done.

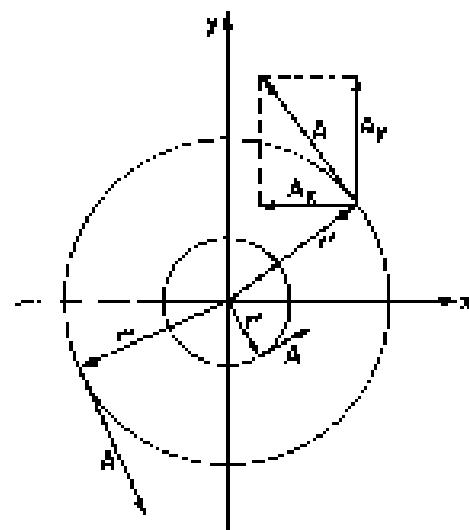


Fig. 14-1. A uniform magnetic field B in the z -direction corresponds to a vector potential A that rotates about the z -axis, with the magnitude $A = Br'/2$ (r' is the displacement from the z -axis).

14-2 The vector potential of known currents

Since B is determined by currents, so also is A . We want now to find A in terms of the currents. We start with our basic equation (14.3):

$$\epsilon^2 \nabla \times B = \frac{j}{\epsilon_0},$$

which becomes, of course, then

$$\epsilon^2 \nabla \times (\nabla \times A) = \frac{j}{\epsilon_0}. \quad (14.12)$$

This equation is for magnetostatics what the equation

$$\nabla \cdot \nabla \phi = -\frac{\rho}{\epsilon_0} \quad (14.13)$$

was for electrostatics.

Our equation (14.12) for the vector potential looks even more like that for ϕ if we rewrite $\nabla \times (\nabla \times A)$ using the vector identity Eq. (2.58):

$$\nabla \times (\nabla \times A) = \nabla(\nabla \cdot A) - \nabla^2 A. \quad (14.14)$$

Since we have chosen to make $\nabla \cdot A = 0$ (and now you see why), Eq. (14.12) becomes

$$\nabla^2 A = - \frac{j}{\epsilon_0 c^2}. \quad (14.15)$$

This vector equation consists, of course, of three equations:

$$\nabla^2 A_x = - \frac{j_x}{\epsilon_0 c^2}, \quad \nabla^2 A_y = - \frac{j_y}{\epsilon_0 c^2}, \quad \nabla^2 A_z = - \frac{j_z}{\epsilon_0 c^2}. \quad (14.16)$$

And each of these equations is *mathematically identical* to

$$\nabla^2 \phi = - \frac{\rho}{\epsilon_0}. \quad (14.17)$$

All we have learned about solving for potentials when ρ is known can be used for solving for each component of A when j is known!

We have seen in Chapter 4 that a general solution for the electrostatic equation (14.17) is

$$\phi(r) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(2) dV_2}{r_{12}},$$

So we know immediately that a general solution for A_x is

$$A_x(1) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{j_x(2) dV_2}{r_{12}}, \quad (14.18)$$

and similarly for A_y and A_z . (Figure 14-2 will remind you of our conventions for r_{12} and dV_2 .) We can combine the three solutions in the vector form

$$A(1) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{\mathbf{j}(2) dV_2}{r_{12}}. \quad (14.19)$$

(You can verify if you wish, by direct differentiation of components, that this integral for A satisfies $\nabla \cdot A = 0$ so long as $\nabla \cdot j = 0$, which, as we saw, must happen for steady currents.)

We have, then, a general method for finding the magnetic field of steady currents. The principle is: the x -component ∇^2 vector potential arising from a current density j is the same as the electric potential ϕ that would be produced by a charge density ρ equal to j_x/c^2 , and similarly for the y - and z -components. (This principle works only with components in fixed directions. The "radial" component of A does not come in the same way from the "radial" component of j , for example.) So from the vector current density j , we can find A using Eq. (14.19), that is, we find each component of A by solving three imaginary electrostatic problems for the charge distributions $\rho_x = j_x/c^2$, $\rho_y = j_y/c^2$, and $\rho_z = j_z/c^2$. Then we get B by taking various derivatives of A to obtain $\nabla \times A$. It's a little more complicated than electrostatics, but the same idea. We will now illustrate the theory by solving for the vector potential in a few special cases.

14-3 A straight wire

For our first example, we will again find the field of a straight wire --which we solved in the last chapter by using Eq. (14.2) and some arguments of symmetry. We take a long straight wire of radius a , carrying the steady current I . (Unlike the charge on a conductor in the electrostatic case, a steady current in a wire is uniformly distributed throughout the cross section of the wire. If we choose our

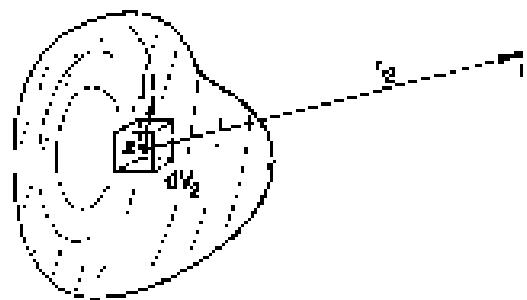


Fig. 14-2. The vector potential A at point 1 is given by an integral over the current elements $j dV$ at all points 2,

coordinates as shown in Fig. 14-3, the current density vector \vec{j} has only a z -component. Its magnitude is

$$j_z = \frac{I}{\pi a^2} z \quad (14.20)$$

inside the wire, and zero outside.

Since j_x and j_y are both zero, we have immediately

$$A_x = 0, \quad A_y = 0.$$

To get A_z we can use our solution for the electrostatic potential ϕ of a wire with a uniform charge density $\rho = j_z/c^2$. For points outside an infinite charged cylinder, the electrostatic potential is

$$\phi = -\frac{\lambda}{2\pi\epsilon_0 c^2} \ln r'$$

where $r' = \sqrt{x^2 + y^2}$ and λ is the charge per unit length, $\pi a^2 \rho$. So A_z must be

$$A_z = -\frac{\pi a^2 j_z}{2\pi\epsilon_0 c^2} \ln r'$$

For points outside a long wire carrying a nonuniform current. Since $\pi a^2 j_z = I$, we can also write

$$A_z = -\frac{I}{2\pi\epsilon_0 c^2} \ln r'. \quad (14.21)$$

Now we can find B from (14.4). There are only two of the six derivatives that are not zero. We get

$$B_x = -\frac{I}{2\pi\epsilon_0 c^2} \frac{\partial}{\partial y} \ln r' = -\frac{I}{2\pi\epsilon_0 c^2} \frac{y}{r'^3}, \quad (14.22)$$

$$B_y = \frac{I}{2\pi\epsilon_0 c^2} \frac{\partial}{\partial x} \ln r' = \frac{I}{2\pi\epsilon_0 c^2} \frac{x}{r'^3}, \quad (14.23)$$

$$B_z = 0.$$

We get the same result as before: B is the same around the wire, and has the magnitude

$$B = \frac{1}{4\pi\epsilon_0 c^2} \frac{2I}{r'}. \quad (14.24)$$

14-4 A long solenoid

Next, we consider again the infinitely long solenoid with a cylindrical differential current of dI per unit length. (We imagine this to be a series of short segments of length dz , carrying the current I , and we neglect the slight pitch of the winding.)

Just as we have defined a "surface charge density" σ , we define here a "surface current density" J equal to the current per unit length on the surface of the solenoid (which is, of course, just the average j times the thickness of the thin winding). The magnitude of J is, here, dz . This surface current (see Fig. 14-4) has the components

$$J_x = -J \sin \phi, \quad J_y = J \cos \phi, \quad J_z = 0.$$

Now we could find A for such a current distribution.

Then, we wish to find A_z for points outside the solenoid. The result is the same as the electrostatic potential outside a cylinder with a similar charge

$$\sigma = \sigma_0 \sin \phi,$$

with $\sigma_0 = J/c^2$. We have not solved such a charge distribution, but we have done something similar. This charge distribution is equivalent to two small cylinders of charge, one positive and one negative, with a slight relative displacement of their

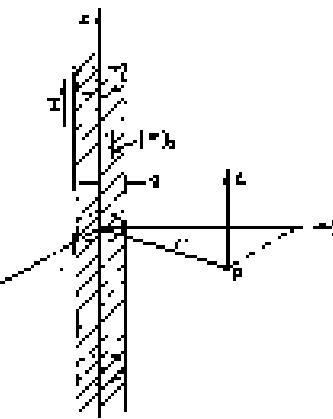


Fig. 14-3. A long cylindrical wire along the z -axis with a uniform current density j .

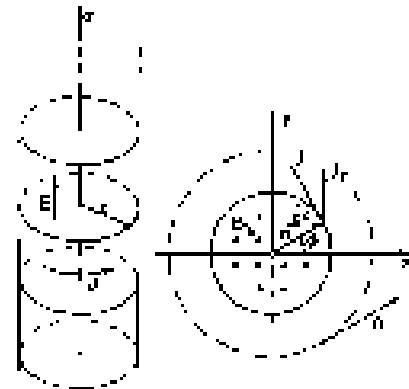


Fig. 14-4. A long solenoid with a surface current density J .

axes in the y -direction. The potential of such a pair of cylinders is proportional to the derivative with respect to y of the potential of a single uniformly charged cylinder. We could work out the constant of proportionality, but let's not worry about it for the moment.

The potential of a cylinder of charge is proportional to $\ln r'$; the potential of the pair is then

$$\psi \propto \frac{\partial \ln r'}{\partial y} = \frac{y}{r'^2}$$

So we know that

$$A_x = -K \frac{y}{r'^2}, \quad (14.25)$$

where K is some constant. Following the same argument, we would find

$$A_y = K \frac{z}{r'^2}. \quad (14.26)$$

Although we said before that there was no magnetic field outside a solenoid, we find now that there is an A -field which circulates around the z -axis, as in Fig. 14-4. The question is: Is it real zero?

Clearly, B_x and B_y are zero, and

$$\begin{aligned} B_z &= \frac{\partial}{\partial z} \left(K \frac{x}{r'^2} \right) - \frac{\partial}{\partial y} \left(-K \frac{y}{r'^2} \right) \\ &= K \left(\frac{1}{r'^4} - \frac{2z^2}{r'^4} + \frac{1}{r'^2} - \frac{2y^2}{r'^4} \right) = 0. \end{aligned}$$

So the magnetic field outside a very long solenoid is indeed zero, even though the vector potential is not.

We can check our result against something else we know: The circulation of the vector potential around the solenoid should be equal to the flux of \mathbf{B} inside the coil (Eq. 14-11). The circulation is $A \cdot 2\pi a$ or, since $A = K/r'$, the circulation is $2\pi K$. Notice that A is independent of r' . That is just as it should be if there is no \mathbf{B} outside, because the flux is just the magnitude of \mathbf{B} inside the solenoid, times πa^2 . It is the same for all circles of radius $r' > a$. We have found in the last chapter that the field inside is $4\pi/\sigma r'^2$, so we can determine the constant K :

$$2\pi K = \pi a^2 \frac{4\pi}{\sigma r'^2},$$

or

$$K = \frac{4\pi^2}{3\sigma a^2}.$$

So the vector potential *outside* has the magnitude

$$A = \frac{4\pi^2}{3\sigma a^2} \frac{1}{r'}, \quad (14.27)$$

and is always perpendicular to the vector \mathbf{r}' .

We have been thinking of a solenoidal coil of wire, but we would get the same fields if we rotated a long cylinder with an electrostatic charge on its surface. If we have a thin cylindrical shell of radius a with a surface charge σ , rotating the cylinder makes a surface current $J = \sigma v$, where $v = \omega a$ is the velocity of the surface charge. There will then be a magnetic field $B = \sigma v a / c r'^2$ inside the cylinder.

Now we can raise an interesting question. Suppose we put a short piece of wire W perpendicular to the axis of the cylinder, extending from the axis out to the surface and fastened to the cylinder so that it rotates with it, as in Fig. 14-5. This wire is moving in a magnetic field, so the $v \times \mathbf{B}$ force will cause the ends of the wire to be charged (they will charge up until the E -field from the charges just balances the $v \times \mathbf{B}$ force). If the cylinder has a positive charge, the end of the wire at the axis will have a negative charge. By measuring the charge on the end of the

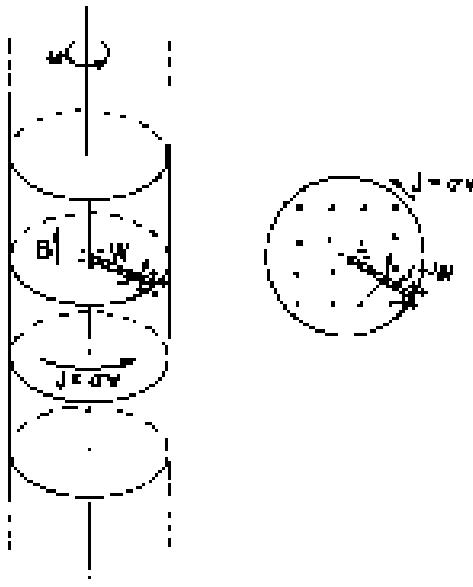


Fig. 14-5. A rotating charged cylinder produces a magnetic field inside. A short radial wire rotating with the cylinder has charges induced on its ends.

wire, we could measure the speed of rotation of the system. We would have an "angular-velocity meter"!

But are you wondering: "What if I put myself in the frame of reference of the rotating cylinder? Then there is just a charged cylinder at rest, and I know that the electrostatic equations say there will be no electric fields inside, so there will be no force pushing charges to the center. So something must be wrong." But there is nothing wrong. There is no "relativity of rotation." A rotating system is not an inertial frame, and the laws of physics are different. We must be sure to use equations of electromagnetism only with respect to inertial coordinate systems.

It would be nice if we could measure the absolute rotation of the earth with such a charged cylinder, but unfortunately the effect is much too small to observe even with the most delicate instruments now available.

14-5 The field of a small loop; the magnetic dipole

Let's use the vector-potential method to find the magnetic field of a small loop of current. As usual, by "small" we mean simply that we are interested in the fields only at distances large compared with the size of the loop. It will turn out that any small loop is a "magnetic dipole." That is, it produces a magnetic field like the electric field from an electric dipole.

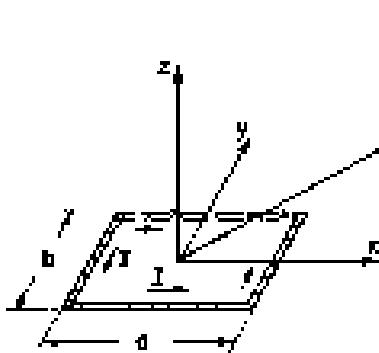


Fig. 14-6. A rectangular loop of wire with side lengths a and b , carrying a current I . What is the magnetic field at P ? ($R \gg a$, or b .)

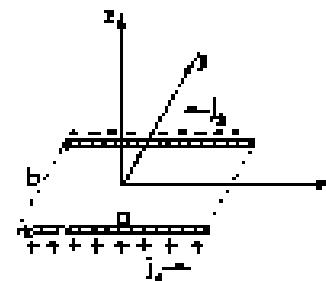


Fig. 14-7. The distribution of J_x in the current loop of Fig. 14-6.

We take first a rectangular loop, and choose our coordinates as shown in Fig. 14-6. There is no current in the z -direction, so A_z is zero. There are currents in the x -direction on the two sides of length a . In each leg, the current density (the current) is uniform. So the situation for A_x is just like the electrostatic potential from two charged rods (see Fig. 14-2). Since the rods have opposite charges, their electric potential at large distances would be just the dipole potential (Section 6-5). At the point P in Fig. 14-6, the potential would be

$$\phi = \frac{L}{4\pi\epsilon_0} \frac{\rho}{R^2}, \quad (14.28)$$

where ρ is the dipole moment of the charge distribution. The dipole moment, in this case, is the total charge on one rod times the separation between them:

$$\rho = qab. \quad (14.29)$$

The dipole moment points in the negative y -direction, so the cosine of the angle between R and ρ is $-y/R$ (where y is the coordinate of P). So we have

$$\phi = -\frac{1}{4\pi\epsilon_0} \frac{\lambda ab}{R^3} \frac{y}{R}.$$

We get A_x simply by replacing λ by I/c^2 :

$$A_x = -\frac{Iab}{4\pi\epsilon_0 c^2} \frac{y}{R^3}. \quad (14.30)$$

By the same reasoning,

$$A_x = \frac{Iab}{4\pi\epsilon_0 c^2} \frac{x}{R^3}. \quad (14.31)$$

Again, A_x , proportional to x and A_z , is proportional to $-y$, so the vector potential (at large distances) goes in circles around the z -axis, circulating in the same sense as I in the loop, as shown in Fig. 14.8.

The strength of A is proportional to Iab , which is the current times the area of the loop. This product is called the *magnetic dipole moment* (or, often, just "magnetic moment") of the loop. We represent it by μ :

$$\mu = Iab. \quad (14.32)$$

The vector potential of a small plane loop of any shape (circle, triangle, etc.) is also given by Eqs. (14.30) and (14.31) provided we replace Iab by

$$\mu = i \cdot (\text{area of loop}). \quad (14.33)$$

We leave the proof of this to you.

We can put our equation in vector form if we define the direction of the vector μ to be the normal to the plane of the loop, with a positive sense given by the right-hand rule (Fig. 14.8). Then we can write

$$A = \frac{1}{4\pi\epsilon_0 c^2} \frac{\mu \times R}{R^3} = \frac{1}{4\pi\epsilon_0 c^2} \frac{\mu \times e_R}{R^3}. \quad (14.34)$$

We have still to find B . Using (14.33) and (14.34), together with (14.4), we get

$$B_x = -\frac{\partial}{\partial z} \frac{\mu}{4\pi\epsilon_0 c^2} \frac{z}{R^3} = -\dots \frac{3xz}{R^5} \quad (14.35)$$

(where by \dots we mean $\mu/(2\pi\epsilon_0 c^2)$).

$$\begin{aligned} B_y &= \frac{\partial}{\partial x} \left(\dots \frac{y}{R^3} \right) = \dots \frac{3yz}{R^5}, \\ B_z &= \frac{\partial}{\partial y} \left(\dots \frac{y}{R^3} \right) - \frac{\partial}{\partial y} \left(\dots \frac{y}{R^3} \right) \\ &= \dots \left(\frac{1}{R^3} - \frac{3x^2}{R^5} \right). \end{aligned} \quad (14.36)$$

The components of the B -field behave exactly like those of the E -field for a dipole oriented along the z -axis. (See Eqs. (6.14) and (6.15); also Fig. 6-5.) That's why we call the loop a magnetic dipole. The word "dipole" is slightly misleading when applied to a magnetic field because there are no magnetic "poles" that correspond to electric charges. The magnetic "dipole field" is not produced by two "charges," but by an elementary current loop.

It is curious, though, that starting with completely different laws, $\nabla \cdot E = \rho/\epsilon_0$ and $\nabla \times B = \mu_0 J$, we can end up with the same kind of a field. Why should that be? It is because the dipole fields appear only when we are far away from all charges or currents. So through most of the relevant space the equations for E and B are identical: both have zero divergence and zero curl. So they give the same solutions. However, the sources whose configuration we summarize by the dipole moments are physically quite different—in one case, it's a circulating current; in the other, a pair of charges, one above and one below the plane of the loop for the corresponding field.

14-6 The vector potential of a closed loop

We are often interested in the magnetic fields produced by circuits of wire in which the diameter of the wire is very small compared with the dimensions of the whole system. In such cases, we can simplify the equations for the magnetic field.

For a thin wire we can write out volume elements as

$$dV = S ds,$$

where S is the cross-sectional area of the wire and ds is the element of distance along the wire. In fact, since the vector ds is in the same direction as j , as shown in Fig. 14-9 (and we can assume that j is constant across any given cross section), we can write a vector equation:

$$j dV = jS ds \quad (14.57)$$

But jS is just what we call the current I in a wire, so our integral for the vector potential (14.29) becomes

$$A(1) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{I ds}{r_{12}} \quad (14.58)$$

(see Fig. 14-10). (We assume that j is the same throughout the circuit. If there are several branches with different currents, we should, of course, use the appropriate I for each branch.)

Again, we can find the fields from (14.38) either by integrating directly or by solving the corresponding electrostatic problems.

14-7 The Law of Biot and Savart

In studying electrostatics we found that the electric field of a known charge distribution could be obtained directly with the integral (Eq. 4-16):

$$\mathbf{E}(1) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(Q) \mathbf{e}_{12} dV_2}{r_{12}^2},$$

As we have seen, it is usually much easier to evaluate this integral—there are really three integrals, one for each component—but to do the integral for the potential and take its gradient.

There is a similar integral which relates the magnetic field to the currents. We already have an integral for A , Eq. (14.58), we can get an integral for B by taking the curl of both sides:

$$\mathbf{B}(1) = \nabla \times A(1) = \nabla \times \left[\frac{1}{4\pi\epsilon_0 c^2} \int \frac{j(2) dV_2}{r_{12}} \right]. \quad (14.59)$$

Now we must be careful! The curl operator means taking the derivatives of $A(1)$, that is, it operates only on the coordinates (x_1, y_1, z_1) . We can move the $\nabla \times$ operator inside the integral sign if we remember that it operates only on variables with the subscript 1, which of course appears only in

$$r_{12} = [(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2]^{1/2}. \quad (14.60)$$

We have, for the x -component of B ,

$$\begin{aligned} B_x &= \frac{\partial A_x}{\partial y_1} - \frac{\partial A_y}{\partial z_1} \\ &= \frac{1}{4\pi\epsilon_0 c^2} \int \left[j_x \frac{\partial}{\partial y_1} \left(\frac{1}{r_{12}} \right) - j_y \frac{\partial}{\partial z_1} \left(\frac{1}{r_{12}} \right) \right] dV_2, \quad (14.61) \\ &= \frac{1}{4\pi\epsilon_0 c^2} \int \left[j_x \frac{y_1}{r_{12}^2} - j_y \frac{z_1}{r_{12}^2} \right] dV_2. \end{aligned}$$

The quantity in brackets is just the x -component of

$$\frac{\mathbf{j} \times \mathbf{r}_{12}}{r_{12}^2} \pm \frac{\mathbf{j} \times \mathbf{r}_{12}}{r_{12}^2}.$$

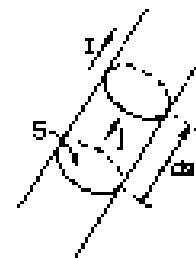


Fig. 14-9. For a fine wire $\int dV$ is the same as $\int ds$.

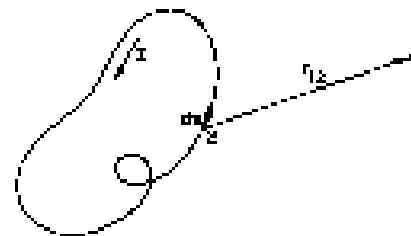


Fig. 14-10. The magnetic field of a wire can be obtained from an integral around the circuit.

Corresponding results will be found for the other components, so we have

$$\mathbf{B}(1) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{\mathbf{J}(2) \times \mathbf{r}_{12}}{r_{12}^3} dV_2. \quad (14.42)$$

The integral gives \mathbf{B} directly in terms of the known currents. The geometry involved is the same as that shown in Fig. 14-2.

If the currents exist on y in circles of small wires we can, as in the last section, immediately do the integral across the wire, replacing dV by Ids , where ds is an element of length of the wire. Then, using the symbol in Fig. 14-10,

$$\mathbf{B}(1) = -\frac{1}{4\pi\epsilon_0 c^2} \int \frac{I \mathbf{e}_{12} \times ds}{r_{12}^3}. \quad (14.43)$$

(The minus sign appears because we have reversed the order of the cross product.) This equation for \mathbf{B} is called the Biot-Savart law, after its discoverers. It gives a formula for obtaining directly the magnetic field produced by wires carrying currents.

You may wonder: "What is the advantage of the vector potential if we can find \mathbf{B} directly with a vector integral?" After all, A also involves three integrals!" Because of the cross product, the integrals for \mathbf{B} are usually more complicated, as is evident from Eq. (14.41). Also, since the integrals for A are like those of electrostatics, we may already know them. Finally, we will see that in more advanced theoretical matters (in relativity, in advanced foundations of the laws of mechanics, like the principle of least action to be discussed later, and in quantum mechanics) the vector potential plays an important role.

The Vector Potential

15-1 The forces on a current loop; energy of a dipole

In the last chapter we studied the magnetic field produced by a small rectangular current loop. We found that it is a dipole field, with the dipole moment given by

$$\mu = IA, \quad (15.1)$$

where I is the current and A is the area of the loop. The direction of the moment is normal to the plane of the loop, so we can also write

$$\mu = IA\hat{n},$$

where \hat{n} is the unit normal to the area A .

A current loop—in magnetic dipole—not only produces magnetic fields, but will also experience forces when placed in the magnetic field of other currents. We will look first at the forces on a rectangular loop in a uniform magnetic field. Let the x -axis be along the direction of the field and the plane of the loop be placed through the y -axis, making the angle θ with the xy -plane as in Fig. 15-1. Then the magnetic moment of the loop—which is normal to its plane—will make the angle θ with the magnetic field.

Since the currents are opposite on opposite sides of the loop, the forces are also opposite, so there is no net force on the loop (when the field is uniform). Because of forces on the two sides marked 1 and 2 in the figure, however, there is a torque which tends to rotate the loop about the y -axis. The magnitude of these forces F_1 and F_2 is

$$F_1 = F_2 \approx IBa.$$

Their moment arm is

$$a \sin \theta,$$

so the torque is

$$\tau = IabB \sin \theta,$$

or, since IBa is the magnetic moment of the loop,

$$\tau = \mu B \sin \theta.$$

The torque can be written in vector notation:

$$\tau = \mu \times B. \quad (15.2)$$

Although we have only shown that the torque is given by Eq. (15.2) in one rather special case, the result is right for a small loop of any shape, as we will see. You will remember that we found the same kind of relation for the torque on an electric dipole:

$$\tau = p \times E.$$

We now ask about the mechanical energy of our current loop. Since there is a torque, the energy evidently depends on the orientation. The principle of virtual work says that the torque is the rate of change of energy with angle, so we can write

$$dU = -\tau d\theta.$$

15-2 The forces on a current loop; energy of a dipole

15-3 Mechanical and electrical energies

15-4 B times A

15-5 The vector potential and quantum mechanics

15-6 What is true for statics is false for dynamics

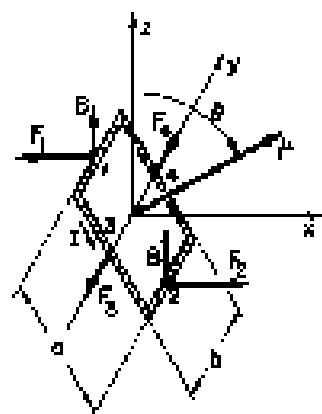


Fig. 15-1. A rectangular loop carrying the current I sits in a uniform field B (in the x -direction). The torque on the loop is $\tau = \mu \times B$, where the magnetic moment $\mu = Ib\hat{n}$.

Setting $\tau = -\mu B \sin \theta$, and integrating, we can write for the energy

$$U = -\mu B \cos \theta + \text{a constant.} \quad (15.3)$$

(The sign is negative because the torque tries to line up the moment with the field; the energy is lowest when μ and B are parallel.)

For reasons which we will discuss later, this energy is not the total energy of a current loop. (We have, for one thing, not taken into account the energy required to maintain the current in the loop.) We will, therefore, call this energy U_{loop} , to remind us that it is only part of the energy. Also, since we are leaving out some of the energy anyway, we can set the constant of integration equal to zero in Eq. (15.3). So we rewrite the equation:

$$U_{\text{loop}} = -\mu \cdot B \quad (15.4)$$

Again, this corresponds to our result for an electric dipole:

$$U = -\mu \cdot E. \quad (15.5)$$

Now the electrostatic energy U in Eq. (15.5) is the *true* energy, but U_{loop} in (15.4) is not the real energy. It can, however, be used in computing forces, by the principle of virtual work, supposing that the current in the loop—or at least μ —is kept constant.

We can show for our rectangular loop that U_{loop} also corresponds to the mechanical work done in bringing the loop into the field. The total force on the loop is zero only in a uniform field; in a non-uniform field there are net forces on a current loop. In putting the loop into a region with a field, we must have gone through places where the field was not uniform, and so work was done. To make the calculation simple, we shall imagine that the loop is brought into the field with its moment pointing along the field. (It can be rotated to its final position after it is in place.)

Imagine that we want to move the loop in the x -direction—toward a region of stronger field—and that the loop is oriented as shown in Fig. 15-2. We start somewhere where the field is zero and integrate the force times the distance as we bring the loop into the field.

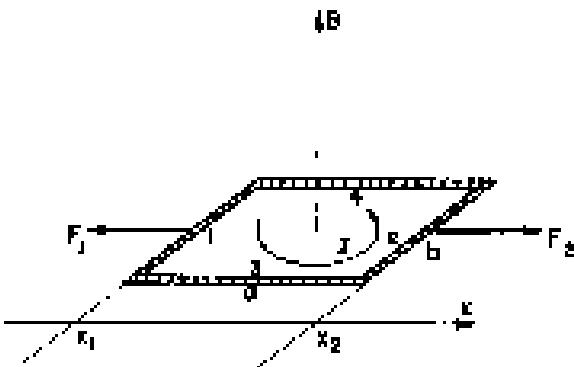


Fig. 15-2. A loop is carried along the x -direction through the field B_x at right angles to x .

First, let's compute the work done on each side separately and then take the sum (rather than adding the forces before integrating). The forces on sides 3 and 4 are at right angles to the direction of motion, so no work is done on them. The force on side 2 is $R B(x)$ in the x -direction, and to get the work done against the magnetic forces we must integrate this from some x where the field is zero, say at $x = -\infty$, to x_2 , its present position:

$$W_2 = - \int_{-\infty}^{x_2} F_2 dx = -ib \int_{-\infty}^{x_2} B(x) dx. \quad (15.6)$$

Similarly, the work done against the forces on side 1 is

$$W_1 = - \int_{-\infty}^{x_1} F_1 dx = ib \int_{-\infty}^{x_1} B(x) dx. \quad (15.7)$$

To find each integral, we need to know how $B(x)$ depends on x . We notice that side 1 follows along right behind side 2, so that its integral includes most of the work done on side 2. In fact, the sum of (15.6) and (15.7) is just

$$W = -\mu \int_{x_1}^{x_2} B(x) dx. \quad (15.8)$$

But if we are in a region where B is nearly the same on both sides 1 and 2, we can write the integral as

$$\int_{x_1}^{x_2} B(x) dx = (x_2 - x_1)B = aB,$$

where B is the field at the center of the loop. The total mechanical energy we have just is

$$U_{\text{mech}} = W = -\mu a B = -\mu B. \quad (15.9)$$

The result agrees with the energy we took for Eq. (15.4).

We would, of course, have gotten the same result if we had added the forces on the loop before integrating to find the work. If we let B_1 be the field at side 1 and B_2 be the field at side 2, then the total force in the \hat{x} -direction is

$$F_x = \mu(B_2 - B_1).$$

If the loop is “vertical,” that is, if B_2 and B_1 are not too different, we can write

$$B_2 \approx B_1 + \frac{\partial B}{\partial x} \Delta x = B_1 + \frac{\partial B}{\partial x} a.$$

So the force is

$$F_x = \mu a \frac{\partial B}{\partial x}. \quad (15.10)$$

The total work done on the loop by external forces is

$$-\int_{x_1}^{x_2} F_x dx = -\mu a \int \frac{\partial B}{\partial x} dx = -\mu a B,$$

which is again just $-\mu B$. Only now we see why it is that the force on a small current loop is proportional to the derivative of the magnetic field, as we would expect from

$$F_x \Delta x = -4C_{\text{max}} = -4(-\mu \cdot B). \quad (15.11)$$

Our result, then, is that even though $C_{\text{max}} = -\mu \cdot B$ may not include all the energy of a system—it is a fake kind of energy—it can still be used with the principle of virtual work to find the forces on steady current loops.

15-2 Mechanical and electrical energies

We want now to show why the energy U_{mech} discussed in the previous section is not the correct energy associated with steady currents—that it does not keep track of the total energy in the world. We have, indeed, emphasized that it can be used like the energy, for calculating forces from the principle of virtual work; we noted that the current in the loop (and all other currents) do not change. Let's see why not this works.

Imagine that the loop in Fig. 15-2 is moving in the $-\hat{x}$ -direction and take the z -axis in the direction of \mathbf{B} . The conduction electrons on side 2 will experience a force along the wire, in the y -direction. But because of their flow as an electric current, there is a component of their motion in the same direction as the force. Each electron is, therefore, doing work done $v_y t$ at the rate $F_y v_y$, where v_y is the component of the electron velocity along the wire. We will call this work done on the electrons’ electrical work. Now it turns out that if the loop is moving in a uniform field, the total electrical work is zero, since positive work is done on some parts of the loop and an equal amount of negative work is done on other parts.

But this is not true if the circuit is moving in a non-uniform field—then there will be a net amount of work done on the electrons. In general, this work would tend to change the flow of the electrons, but if the current is being held constant, energy must be absorbed or delivered by the battery or other source that is keeping the current steady. This energy was not included when we computed $\langle U_{\text{elect}} \rangle$ in Eq. (15.9), because our computations included only the mechanical forces on the body of the wire.

You may be thinking: But the force on the electrons depends on how fast the wire is moved; perhaps if the wire is moved slowly enough this electrical energy can be neglected. It is true that the rate at which the electrical energy is delivered is proportional to the speed of the wire, but the total energy delivered is proportional also to the time that this rate gets on. So the total electrical energy is proportional to the velocity times the time, which is just the distance moved. For a given distance moved in a field the same amount of electrical work is done.

Let's consider a segment of wire of unit length carrying the current I and moving in a direction perpendicular to itself and to a magnetic field B with the speed v_{wire} . Because of the current the electrons will have a drift velocity v_{drift} along the wire. The component of the magnetic force on each electron in the direction of the drift is $qv_{\text{drift}}B$. So the rate at which electrical work is being done is $P_{\text{elect}} = (qv_{\text{drift}}B)v_{\text{wire}}$. If there are N conduction electrons in the unit length of the wire, the total rate at which electrical work is being done is

$$\frac{dU_{\text{elect}}}{dt} = Nqv_{\text{drift}}Bv_{\text{wire}}$$

But $Nqv_{\text{drift}} = I$, the current in the wire, so

$$\frac{dU_{\text{elect}}}{dt} = Iv_{\text{wire}}B.$$

Now since the current is held constant, the forces on the conduction electrons do not cause them to accelerate; the electrical energy is not going into the electrons but from the source that is keeping the current constant.

But notice that the force on the wire is IB , so IBv_{wire} is also the rate of mechanical work done on the wire, $dU_{\text{mech}}/dt = IBv_{\text{wire}}$. We conclude that the mechanical work done on the wire is just equal to the electrical work done by the current source, so the energy of the loop is a *constant*!

This is not a coincidence, but a consequence of the law we already know. The total force on each charge in the wire is

$$F = q(E + v \times B).$$

The rate at which work is done is

$$v \cdot F = q[v \cdot E + v \cdot (v \times B)]. \quad (15.12)$$

If there are no electric fields we have only the second term, which is always zero. We shall see later that changing magnetic fields produce electric fields, so our reasoning applies only to moving wires in steady magnetic fields.

How is it then that the principle of virtual work gives the right answer? Because we still have not taken into account the total energy of the world. We have not included the energy of the currents that are producing the magnetic field we start out with.

Suppose we imagine a complete system such as that drawn in Fig. 15-3(a), in which we are moving our loop with the current I_1 into the magnetic field B_1 produced by the current I_2 in a coil. Now the current I_1 in the loop will also be producing some magnetic field B_2 at the coil. If the loop is moving, the field B_2 will be changing. As we shall see in the next chapter, a changing magnetic field generates an E -field; and this E -field will do work on the charges in the coil. This energy must also be included in our balance sheet of the total energy.

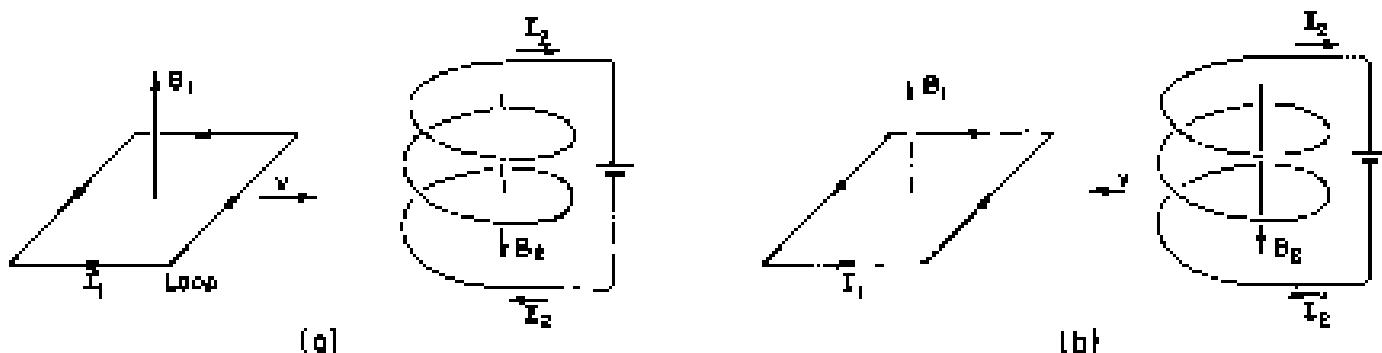


Fig. 15-3. Finding the energy of a small loop in a magnetic field.

We could wait until the next chapter to find out about this new energy term, but we can also see what it will be if we use the principle of relativity in the following way. When we are moving the loop toward the stationary coil we know that its electrical energy is just equal and opposite to the mechanical work done. So

$$U_{\text{mech}} + U_{\text{electr(loop)}} = 0.$$

Suppose now we look at what is happening from a different point of view, in which the loop is at rest, and the coil is moved toward it. The coil is then moving into the field produced by the loop. The same arguments would give that

$$U_{\text{mech}} + U_{\text{electr(coil)}} = 0.$$

The mechanical energy is the same in the two cases because it comes from the force between the two circuits.

The sum of the two equations gives

$$2U_{\text{mech}} + U_{\text{electr(loop)}} + U_{\text{electr(coil)}} = 0.$$

The total energy of the whole system is, of course, the sum of the two electrical energies plus the mechanical energy taken only once. So we have

$$U_{\text{total}} = U_{\text{electr(loop)}} - U_{\text{electr(coil)}} + U_{\text{mech}} = -U_{\text{mech}}. \quad (15.13)$$

The total energy of the world is really the negative of U_{mech} . If we want the true energy of a magnetic dipole, for example, we should write

$$U_{\text{total}} = -g \cdot B.$$

It is only if we make the condition that all currents are constant that we can use only a part of the energy, U_{mech} (which is always the negative of the true energy), to find the mechanical forces. In a more general problem, we must be careful to include all energies.

We saw an analogous situation in electrodynamics. We showed⁴ that the energy of a capacitor is equal to $Q^2/2C$. When we use the principle of virtual work to find the force between the plates of the capacitor, the change in energy is equal to $Q^2/2$ times the change in $1/C$. That is,

$$\Delta E = \frac{Q^2}{2} \Delta \left(\frac{1}{C} \right) = -\frac{Q^2}{2} \frac{\Delta C}{C^2}. \quad (15.14)$$

Now suppose that we were to calculate the work done in moving two conductors subject to the different condition that the voltage between them is held constant. Then we can get the right answers for forces from the principle of virtual work if we do something artificial. Since $(Q = CV)$, the real energy is $\frac{1}{2}CV^2$. But if we define an artificial energy equal to $-\frac{1}{2}CV^2$, then the principle of virtual work can be used to get forces by setting the change in the artificial energy equal to the

mechanical work, provided that we insist that the voltage V be held constant. Then

$$\delta U_{\text{elect}} = \delta \left(-\frac{CV^2}{2} \right) = -\frac{V^2}{2} \delta C, \quad (15.15)$$

which is the same as Eq. (15.14). We get the correct result even though we are neglecting the work done by the electrical system to keep the voltage constant. Again, this electrical energy is just twice as big as the mechanical energy and of the opposite sign.

Thus if we calculate artificially, disregarding the fact that the source of the potential has to do work to maintain the voltage constant, we get the right answer. It is exactly analogous to the situation in magnetostatics.

15-3 The energy of steady currents

We can now use our knowledge that $U_{\text{elect}} = -P_{\text{loss}}$ to find the true energy of steady currents in magnetic fields. We can begin with the true energy of a small current loop. Calling U_{loop} just U , we write

$$U = \mu \cdot B. \quad (15.16)$$

Although we calculated this energy for a plane rectangular loop, the same result holds for a small plane loop of any shape.

We can find the energy of a circuit of any shape by imagining that it is made up of small current loops. Say we have a wire in the shape of the loop Γ of Fig. 15-4. We pull in this curve with the surface S , and on the surface mark out a large number of small loops, each of which can be considered plane. If we set the current I (constant) around each of the little loops, the net result will be the same as a current stream I , since the currents will cancel on all lines intersecting Γ . Physically, the system of little currents is indistinguishable from the original circuit. The energy will also be the same, and so is just the sum of the energies of the little loops.

If the area of each little loop is Δa , its energy is $I^2 B_n \Delta a$, where B_n is the component normal to Δa . The total energy is

$$U = \sum I B_n \Delta a.$$

Going to the limit of infinitesimal loops, the sum becomes an integral, and

$$U = \int B_n da = \int B \cdot n da, \quad (15.17)$$

where n is the unit normal to da .

If we set $B = \nabla \times A$, we can connect the surface integral to a line integral, using Stokes' theorem,

$$\int_S (\nabla \times A) \cdot n da = \oint_C A \cdot ds, \quad (15.18)$$

where ds is the line element along Γ . So we have the energy for a circuit of any shape:

$$U = \int_{\text{circuit}} A \cdot ds. \quad (15.19)$$

In this expression A refers, of course, to the vector potential due to those currents (other than the I in the wire) which produce the field B at the wire.

Now any distribution of steady currents can be imagined to be made up of filaments that run parallel to the lines of current flow. For each pair of such circuits, the energy is given by (15.19), where the integral is taken around one circuit, using the vector potential A from the other circuit. For the total energy we want the sum of all such pairs. If instead of keeping track of the pairs, we take the complete sum over all the filaments, we would be enumerating the energy twice (we saw a similar effect in electrodynamics), so the total energy can be written

$$U = \frac{1}{2} \int B \cdot A dV. \quad (15.20)$$

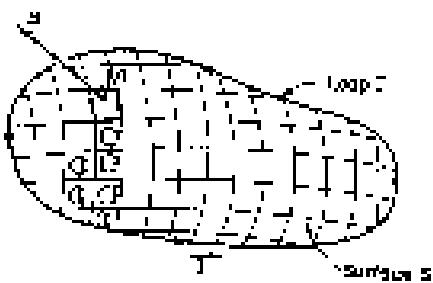


Fig. 15-4. The energy of a large loop in a magnetic field can be considered as the sum of energies of smaller loops.

This formula corresponds to the result we found for the electrostatic energy:

$$U = \frac{1}{2} \int \rho \phi dV. \quad (15.21)$$

So we may if we wish think of A as a kind of potential energy for currents in magnetostatics. Unfortunately, this idea is not too useful because it is true only for static fields. In fact, neither of the equations (15.20) and (15.21) gives the correct energy when the fields change with time.

15-4 \mathbf{A} versus \mathbf{A}

In this section we would like to discuss the following questions: Is the vector potential merely a device which is useful in making calculations—as the scalar potential is useful in electrostatics—or is the vector potential a “real” field? Isn’t the magnetic field the “real” field, because it is responsible for the force on a moving particle? First we should say that the phrase “a real field” is not very meaningful. For one thing, you probably don’t feel that the magnetic field is very “real” anyway, because even the whole idea of a field is a rather abstract thing. You cannot put out your hand and feel the magnetic field. Furthermore, the value of the magnetic field is not very definite; by choosing a suitable moving coordinate system, for instance, you can make a magnetic field at a given point disappear.

What we mean here by a “real” field is this: a real field is a mathematical function we use for avoiding the idea of action at a distance. If we have a charged particle at the position P , it is affected by other charges located at some distance from P . Our way to describe the interaction is to say that the other charges make some “condition”—whatever it may be—in the environment at P . If we know that condition, which we describe by giving the electric and magnetic fields, then we can determine completely the behavior of the particle—with no further reference to how those conditions come about.

In other words, if those other charges were altered in some way, but the conditions at P that are described by the electric and magnetic field at P remain the same, then the motion of the charge will also be the same. A “real” field is then a set of numbers we specify in such a way that what happens at P depends only on the numbers at that point. We do not need to know any more about what’s going on at other places. It is in this sense that we will discuss whether the vector potential is a “real” field.

You may be wondering about the fact that the vector potential is not unique: that it can be changed by adding the gradient of any scalar with no change at all in the forces on particles. That has got, however, nothing to do with the question of reality in the sense that we are talking about. For instance, the magnetic field is in a sense altered by a relativity change (as are also E and A). But we are not worried about what happens if the field was changed in this way. That doesn’t really make any difference; that has nothing to do with the question of whether the vector potential is a proper “real” field for describing magnetic effects, or whether it is just a useful mathematical tool.

We should also make some remarks on the usefulness of the vector potential A . We have seen that it can be used in a formal procedure for calculating the magnetic fields of known currents, just as ϕ can be used to find electric fields. In electrostatics we saw that ϕ was given by the scalar integral

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(2)}{r_{12}} dV_2. \quad (15.22)$$

From this ϕ , we get the three components of E by three differential operations. This procedure is usually easier to handle than evaluating the three integrals in the vector formula

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(2)\mathbf{e}_{12}}{r_{12}^2} dV_2. \quad (15.23)$$

First, there are three integrals; and second, each integral is in general somewhat more difficult.

The advantages are much less clear for magnetostatics. The integral for A is already a vector integral:

$$A(\mathbf{r}) = \frac{1}{4\pi\mu_0 c^2} \int \frac{\mathbf{j}(\mathbf{r}') dV_s}{r_{12}}, \quad (15.24)$$

which is, of course, three integrals. Also, when we take the curl of A to get \mathbf{B} , we have six derivatives to do and combine by pairs. It is not immediately obvious whether in most problems this procedure is really any easier than computing \mathbf{B} directly from

$$\mathbf{B}(\mathbf{r}) = \frac{1}{4\pi\mu_0 c^2} \int \frac{\mathbf{j}(\mathbf{r}') \times \mathbf{r}_{12}}{r_{12}^3} dV_s. \quad (15.25)$$

Using the vector potential is often more difficult for simple problems for the following reason. Suppose we are interested only in the magnetic field \mathbf{B} at one point, and that the problem has some axis symmetry—say we want the field at a point on the axis of a ring of current. Because of the symmetry, we can easily get \mathbf{B} by doing the integral of Eq. (15.25). If, however, we hope to find A first, we would have to compute \mathbf{B} from derivatives of A , so we must know what A is at all points in the neighborhood of the point of interest. At most c^2 of these points are off the axis of symmetry, so the integral for A gets complicated. In the ring problem, for example, we would need to use elliptic integrals. In such problems, A is clearly not very useful. It is true that in many complex problems it is easier to work with A , but it would be hard to argue that this use of technique would justify making you learn about one more vector field.

We have introduced A because it does have an important physical significance. Not only is it related to the energies of currents, as we saw in the last section, but it is also a "real" physical field in the sense that we described above. In classical mechanics it is clear that we can write the force on a particle as

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \quad (15.26)$$

so that, given the forces, everything about the motion is determined. In any region where $\mathbf{B} = 0$ even if A is not zero, such as outside a solenoid, there is no discernible effect of A . Therefore for a long time it was believed that A was not a "real" field. It turns out, however, that there are phenomena involving quantum mechanics which show that the field A is in fact a "real" field in the sense we have defined it. In the next section we will show you how that works.

15-5 The vector potential and quantum mechanics

There are many changes in what concepts are important when we go from classical to quantum mechanics. We have already discussed some of them in Vol. 1. In particular, the force concept gradually fades away, while the concepts of energy and momentum become of paramount importance. You remember that instead of particle motions, one deals with probability amplitudes which vary in space and time. In these amplitudes there are strengths related to momenta, and quantities related to energies. The momenta and energies, which determine the phases of wave functions, are therefore the important quantities in quantum mechanics. Instead of forces, we deal with the way interactions change the wavelength(s) of the waves. The idea of a force becomes quite secondary—if it is there at all. When people talk about nuclear forces, for example, what they usually analyze and work with are the energies of interaction of two nucleons, and not the force between them. Nobody ever calculates the energy to find out what the force looks like. In this section we want to describe how the vector and scalar potentials enter into quantum mechanics. It is, in fact, just because momentum and energy play a central role in quantum mechanics that A and ϕ provide the most direct way of introducing electrodynamic effects into quantum descriptions.

We must review a little how quantum mechanics works. We will consider again the imaginary experiment described in Chapter 17 of Vol. 1, in which elec-

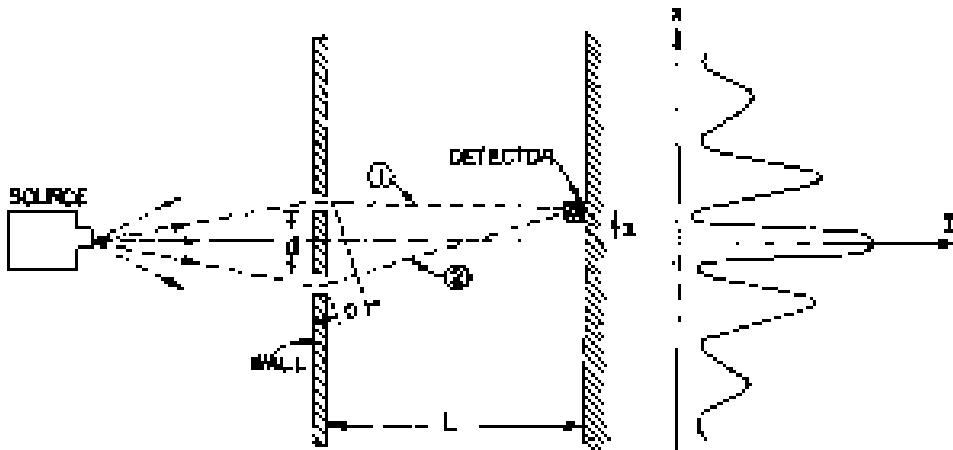


Fig. 15-5. An interference experiment with electrons
(see also Chapter 37 of Vol. II).

itrons are diffracted by two slits. The arrangement is shown again in Fig. 15-5. Electrons, all of nearly the same energy, leave the source and travel toward a wall with two narrow slits. Beyond the wall is a "backstop" with a movable detector. The detector measures the rate, which we call I , at which electrons arrive at a small region of the backstop at the distance x from the axis of symmetry. The rate is proportional to the probability that an individual electron that leaves the source will reach that region of the backstop. This probability has the complicated-looking distribution shown in the figure, which we understand as due to the interference of two amplitudes, one from each slit. The interference of the two amplitudes depends on their phase difference. That is, if the amplitudes are $C_1 e^{i\Phi_1}$ and $C_2 e^{i\Phi_2}$, the phase difference $\delta = \Phi_1 - \Phi_2$ determines their interference pattern [see Eq. 429.12 in Vol. II]. If the distance between the screen and the slits is L , and if the difference in the path lengths for electrons going through the two slits is a , as shown in the figure, then the phase difference of the two waves is given by

$$\delta = \frac{a}{\lambda}. \quad (15.27)$$

As usual, we let $\delta = \lambda/2\pi$, where λ is the wavelength of the space variation of the probability amplitude. For simplicity, we will consider only values of x much less than L ; then we can set

$$a = \frac{x}{L} d$$

and

$$\delta = \frac{x}{L} \frac{d}{\lambda}. \quad (15.28)$$

When x is zero, δ is zero: the waves are in phase, and the probability has a maximum. When x is λ , the waves are out of phase, they interfere destructively, and the probability is a minimum. So we get the wavy function for the electron intensity.

Now we would like to state the law that for quantum mechanics replaces the force law $F = -q\mathbf{E} \times \mathbf{B}$. It will be the law that determines the behavior of quantum-mechanical particles in an electromagnetic field. Since what happens is determined by amplitudes, the law must tell us how the magnetic influences affect the amplitudes; we are no longer dealing with the acceleration of a particle. The law is the following: the phase of the amplitude to arrive via any trajectory is changed by the presence of a magnetic field by an amount equal to the integral of the vector potential along the whole trajectory times the charge of the particle over Planck's constant. That is,

$$\text{Magnetic change in phase} = \frac{q}{h} \int_{\text{trajectory}} \mathbf{A} \cdot d\mathbf{s}. \quad (15.29)$$

If there were no magnetic field there would be a certain phase of arrival. If there is a magnetic field \mathbf{B} , then the phase of the arriving wave is increased by the integral in Eq. (15.29).

Although we will not need to use it for our present discussion, we note that the effect of an electrostatic field is to produce a phase change given by the negative of the time integral of the scalar potential ϕ :

$$\text{Electric charge in phase} = -\frac{q}{\hbar} \int \phi \, dt.$$

These two expressions are correct for only for static fields, but together give the correct result for any alternating field, static or dynamic. This is the law that replaces $P = g(\mathbf{E} + \mathbf{v} \times \mathbf{B})$. We want now, however, to consider only a static magnetic field.

Suppose that there is a magnetic field present in the two-slit experiment. We want to ask for the phase of arrival at the screen of the two waves whose paths pass through the two slits. Their interference determines where the maxima in the probability will be. We may call Φ_1 the phase of the wave along trajectory (1). If $\Phi_1(B=0)$ is the phase without the magnetic field, then when the field is turned on the phase will be

$$\Phi_1 = \Phi_1(B=0) + \frac{q}{\hbar} \int_{(1)} A \cdot d\mathbf{s}. \quad (15.30)$$

Similarly, the phase for trajectory (2) is

$$\Phi_2 = \Phi_2(B=0) + \frac{q}{\hbar} \int_{(2)} A \cdot d\mathbf{s}. \quad (15.31)$$

The interference of the waves at the detector depends on the phase difference

$$\delta = \Phi_2(B=0) - \Phi_1(B=0) + \frac{q}{\hbar} \int_{(2)-1} A \cdot d\mathbf{s} - \frac{q}{\hbar} \int_{(1)} A \cdot d\mathbf{s}. \quad (15.32)$$

The no field difference we will call $\delta(B=0)$, it is just the phase difference we have calculated above in Eq. (15.28). Also, we notice that the two integrals can be written as one integral that goes forward along (1) and back along (2); we call this the closed path (1-2). As we have

$$\delta = \delta(B=0) - \frac{q}{\hbar} \oint_{(1-2)} A \cdot d\mathbf{s}. \quad (15.33)$$

This equation tells us how the electron motion is changed by the magnetic field; with it we can find the new positions of the intensity maxima and minima at the backstop.

Before we do that, however, we want to raise the following interesting and important point. You remember that the vector potential function has some arbitrariness. Two different vector potential functions A and A' whose difference is the gradient of some scalar function ψ , both represent the same magnetic field, since the curl of a gradient is zero. They give, therefore, the same classical forces $q\mathbf{v} \times \mathbf{B}$. In quantum mechanics the effects depend on the vector potential, which of the many possible A -functions is correct?

The answer is that the same arbitrariness in A continues to exist for quantum mechanics. If in Eq. (15.33) we change A to $A' = A + \nabla\psi$, the integral on A becomes

$$\oint_{(1-2)} A \cdot d\mathbf{s} = \oint_{(1-2)} A' \cdot d\mathbf{s} + \oint_{(1-2)} \nabla\psi \cdot d\mathbf{s}.$$

The integral of $\nabla\psi$ is around the closed path (1-2), but the integral of the tangential component of a gradient on a closed path is always zero, by Stokes' theorem. Therefore both A and A' give the same phase differences and the same quantum-mechanical interference effects. In both classical and quantum theory it is only the curl of A that matters; any choice of the function of A which has the correct curl gives the correct physics.

The same conclusion is evident if we use the results of Section 14-1. There we found that the line integral of A around a closed path is the flux of B through the path, which here is the flux between points (1) and (2). Equation (15.33) can, if we wish, be written as

$$i = \mu(B - 0) + \frac{q}{\pi} [\text{Flux of } B \text{ between (1) and (2)}], \quad (15.34)$$

where by the flux of B we mean, as usual, the surface integral of the normal component of B . The result depends only on B , and therefore only on the curl of A .

Now because we can write the result in terms of B as well as in terms of A , you might be inclined to think that the B holds its own as a "real" field and that the A can still be thought of as an artificial construction. But the definition of "real" field that we originally proposed was based on the idea that a "real" field would not act on a particle from a distance. We can, however, give an example in which B is zero—or at least arbitrarily small—at any place where there is some chance to find the particle, so that it is not possible to think of it acting directly on them.

You remember that for a long solenoid carrying an electric current there is a B -field inside but none outside, while there is lots of A circulating around outside, as shown in Fig. 15-6. If we assume a situation in which electrons are to be found only outside of the solenoid—only where there is A —there will still be an influence on the motion, according to Eq. (15.33). Classically, that is impossible. Classically, the force depends only on B ; in order to know that the solenoid is carrying current, the particle must go through it. But quantum-mechanically you could find out that there is a magnetic field inside the solenoid by going around it—without ever going close to it!

Suppose that we put a very long solenoid of small diameter just behind the wall and between the two slits, as shown in Fig. 15-7. The diameter of the solenoid is to be much smaller than the distance d between the two slits. In these circumstances, the deflection of the electrons at the slit gives no appreciable probability that the electrons will get near the solenoid. What will be the effect on our interference experiment?

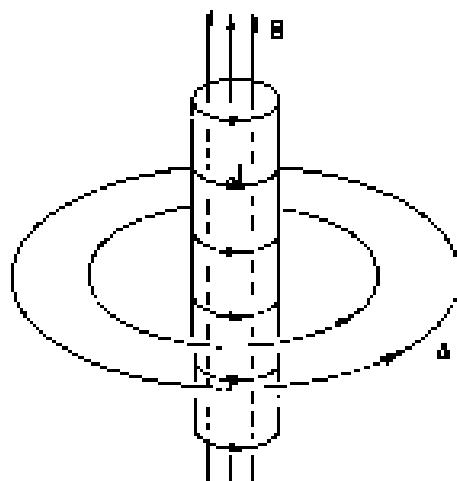


Fig. 15-6. The magnetic field and vector potential of a long solenoid.

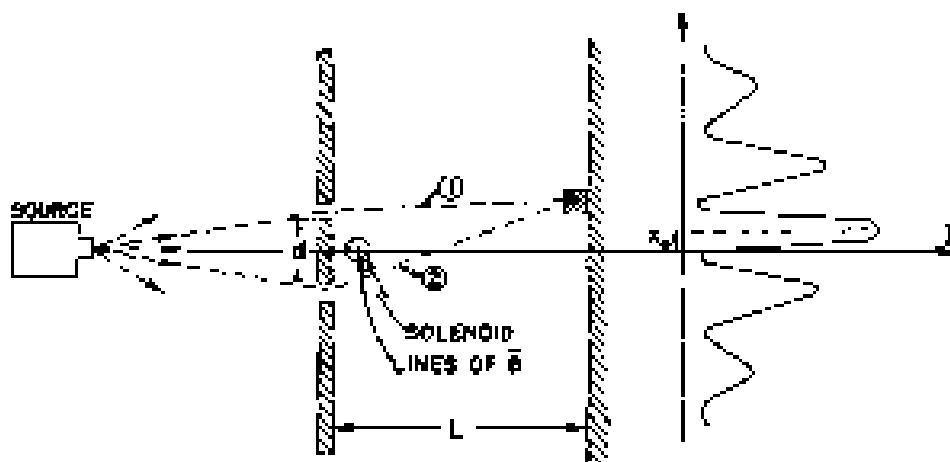


Fig. 15-7. A magnetic field can influence the motion of electrons even though it exists only in regions where there is an arbitrarily small probability of finding the electrons.

We compare the situation with and without a current through the solenoid. If we have no current, we have no B or A and we get the original pattern of electron intensity at the backstop. If we turn the current on in the solenoid and build up a magnetic field B inside, then there is an A outside. There is a shift in the phase difference proportional to the circulation of A outside the solenoid, which will mean that the pattern of maxima and minima is shifted to a new position. In fact, since the flux of B inside is a constant for any pair of paths, so also is the circulation of A . For every arrival point there is the same phase change! This corresponds

to shifting the entire pattern in x by a congruent amount, say x_0 , that we can easily calculate. The maximum intensity will occur where the phase difference between the two waves is zero. Using Eq. (15.32) or Eq. (15.33) for δ and Eq. (15.28) for $\delta(\theta = 0)$, we have

$$x_0 = -\frac{L}{\lambda} \lambda \frac{q}{k} \int_{n=2}^{\infty} A \cdot d\mathbf{r}, \quad (15.35)$$

or

$$x_0 = -\frac{I}{q} \lambda \frac{q}{k} [\text{flux of } \mathbf{B} \text{ between (1) and (2)}]. \quad (15.36)$$

The pattern with the solenoid in place should appear* as shown in Fig. 15-7. At least, that is the prediction of quantum mechanics.

Precisely this experiment has recently been done. It is a very, very difficult experiment. Because the wavelength of the electrons is so small, the apparatus must be on a tiny scale to observe the interference. The slits must be very close together, and that means that one needs an exceedingly small solenoid. It turns out that in certain circumstances, iron crystals will grow in the form of very long, microscopically thin filaments called whiskers. When these iron whiskers are magnetized they are like a tiny solenoid, and there is no field outside except near the ends. The electron interference experiment was done with such a whisker between two slits, and the predicted displacement in the pattern of electrons was observed.

In our sense then, the A -field is "real." You may say: "But there was a magnetic field." There was, but remember our original idea—that a field is "real" if it is what must be specified at the position of the particle in order to get the motion. The B -field in the whisker acts at a distance. If we want to describe its influence not as action-at-a-distance, we must use the vector potential.

This subject has an interesting history. The theory we have described was known from the beginning of quantum mechanics in 1926. The fact that the vector potential appears in the wave equation of quantum mechanics (called the Schrödinger equation) was obvious from the day it was written. That it cannot be replaced by the magnetic field in any easy way was observed by one soon after the other who tried to do so. This is also clear from our example of electrons moving in a region where there is no field and being affected nevertheless. But because in classical mechanics A did not appear to have any direct importance and, furthermore, because it could be changed by adding a gradient, people repeatedly said that the vector potential had no direct physical significance—that only the magnetic and electric fields are "right" over in quantum mechanics. It seems strange in retrospect that no one thought of discussing this experiment until 1956, when Bohm and Aharonov first suggested it and made the whole question crystal clear. The implication was there all the time, but no one paid attention to it. Thus many people were rather shocked when the matter was brought up. That's why someone thought it would be worth while to do the experiment to see that it really was right, even though quantum mechanics, which had been believed for so many years, gave an unequivocal answer. It is interesting that something like this can be around for thirty years but, because of certain prejudices of what is real and is not significant, continues to be ignored.

Now we wish to continue in our analysis a little further. We will show the connection between the quantum-mechanical formula and the classical formula to show why it turns out that if we look at things on a large enough scale it will look as though the particles are acted on by a force equal to $qv \times \mathbf{B}$ or the curl of A . To get classical mechanics from quantum mechanics, we need to consider cases in which all the wavelengths are very small compared with distances over which external conditions, like L and a , vary appreciably. We shall not prove the result in great generality, but only in a very simple example, to show how it works. Again we consider the same slit experiment. But instead of putting all the magnetic field in a very tiny region between the slits, we imagine a magnetic field that extends

* If the field \mathbf{B} curves out of the plane of the figure, the flux as we have defined it is negative and x_0 is positive.

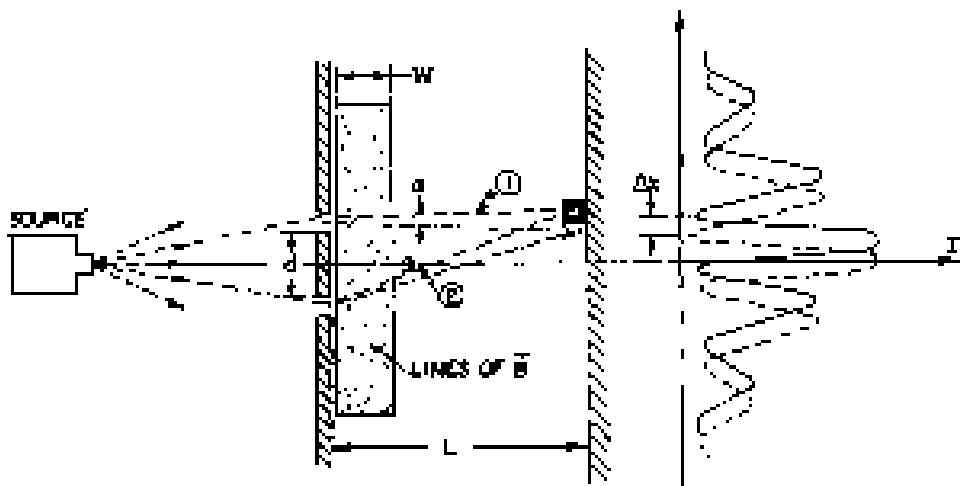


Fig. 15-8. The shift of the interference pattern due to a strip of magnetic field.

over a larger region behind the slits, as shown in Fig. 15-8. We will take the idealized case where we have a magnetic field which is uniform in a narrow strip of width w , considered small, as compared with L . (This can easily be arranged; the magnet can be put as far out as we want.) In order to calculate the shift in phase, we must take the two integrals of A along the two trajectories (1) and (2). They differ, as we have seen, merely by the flux of B between the paths. To our approximation, the flux is fixed. The phase difference for the two paths is then

$$\delta = \delta(B=0) + \frac{q}{h} B w d. \quad (15.37)$$

We note that, to our approximation, the phase shift is independent of the angle. So again, the effect will be to shift the whole pattern upward by an amount Δx . Using Eq. (15.28),

$$\Delta x = \frac{L\lambda}{\theta} \Delta\phi = \frac{L\lambda}{d} [\delta - \delta(B=0)].$$

Using (15.37) for $\delta - \delta(B=0)$,

$$\Delta x = L\lambda \frac{q}{h} B w. \quad (15.38)$$

Such a shift is equivalent to deflecting all the trajectories by the small angle α (see Fig. 15-8), where

$$\alpha = \frac{\Delta x}{L} = \frac{\lambda}{\theta} q B w. \quad (15.39)$$

Now classically we would also expect a thin strip of magnetic field to deflect all trajectories through some small angle, say α' , as shown in Fig. 15-9(a). As the electrons go through the magnetic field, they feel a transverse force $qv \times B$ which leads to a finite v/v . The change in their transverse momentum is just equal to this impulse, so

$$\Delta p_x = qvB. \quad (15.40)$$

The angular deflection [Fig. 15-9(b)] is equal to the ratio of this transverse momentum to the total momentum p . We get that

$$\alpha' = \frac{\Delta p_x}{p} = \frac{qvB}{p}. \quad (15.41)$$

We can compare this result with Eq. (15.39), which gives the same quantity computed quantum-mechanically. But the connection between classical mechanics and quantum mechanics is this: A particle of momentum p corresponds to a quan-

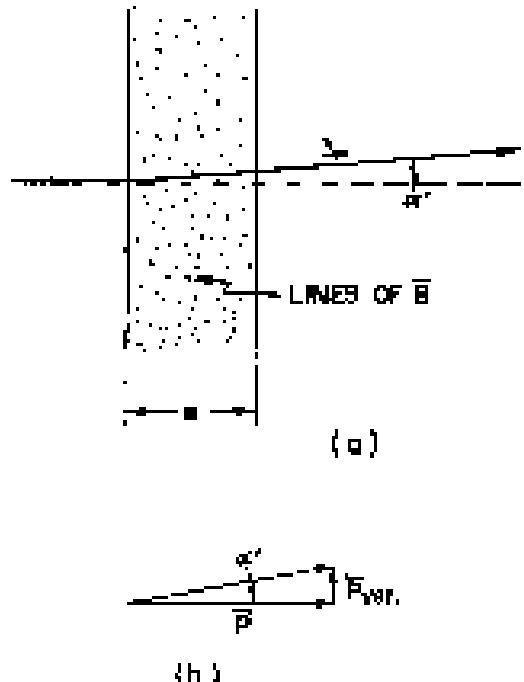


Fig. 15-9. Deflection of a particle due to passage through a strip of magnetic field.

term amplitude varying with the wavelength $\lambda = \hbar/\omega$. With this equality, ϕ and ψ are identical; the electrical and quantum calculations give the same result.

From the analysis we see how it is that the vector potential which appears in quantum mechanics in an explicit form produces a classical force which depends only on its derivatives. In quantum mechanics what matters is the interference between nearby paths; it always turns out that the effects depend only on how much the field A changes from point to point, and therefore only on the derivatives of A and not on the value itself. Nevertheless, the vector potential A together with the scalar potential ϕ that goes with it happens to give the most direct description of the physics. This becomes more and more apparent the more deeply we go into the quantum theory. In the general theory of quantum electrodynamics, one takes the vector and scalar potentials as the fundamental quantities in a set of equations that replace the Maxwell equations; E and B are slowly disappearing from the modern expression of physical laws; they are being replaced by A and ϕ .

15-6 What is true for statics is false for dynamics

We are now at the end of our exploration of the subject of static fields. Already in this chapter we have come perilously close to having to worry about what happens when fields change with time. We were barely able to avoid it in our treatment of magnetic energy by taking refuge in a relativistic argument. Even so, our treatment of the energy problem was somewhat artificial and perhaps even misleading, because we ignored the fact that moving coils must, in fact, produce changing fields. It is now time to take up the treatment of time-varying fields—the subject of electrodynamics. We will do so in the next chapter. First, however, we would like to emphasize a few points.

Although we began this course with a presentation of the complete and correct equations of electromagnetism, we immediately began to study some incomplete pieces—because that was easier. There is a great advantage in starting with the simpler theory of static fields, and proceeding only later to the more complicated theory which includes dynamic fields. There is less new material to learn all at once, and there is time for you to develop your intellectual muscles in preparation for the bigger task.

But there is one danger in this process that before we get to see the complete story, the incomplete truths learned en route may become ingrained and taken as the whole truth—that what is true and what is only sometimes true will become confused. So we give in Table 15-1 a summary of the important formulas we have covered, separating those which are true in general from those which are true for statics, but false for dynamics. This summary also shows, in part, where we are going, since as we treat dynamics we will be developing in detail what we must just state here without proof.

It may be useful to make a few remarks about the table. First, you should notice that the equations we started with are the *true* equations; we have not misled you there. The electromagnetic force law, called the *Ampère law*, $F = q(E + v \times B)$ is true. It is only Coulomb's law that is false, in as much only for statics. The four Maxwell equations for E and B are also true. The equations we took for statics are false, of course, because we left off all terms with time derivatives.

Gauss' law, $\nabla \cdot E = \rho/\epsilon_0$, continues, but the curl of E is not zero in general. So E cannot always be equivalent to the gradient of a scalar—the electrostatic potential. We will see that a scalar potential will remain, but it is a time-varying quantity that must be used together with vector potentials for a complete description of the electric field. The equations governing this new scalar potential are, necessarily, also new.

We must also give up the idea that E is zero in conductors. When the fields are changing, the charges in conductors do not, in general, have time to rearrange themselves to make the field zero. They are set in motion, but never reach equilibrium. The only general statement is: electric fields in conductors produce currents.

Table 15-1

ELECTRICITY IN GENERAL (true only for statics)		TRUE ALWAYS	
$E = \frac{1}{4\pi\epsilon_0} \frac{q\phi}{r^2}$	(Coulomb's law)	$E = q(E + v \times B)$	(Lorentz force)
		$\rightarrow \nabla \cdot E = \frac{\rho}{\epsilon_0}$	(Gauss' law)
$\nabla \times E = 0$		$\rightarrow \nabla \times E = -\frac{\partial B}{\partial t}$	(Faraday's law)
$E = -\nabla\phi$		$E = -\nabla\phi = \frac{\partial A}{\partial r}$	
$B(t) = \frac{1}{4\pi\epsilon_0} \frac{\rho(2)\phi_{ext}}{r_{ext}^2} dV_2$			
For conductors, $E = 0$, $\phi = \text{constant}$, $Q = CV$		In a conductor, E makes current j	
$c^2 \nabla \times B = \frac{j}{\epsilon_0}$	(Ampere's law)	$\rightarrow \nabla \cdot B = 0$	(No magnetic charges)
$B(t) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{\rho(2)}{r_{ext}^2} \frac{\phi_{ext}}{r_{ext}^2} dV_2$		$B = \nabla \times A$	
$\nabla^2 \phi = -\frac{\rho}{\epsilon_0}$	(Poisson's equation)	$\rightarrow c^2 \nabla \times B = \frac{1}{\epsilon_0} j + \frac{\partial A}{\partial t}$	
$\left. \begin{array}{l} \nabla^2 A = -\frac{j}{\epsilon_0 c^2} \\ \text{with} \\ \nabla \cdot A = 0 \end{array} \right\}$		$\left. \begin{array}{l} \nabla^2 \phi = \frac{1}{\epsilon_0} \frac{\partial^2 \phi}{\partial t^2} = -\frac{\rho}{\epsilon_0} \\ \text{and} \\ \nabla^2 A = \frac{1}{\epsilon_0} \frac{\partial^2 A}{\partial t^2} = -\frac{j}{\epsilon_0 c^2} \\ \text{with} \\ c^2 \nabla \cdot A + \frac{\partial \phi}{\partial t} = 0 \end{array} \right\}$	
$\phi(t) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(2)}{r_{ext}} dV_2$		$\left. \begin{array}{l} \phi(t) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(2, t')}{r_{ext}} dV_2 \\ \text{and} \\ A(t, t) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{j(2, t')}{r_{ext}} dV_2 \\ \text{with} \\ t' = t - \frac{r_{ext}}{c} \end{array} \right\}$	
$A(t) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{j(2)}{r_{ext}} dV_2$			
$V = \frac{1}{2} \int \rho\phi dV + \frac{1}{2} \int j \cdot A dV$		$W = \int \left(\frac{\epsilon_0}{2} E \cdot E + \frac{\mu_0}{2} B \cdot B \right) dV$	

The equations marked by an arrow (\rightarrow) are Maxwell's equations.

tents. So in varying fields a conductor is not an equipotential. It also follows that the idea of a capacitor is no longer precise.

Since there are no magnetic charges, the divergence of \mathbf{B} is always zero. So \mathbf{B} can always be equated to $\nabla \times \mathbf{A}$. (Everything doesn't change!) But the generation of \mathbf{B} is not only from currents: $\nabla \times \mathbf{B}$ is proportional to the current density plus a new term $\partial \mathbf{A} / \partial t$. This means that \mathbf{A} is related to currents by a new equation. It is also related to ϕ . If we make use of our freedom to choose $\nabla \cdot \mathbf{A}$ for our own convenience, the equations for \mathbf{A} and ϕ can be arranged to take on a simple and elegant form. We therefore make the condition that $\nabla^2 \nabla \cdot \mathbf{A} = -\partial \mathbf{A} / \partial t$, and the differential equations for \mathbf{A} and ϕ appear as shown in the table.

The potentials A and ϕ can still be found by integrals over the currents and charges, but not the same integrals as for statics. Most wonderfully, though, the true integrals are like the static ones, with only a small and physically appealing modification. When we do the integrals to find the potentials at some point, say point (1) in Fig. 15-10, we must use the values of J and ρ at the point (2) at an earlier time $t' = t - r_{12}/c$. As you would expect, the influences propagate from point (2) to point (1) at the speed c . With this small change, one can solve for the fields of varying currents and charges, because once we have A and ϕ , we get \mathbf{E} from $\nabla \times \mathbf{A}$, as before, and \mathbf{B} from $\nabla \phi - A_t / c$.

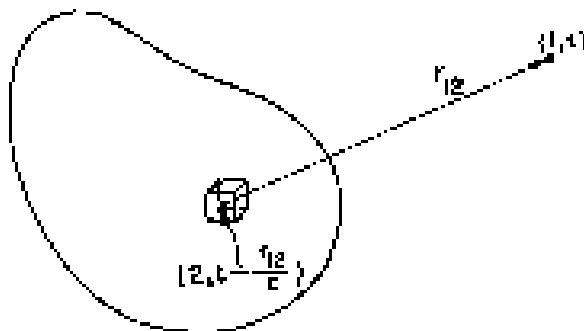


Fig. 15-10. The potentials at point (1) and at the time t are given by summing the contributions from each element of the source at the moving point (2), using the currents and charges which were present at the earlier time $t - r_{12}/c$.

Finally, you will notice that some results—for example, that the energy density in an electric field is $\epsilon_0 E^2/2$ —are true for electrodynamics as well as for statics. You should not be misled into thinking that this is at all “natural.” The validity of any formula derived in the static case must be demonstrated over again for the dynamic case. A concrete example is the expression for the electrostatic energy, in terms of a volume integral of ρu . This result is true only for statics.

We will consider all these matters in more detail in due time, but it will perhaps be useful to keep in mind this summary, so you will know what you can forget, and what you should remember as always true.

Induced Currents

16-1 Motors and generators

The discovery in 1820 that there was a close connection between electricity and magnetism was very exciting—until then, the two subjects had been considered to quite independent. The first discovery was that currents in wires make magnetic fields; then, in the same year, it was found that wires carrying current in a magnetic field have forces on them.

One of the immediate questions whenever there is a new electrical force is the possibility of using it in an engine to do work. Almost immediately after their discovery, people started to design electric motors using the forces on current-carrying wires. The principle of the electrodynamic motor is shown in bare outline in Fig. 16-1. A rectangular magnet—usually with some pieces of soft iron—is used to produce a magnetic field in two slots. Across each slot there is a north and south pole, as shown. A rectangular coil of copper is placed with one side in each slot. When a current passes through the coil, it flows in opposite directions in the two slots, so the forces are also opposite, producing a torque on the coil about the axis shown. If the coil is mounted so as to turn it can turn, it can be coupled to pulleys or gears and can do work.

The same idea can be used for making a sensitive instrument for electrical measurements. That the moment the force law was discovered the precision of electrical measurements was greatly increased. First, the torque of such a motor can be made much greater for a given current by making the current go around many turns instead of just one. Then the coil can be mounted so that it turns with very little torque either by supporting its shaft on very delicate jewel bearings or by hanging the coil on a very fine wire or a quartz fiber. Then an exceedingly small current will move the coil, and for small angles the amount of rotation will be proportional to the current. The rotation can be measured by gluing a pointer to the coil or, for the most delicate instruments, by attaching a small mirror to the coil and looking at the shift of the image of a scale. Such instruments are called galvanometers. Voltmeters and ammeters work on the same principle.

The same ideas can be applied on a large scale to make large motors for providing mechanical power. The coil can be made to go around and around by arranging that the connections to the coil are reversed each half-turn by contacts mounted on the shaft. Then the torque is always in the same direction. Small dc motors are made just this way. Larger motors, dc or ac, are often made by replacing the permanent magnet by an electromagnet energized from the electrical power source.

With the realization that electric currents make magnetic fields, people immediately suggested that somehow or other, magnetic fields also make electric fields. Various experiments were tried. For example, two wires were placed parallel to each other and a current was passed through one of them in the hope of inducing a current in the other. The thought was that the magnetic field might in some way drag the electrons along in the second wire, giving some sort of law of "like prefers to move alike." With the largest available current and the most sensitive galvanometer to detect any current, the result was negative. Large magnets next to wires also produced no observed effects. Finally, Faraday discovered in 1831 the essential feature that had been missed—that electric effects exist only when there is something changing. If one of a pair of wires has a changing current, a current is induced in the other. Or if a magnet is moved near an electric circuit, there is a current. Why this occurs is not known. This was the induction effect rediscovered

16-1 Motors and generators

16-2 Transformers and inductances

16-3 Forces on induced currents

16-4 Electrical technology

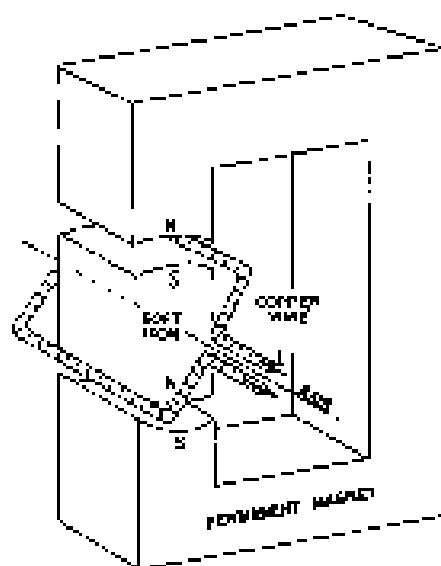


Fig. 16-1. Schematic outline of a simple electromagnetic motor.

by Faraday. It transformed the rather dull subject of static fields into a very exciting dynamic subject with an enormous range of wonderful phenomena. This chapter is devoted to a qualitative description of some of them. As we will see, one can quickly get into fairly complicated situations that are hard to analyze quantitatively in all their details. But never mind, our main purpose in this chapter is just to acquaint you with the phenomena involved. We will take up the detailed analysis later.

We can easily understand one feature of magnetic induction from what we already know, although it was not known in Faraday's time. It comes from the $v \times B$ force on a moving charge that is proportional to its velocity in a magnetic field. Suppose that we have a wire which passes near a magnet, as shown in Fig. 16-2, and that we connect the ends of the wire to a galvanometer. If we move the wire across the end of the magnet the galvanometer pointer moves.

The magnet produces some vertical magnetic field, and when we push the wire across the field, the electrons in the wire feel a sideways force—not right angles to the field and to the wire. The field pushes the electrons along the wire. But why does this move the galvanometer, which is so far from the force? Because when the electrons which feel the magnetic force try to move, they pull—by electric repulsion—the electrons a little farther down the wire; these, in turn, repel the electrons a little further on, and so on for a long distance. An amazing thing.

It was so interesting to Gauss and Weber—who first built a galvanometer—that they tried to see how far the forces in the wire would go. They strung a wire all the way across their city. Mr. Gauss, at one end, connected the wires to a battery (batteries were known before generators) and Mr. Weber watched the galvanometer move. They had a way of signaling long distances—it was the beginning of the telegraph! Of course, this has nothing directly to do with induction—it has to do with the very wires carrying currents, whether the currents are pushed by induction or not.

Now suppose in the setup of Fig. 16-2 we leave the wire alone and move the magnet. We still see no effect on the galvanometer. As Faraday discovered, moving the magnet under the wire—one way—has the same effect as moving the wire over the magnet—the other way. But when the magnet is moved, we no longer have any $v \times B$ force on the electrons in the wire. This is the new effect that Faraday found. Today, we might hope to understand it from a relativistic argument.

We already understand that the magnetic field of a magnet comes from its internal currents. So we expect to observe the same effect if instead of a magnet in Fig. 16-2 we use a coil of wire in which there is a current. If we move the wire past the coil there will be a current through the galvanometer, or also if we move the coil past the wire. But there is now a more exciting thing: If we change the magnetic field of the coil not by moving it, but by changing its current, there is again an effect in the galvanometer. For example, if we have a loop of wire near a coil, as shown in Fig. 16-3, and if we keep both of them stationary but switch off the current, there is a pulse of current through the galvanometer. When we switch the coil on again, the galvanometer kicks in the other direction.

Whenever the galvanometer in a situation such as the one shown in Fig. 16-2, or in Fig. 16-3, has a current, there is a net push on the electrons in the wire in one direction along the wire. There may be pushes in different directions at different places, but there is more push in one direction than another. What counts is the push integrated around the complete circuit. We call this net integrated push the electromotive force (abbreviated emf) in the circuit. More precisely, the emf is defined as the tangential force per unit charge in the wire integrated over length, once around the complete circuit. Faraday's complete discovery was that emf's can be generated in a wire in three different ways: by moving the wire, by moving a magnet near the wire, or by changing a current in a nearby wire.

Let's consider the simple machine of Fig. 16-1 again, only now, instead of putting a current through the wire to make it turn, let's turn the loop by an external force, for example by hand or by a waterwheel. When the coil rotates, its wires are moving in the magnetic field and we will find an emf in the circuit of the coil. The motor becomes a generator.

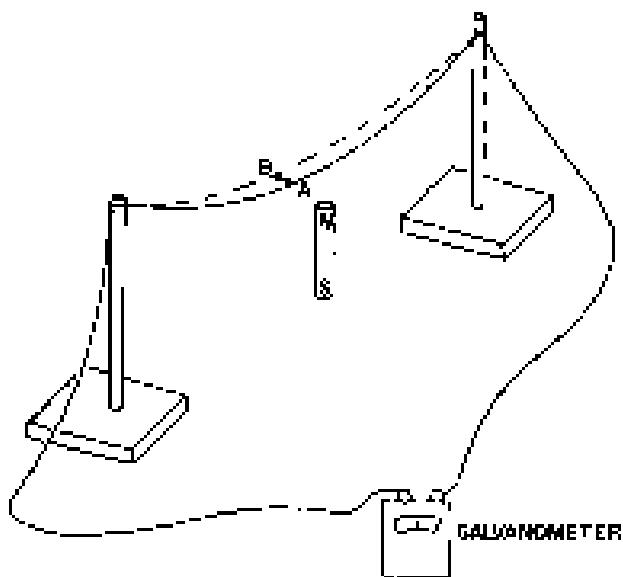


Fig. 16-2. Moving a wire through a magnetic field produces a current, as shown by the galvanometer.

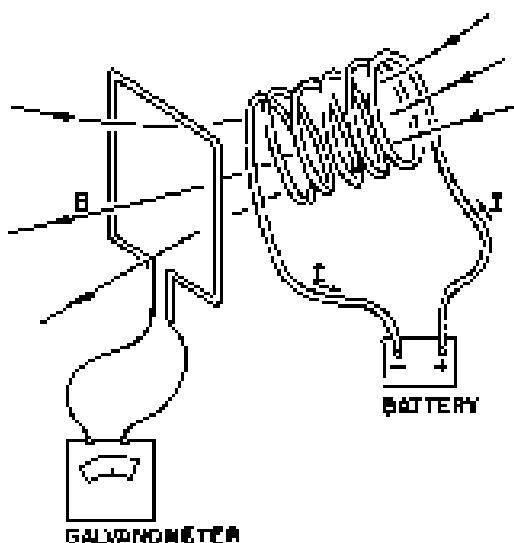


Fig. 16-3. A coil with current produces a current in a second coil if the first coil is moved or if its current is changed.

The coil of the generator has an induced emf from its motion. The amount of the emf is given by a simple rule discovered by Faraday. (We will just state the rule now and wait until later to examine it in detail.) The rule is that when the magnetic flux that passes through the loop that is the normal component of \mathbf{B} integrated over the area of the loop is changing with time, the emf is equal to the rate of change of the flux. We will refer to this as "the DRS rule." You see that when the coil of Fig. 16-1 is rotated, the flux through it changes. At the start some flux goes through one way; then when the coil has rotated 180° the same flux goes through the other way. If we continuously rotate the coil the flux is first positive, then negative, then positive, and so on. The rate of change of the flux must alternate also. So there is an alternating emf in the coil. If we connect the two ends of the coil to outside wires through some sliding contacts called slip-slugs (just so the wires won't get twisted) we have an alternating-current generator.

Or we can also manage, by means of some sliding contacts, that after every one-half rotation, the connection between the coil ends and the outside wires is reversed, so that when the emf reverses, so do the connections. Then the pulses of emf will always push currents in the same direction through the external circuit. We have what is called a direct-current generator.

The machine of Fig. 16-1 is either a motor or a generator. The reciprocity between motors and generators is nicely shown by using two identical dc "motors" of the permanent magnet kind, with their coils connected by two copper wires. When the shaft of one is turned mechanically, it becomes a generator and drives the other as a motor. If the shaft of the second is turned, it becomes the generator and drives the first as a motor. So here is an interesting example of a new kind of equivalence of nature; motor and generator are equivalent. The quantitative equivalence is, in fact, not completely accidental. It is related to the law of conservation of energy.

Another example of a device that can operate either to generate emf's or to respond to emf's is the receiver of a standard telephone—that is, ac "earphone." The original telephone of Bell consisted of two such "earphones," connected by two long wires. The basic principle is shown in Fig. 16-4. A permanent magnet produces a magnetic field in two "yokes" of soft iron, and in a thin diaphragm that is moved by sound pressure. When the diaphragm moves, it changes the amount of magnetic field in the yokes. Therefore a coil of wire wound around one of the yokes will have the flux through it changed when a sound wave hits the diaphragm.

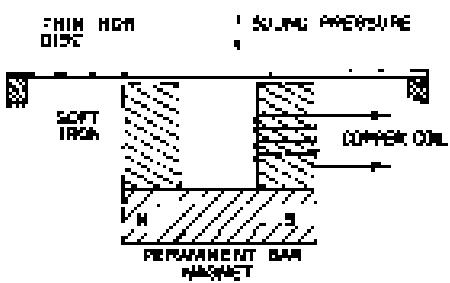


Fig. 16-4. A telephone transducer or receiver.

So there is an emf in the coil. If the ends of the coil are connected to a circuit, a current, which is an electrical representation of the sound is set up.

If the ends of the coil of Fig. 16-4 are connected by two wires to another identical gauge, varying currents will flow in the second coil. These currents will produce a varying magnetic field and will make a varying attraction on the iron diaphragm. The diaphragm will wiggle and make sound waves approximately similar to the ones that moved the original diaphragm. With a few bits of iron and copper the human voice is transmitted over wires!

(The modern home telephone uses a receiver like the one described but uses an improved invention to get a more powerful transmitter. It is the "carbon-bubble microphone," that uses sound pressure to vary the electric current from a battery.)

16-2 Transformers and inductances

One of the most interesting features of Faraday's discovery is just that an coil *causes* a driving coil—which we can understand in terms of the magnetic force $\mu_0 H \times B$ —but that a changing current in one coil makes an emf in a second coil. And quite surprisingly the amount of emf induced in the second coil is given by the same “flux rule”: that the emf is equal to the rate of change of the flux through the coil. Suppose that we take two coils, each wound around separate bundles of iron sheets (these help to make stronger magnetic fields), as shown at Fig. 16-5. Now we connect one of the coils—coil (a)—to an alternating-current generator. The continually changing current produces a continually varying magnetic field. This varying field generates an alternating emf in the second coil—coil (b). It is an emf, for example, produced enough power to light an electric bulb.

The emf alternates in coil (b) at a frequency which is of course, the same as the frequency of the original generator. But the current in coil (a) can be larger or smaller than the current in coil (b). The current in coil (b) depends on the emf induced in it and on the resistance and inductance of the rest of its circuit. The emf can be less than that of the generator if, say, there is little flux change. Or the emf in coil (b) can be much much larger than that in the generator by winding coil (b) with many turns. Even in a given magnetic field the flux through the coil is then greater. (Or if you prefer to look at it another way, the emf is the same in each turn, and since the total emf is the sum of the emfs of the separate turns, many turns in series produce a large emf.)

Such a combination of two coils—usually with an arrangement of iron sheets to guide the magnetic fields—is called a “transformer.” It can “transform” one emf (also called a “voltage”) to another.

There are also induction effects in a single coil. For instance, in the setup in Fig. 16-3 there is a changing flux not only through coil (b), which lights the bulb, but also through coil (a). The varying current in coil (a) produces a varying magnetic field inside铁心 and the flux of this field is continually changing, so there is a self-induced emf in coil (a). There is an emf acting on any current when it is building up a magnetic field—or, in general, when the field is changing in any way. The effect is called *self-inductance*.

When we gave “the flux rule” that the emf is equal to the rate of change of the flux linkage, we didn't specify the direction of the emf. There is a simple rule, called Lenz's rule, for figuring out which way the emf goes: the emf tries to oppose any flux change. That is, the direction of an induced emf is always such that if a current were to flow in the direction of the emf, it would produce a flux of *B* that opposes the change in *B* that produces the emf. Lenz's rule can be used to find the direction of the emf in the generator of Fig. 16-1, or in the transformer windings of Fig. 16-5.

In particular, if there is a changing current in a single coil (or in any wire) there is a “back” emf in the circuit. This emf acts on the charges flowing in coil (a) of Fig. 16-5 to oppose the change in magnetic field, and so in the direction to oppose the change in current. It tries to keep the current constant; it is opposite in the current when the current is increasing, and it is in the direction of the current

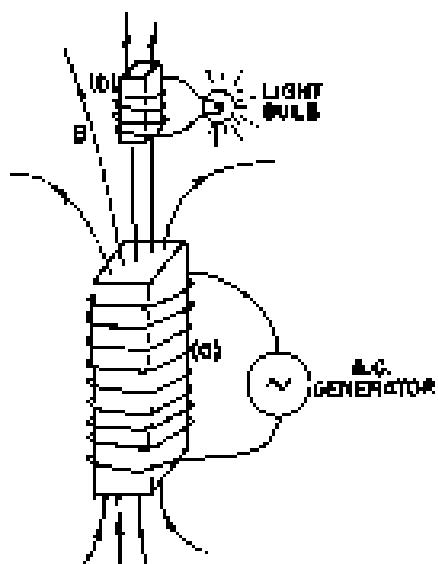


Fig. 16-5. Two coils, wrapped around bundles of iron sheets, allow a generator to light a bulb with no direct connection.

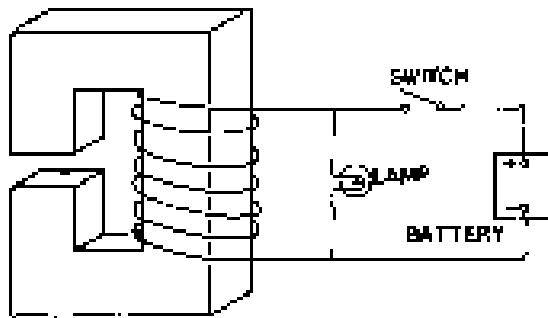


Fig. 16-6. Circuit connections for an electromagnet. The lamp shows the passage of current when the switch is opened, preventing the appearance of excessive emfs.

when it is decreasing. A current in a self-inductance has "inertia," because the inductive effects try to keep the flow constant, just as mechanical inertia tries to keep the velocity of an object constant.

Any large electromagnet will have a large self-inductance. Suppose that a battery is connected to the coil of a large electromagnet, as in Fig. 16-6, and that a strong magnetic field has been built up. (The current reaches a steady value determined by the battery voltage and the resistance of the wire in the coil.) But now suppose that we try to disconnect the battery by opening the switch. If we really opened the circuit, the current would go to zero rapidly, and in doing so it would generate an enormous emf. In most cases this emf would be large enough to develop an arc across the opening contacts of the switch. The high voltage that appears might also damage the insulation of the coil—or you, if you are the person who opens the switch! For these reasons, electromagnets are usually connected in a circuit like the one shown in Fig. 16-6. When the switch is opened, the current does not change rapidly but remains steady, flowing instead through the lamp, being driven by the coil due to the self-inductance of the coil.

16-3 Forces on induced currents

You have probably seen the dramatic demonstration of Lenz's rule made with the paddle shown in Fig. 16-7. It is an electromagnet, just like coil (a) of Fig. 16-5. An aluminum ring is placed on the end of the magnet. When the rod is connected to an alternating-current generator by closing the switch, the ring flies into the air. The force results, of course, from the induced currents in the ring. The fact that the ring flies away shows that the currents in it oppose the change of the field through it. When the magnet is reaching a double-poled stage, the induced current in the ring is making a downward-point north pole. The ring and the coil are repelled over the two magnets with like poles opposite. If a thin radial cut is made in the ring the force disappears, showing that it does indeed come from the currents in the ring.

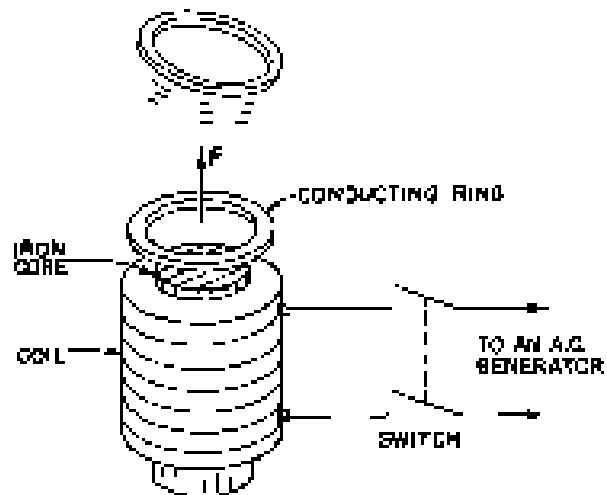


Fig. 16-7. A conducting ring is strongly repelled by an electromagnet with a varying current.

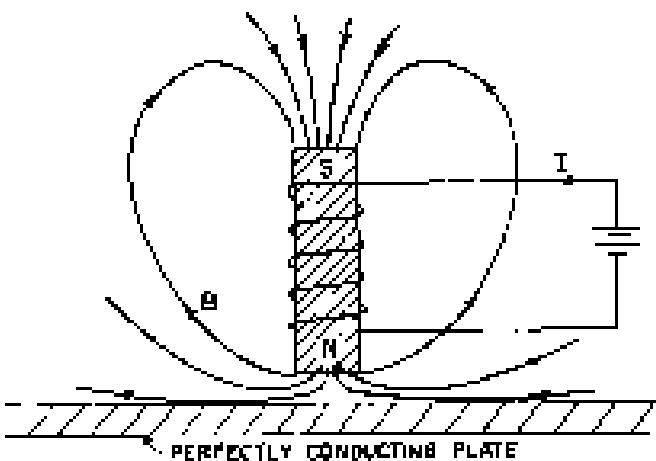


Fig. 16-8. An electromagnet near a perfectly conducting plate.

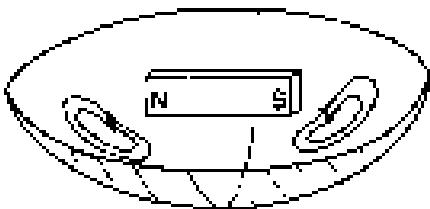


Fig. 16-9. A bar magnet is suspended above a superconducting bowl, by the repulsion of eddy currents.

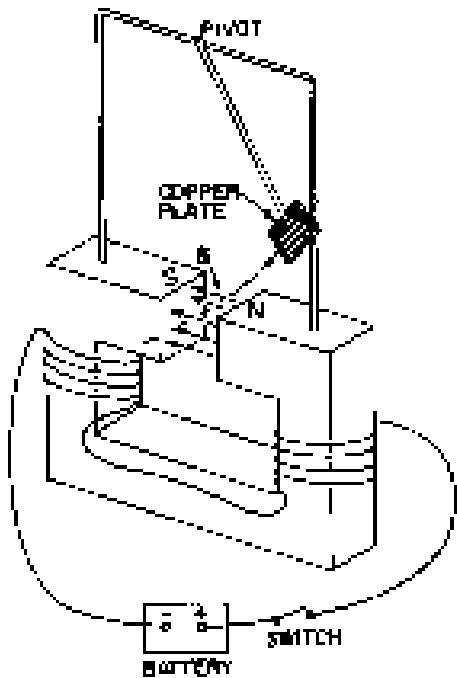


Fig. 16-10. The breaking of the pendulum shows the forces due to eddy currents.

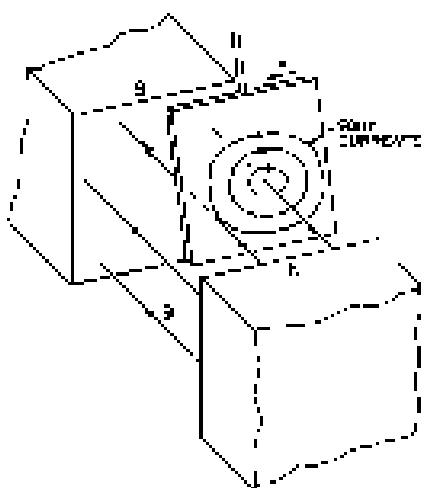


Fig. 16-11. The eddy currents in the copper pendulum.

If, instead of the ring, we place a disc of aluminum or copper across the end of the electromagnet of Fig. 16-7, it is also repelled; induced currents circulate in the material of the disc, and again produce a repulsion.

An interesting effect, similar in origin, occurs with a sheet of a perfect conductor. In a "perfect conductor" there is no resistance whatever to the current. So if currents are generated in it, they can keep going forever. In fact, the slight emf would generate an arbitrarily large current which really means that there can be no emfs at all. Any attempt to make a magnetic flux go through such a sheet generates currents that create opposite B fields—all with infinitesimal emfs, so with no flux entering.

If we have a sheet of a perfect conductor and put an electromagnet next to it, when we turn on the current in the magnet, currents called eddy currents appear in the sheet, so that no magnetic flux enters. The field lines would look as shown in Fig. 16-8. The same thing happens, of course, if we bring a bar magnet near a perfect conductor. Since the eddy currents are creating opposing fields, the magnets are repelled from the conductor. This makes it possible to suspend a bar magnet in air above a sheet of perfect conductor shaped like a dish, as shown in Fig. 16-9. The magnet is suspended by the repulsion of the induced eddy currents in the perfect conductor. These are no perfect conductors at ordinary temperatures, but some materials become perfect conductors at low enough temperatures. For instance, below 1.3°K tin conducts perfectly. It is called a superconductor.

If the conductor in Fig. 16-8 is not quite perfect there will be some resistance to flow of the eddy currents. The currents will tend to die out and the magnet will slowly settle down. The eddy currents in an imperfect conductor need not stop being going, and to have an emf the flux must keep changing. The loss of the magnetic field gradually punishes the conductor.

In a normal conductor, there are not only repulsive forces from eddy currents, but there can also be sideways forces. For instance, if we move a magnet sideways along a conducting surface the eddy currents produce a force of drag, because the induced currents are opposing the change in the location of flux. Such forces are proportional to the velocity and are like a kind of viscous force.

These effects show up nicely in the apparatus shown in Fig. 16-10. A square sheet of copper is suspended on the end of a rod to make a pendulum. The copper swings back and forth between the poles of an electromagnet. When the magnet is turned on, the pendulum motion is suddenly arrested. As the metal plate cuts the flux of the magnet, there is a current induced in the plate which acts to oppose the change in flux through the plate. If the sheet were a perfect conductor, the currents would be so great that they would push the plate out again—it would bounce back. With a copper plate there is some resistance in the plate, so the currents start by bringing the plate almost to a dead stop as it starts to enter the field. Then, as the currents die down, the plate slowly settles to rest in the magnetic field.

The nature of the eddy currents in the copper pendulum is shown in Fig. 16-11. The strength and geometry of the currents are quite sensitive to the shape of the plate. If, for instance, the copper plate is replaced by one which has several narrow slits cut in it, as shown in Fig. 16-12, the eddy-current effects are drastically reduced. The pendulum swings through the magnetic field with only a small retarding force. The reason is that the currents in each section of the copper have less flux to drive them, so the effects of the resistance of each loop are greater. The currents are smaller and the drag is less. The viscous character of the force is seen even more clearly if a sheet of copper is placed between the poles of the magnet of Fig. 16-10 and then released. It doesn't fall; it just sinks slowly downward. The eddy currents exert a strong resistance to the motion—just like the viscous drag in water.

If, instead of dragging a conductor past a magnet, we try to rotate it in a magnetic field, there will be a resistive torque from the same effects. Alternatively, if we rotate a magnet—and over and over—a conducting plate or ring, the ring is dragged around; currents in the ring will create a torque that tends to relax the ring with the magnet.

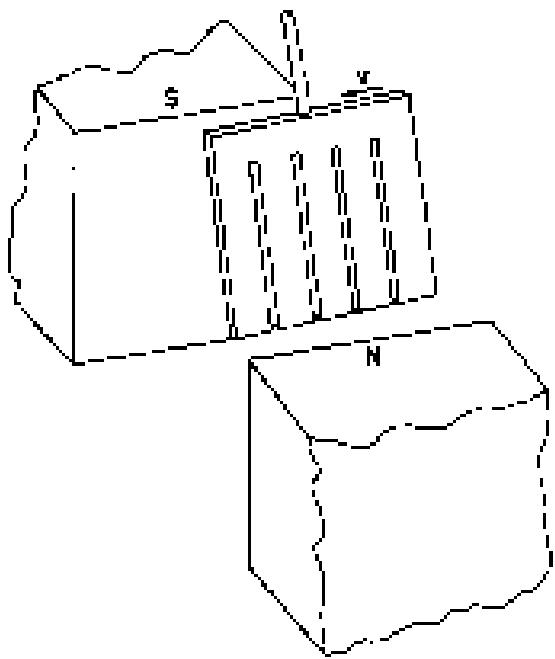


Fig. 16-12. Eddy-current effects are drastically reduced by cutting slots in the plate.

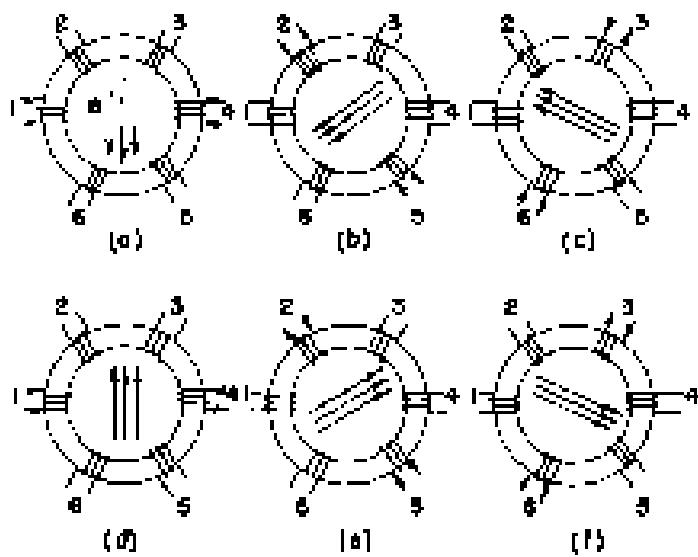


Fig. 16-13. Making a rotating magnetic field.

A field just like that of a rotating magnet can be made with an arrangement of coils such as is shown in Fig. 16-13. We take a torus of iron (that is, a ring of iron like a doughnut) and wind six coils on it. If we put a current, as shown in part (a), through windings (1) and (4), there will be a magnetic field in the direction shown in the figure. If we now switch the current to windings (2) and (5), the magnetic field will be in a new direction, as shown in part (b) of the figure. Continuing the process, we get the sequence of fields shown in the rest of the figure. If the process is done smoothly, we have a "rotating" magnetic field. We can easily get the required sequence of currents by connecting the ends to a three-phase power line, which provides just such a sequence of currents. "Three-phase power" is made in a generator using the principle of Fig. 16-1, except that there are three loops fastened together on the same shaft in a symmetrical way—that is, with an angle of 120° from one loop to the next. When the coils are rotated as a unit, the emf is a maximum in one, then in the next, and so on in a regular sequence. There are many practical advantages of three-phase power. One of them is the possibility of making a rotating magnetic field. The torque produced on a conductor by such a rotating field is easily shown by standing a metal ring on an insulating table just above the torus, as shown in Fig. 16-14. The rotating field causes the ring to spin about a vertical axis. The basic elements seen here are quite the same as those at play in a large commercial three-phase induction motor.

Another form of induction motor is shown in Fig. 16-15. The arrangement shown is not suitable for a practical high-efficiency motor but will illustrate the principle. The electromagnet M , consisting of a bundle of laminated iron sheets wound with a solenoidal coil, is powered with alternating current from a generator. The magnet produces a varying flux of Φ through the aluminum disc. If we have just these two components, as shown in part (a) of the figure, we do not yet have a motor. There are eddy currents in the disc, but they are symmetric and there is no torque. (There will be some heating of the disc due to the induced currents.) If we now cover only one-half of the magnet pole with an aluminum plate, as shown in part (b) of the figure, the disc begins to rotate, and we have a motor. The operation depends on two eddy-current effects. First, the eddy currents in the aluminum plate oppose the changes of flux through it, so the magnetic field above the plate always lags the field above last half of the pole, which is not covered. This so-called "shaded-pole" effect produces a field which is the "skewed" region versus

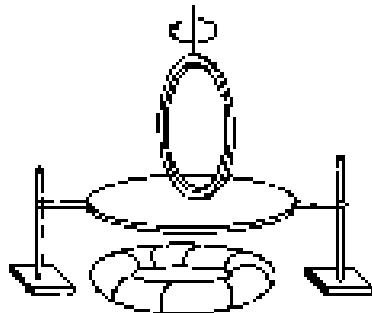


Fig. 16-14. The rotating field of Fig. 16-13 can be used to provide torque on a conducting ring.

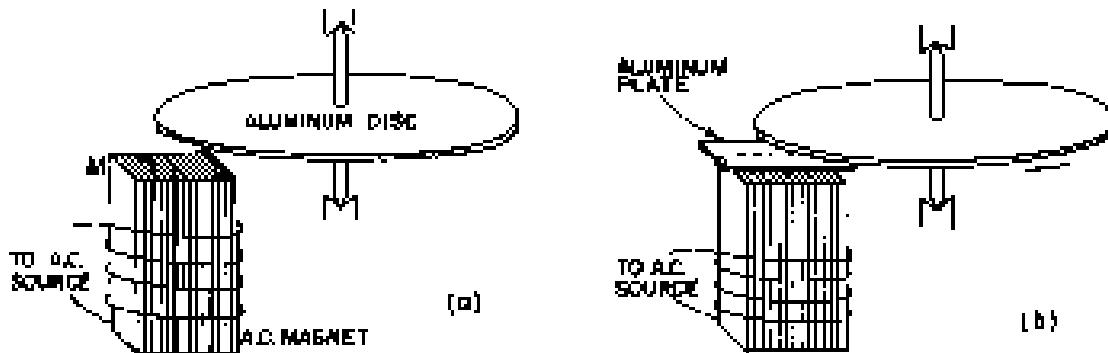


Fig. 16-15. A simple example of a shaded-pole induction motor.

much like that in the "unshaded" region except that it is delayed a constant amount in time. The result is as if there were a magnet only half as wide which is continually being moved from the unshaded region toward the shaded one. Then the varying fields interact with the eddy currents in the disc to produce the torque on it.

16-4 Electrical technology

When Faraday first made public his remarkable discovery that a changing magnetic flux produces an emf, he was asked (as anyone is asked when he discovers a new fact of nature), "What is the use of it?" All he had found was the oddity that a tiny current was produced when he moved a wire near a magnet. Of what possible "use" could that be? His answer was: "What is the use of a newborn baby?"

Yet think of the tremendous practical applications his discovery has led to. What we have been describing are not just rays but complex changes in great cases to represent the principle of some practical machine. For instance, the rotating ring in the turning field is an induction motor. There are, of course, some differences between it and a practical induction motor. The ring has a very small torque; it can be stopped with your hand. For a good motor, things have to be put together more intimately: there shouldn't be so much "wasted" magnetic field out in the air. First, the field is concentrated by using iron. We have not discussed how iron does that, but iron can make the magnetic field tens of thousands of times stronger than copper coils alone could do. Second, the gaps between the pieces of iron are made small; to do that, some iron is even built into the rotating ring. Everything is arranged so as to get the greatest forces and the greatest efficiency—that is, conversion of electrical power to mechanical power until the "ring" can no longer be held still by your hand.

This problem of closing the gaps and making the thing work in the most practical way is engineering. It requires serious study of design problems, although there are no new basic principles from which the forces are obtained. But there is a long way to go from the basic principles to a practical and economic design. Yet it is just such careful engineering design that has made possible such a tremendous thing as Boulder Dam and all that goes with it.

What is Boulder Dam? A huge river is stopped by a concrete wall. But what a wall it is! Shaped with a perfect curve that is very carefully worked out so that the least possible amount of concrete will hold back a whole river. It thickens at the bottom in that wonderful shape that the artists like but that the engineers can appreciate because they know that such thickening is related to the increase of pressure with the depth of the water. But we are getting away from electricity.

Then the water of the river is diverted into a huge pipe. That's a nice engineering accomplishment in itself. The pipe feeds the water into a "waterwheel"—a huge turbine—and makes wheels turn. (Another engineering feat.) But why turn wheels? They are coupled to an exquisitely intricate mess of copper and iron, all

twisted and intertwined. With two parts—one that has and one that doesn't. All a complex intermixture of a few iron and copper but also some paper and shellac for insulation. A revolving magnet thing. A generator. Somewhere out of the mass of copper and iron comes a few special pieces of copper. The dam, the turbine, the iron, the copper, all put there to make something special happen to a few bars of copper—an emf. Then the copper bars go a little way and circle for several times around another piece of iron in a transformer; then their job is done.

But around that same piece of iron coils another cable of copper which has no direct connection whatsoever to the bars from the generator; they have just been influenced because they passed near it to get their emf. The transformer converts the power from the relatively low voltages required for the efficient design of the generator to the very high voltages that are best for efficient transmission of electrical energy over long cables.

And everything must be enormously efficient—there can be no waste, no loss. Why? The power for a metropolis is going through. If a small fraction were lost—one or two percent—think of the energy left behind! If one percent of the power were lost in the transformer, that energy would need to be taken out somewhere. If it appeared as heat, it would quickly melt the whole thing. There is, of course, some small inefficiency, but all that is required are a few pumps which circulate some oil through a radiator to keep the transformer from heating up.

Out of the Boulder Dam come a few dozen rods of copper—long, long, long rods of copper perhaps the thickness of your wrist that go for hundreds of miles in all directions. Small rods of copper carrying the power of a giant river. Then the rods are split to make more rods . . . then to more transformers . . . sometimes to great generators which receive the current in another form . . . sufficient to generate running for big industrial purposes . . . in more transformers . . . then more splitting and spreading . . . until finally the river is spread throughout the whole city—running motors, making heat, making light, working industry. The cascade of hot little jets cold water over 600 miles away—all done with specially arranged pieces of copper and iron. Large motors for rolling steel, or tiny motors for a dentist's drill. Thousands of little wheels, turning in response to the turning of the big wheel at Boulder Dam. Stop the big wheel, and all the wheels stop; the lights go out. They really are connected.

Yet there is more. The same phenomena that take the tremendous power of the river and spread it through the countryside, until a few drops of the water run into the dentist's drill, come again into the building of extremely fine mechanisms . . . for the detection of incredibly small amounts of current . . . for the transmission of voices, music, and pictures . . . for computers . . . for automatic machines of fantastic precision.

All this is possible because of carefully designed arrangements of copper and iron—efficiently oriented magnetic fields . . . blocks of soft iron six feet in diameter whirling with clearances of 1/16 of an inch . . . careful proportions of copper for the optimum efficiency . . . strange shapes all serving a purpose, like the curve of the dam.

If some future archaeologist unearths Boulder Dam, we may guess that he would admire the beauty of its curves. But also the explorers from some great future civilization will look at the generators and transformers and say: "Notice that every iron piece has a beautifully efficient shape. Think of the thought that has gone into every piece of copper!"

This is the power of engineering and the careful design of our electrical technology. There has been created in the generator something which exists nowhere else in nature. It is true that there are forms of induction in other places. Certainly in some places around the sun and stars there are effects of electromagnetic induction. Perhaps also (though it's not certain) the magnetic field of the earth is maintained by an analog of an electric generator that operates on circulating currents in the interior of the earth. But nowhere have there been pieces put together with moving parts to generate electrical power as is done in the generator—with great efficiency and regularity.

You may think that designing electric generators is no longer an interesting subject, that it is a dead subject because they are all designed. Almost perfect generators or rotors can be taken from a shell. Even if this were true, we can admire the wonderful accomplishment of a problem solved to near perfection. But there remain as many unfinished problems. Even generators and transformers are not living as problems. It is likely that the whole field of low temperatures and superconductors will soon be applied to the problem of electric power distribution. With a radically new factor in the problem, new optimum designs will have to be evolved. Power networks of the future may have little resemblance to those of today.

You can see that there is an endless number of applications and problems that one could take up while studying the laws of induction. The study of the design of electrical machinery is a big work in itself. We cannot go very far in that direction, but we should be aware of the fact that when we have discovered the law of induction, we have suddenly connected our theory to an enormous practical development. We must, however, leave that subject to the engineers and applied scientists who are interested in working out the details of particular applications. Physics only supplies the base—the basic principles don't apply, so to speak what. (We have not yet completed the base, because we have yet to consider in detail the properties of areas of copper. Physics has something to say about these as we will see a little later.)

Modern electrical technology began with Faraday's discoveries. The useless hobby developed into a prodigy and changed the face of the earth in ways no man farther could have imagined.

The Laws of Induction

17-1 The physics of induction

In the last chapter we described many phenomena which show that the effects of induction are quite complicated and interesting. Now we want to discuss the fundamental principles which govern these effects. We have already defined the emf in a conducting circuit as the total accumulated force on the charges throughout the length of the loop. More specifically, it is the tangential component of the force per unit charge, integrated along the wire once around the circuit. This quantity is equal, therefore, to the total work done on a single charge that travels once around the circuit.

We have also given the "flux rule," which says that the emf is equal to the rate at which the magnetic flux through such a conducting circuit is changing. Let's see if we can understand why that might be. First, we'll consider a case in which the flux changes because a circuit is moved in a steady field.

In Fig. 17-1 we show a simple loop of wire whose dimensions can be changed. The loop has two parts, a fixed U-shaped part (a) and a movable crossbar (b) that can slide along the two legs of the U. There is always a complete circuit, but its area is variable. Suppose we now place the loop in a uniform magnetic field with the plane of the U perpendicular to the field. According to the rule, when the crossbar is moved there should be in the loop an emf that is proportional to the rate of change of the flux through the loop. This emf will cause a current in the loop. We will assume that there is enough resistance in the wire that the currents are small. Then we can neglect any magnetic field from this current.

The flux through the loop is Φ_B , so the "flux rule" would give for the emf—which we write as ϵ

$$\epsilon = -\Phi_B \frac{dI}{dt} = -\Phi_B v,$$

where v is the speed of translation of the crossbar.

Now we should be able to understand this result from the magnetic $v \times B$ forces on the charges in the moving crossbar. These charges will feel a force, tangential to the wire, equal to vB per unit charge. To is constant along the length a of the crossbar and zero elsewhere, so the integral is

$$\epsilon = -vAB,$$

which is the same result we got from the rate of change of the flux.

The argument just given can be extended to any case where there is a fixed magnetic field and the wires are moved. One can prove, in general, that for any circuit whose parts move in a fixed magnetic field the emf is the time derivative of the flux, regardless of the shape of the circuit.

On the other hand, what happens if the loop is stationary and the magnetic field is changed? We cannot deduce the answer to this question from the same argument. It was Faraday's discovery, from experiment—that the "flux rule" is still correct no matter why the flux changes. The force on electric charges is given in complete generality by $F = q(E + v \times B)$; there are no new special "forces due to changing magnetic fields." Any forces on charges at rest in a stationary wire come from the E term. Faraday's observations led to the discovery that electric and magnetic fields are related by a new law: in a region where the magnetic field is changing with time, electric fields are generated. It is this electric

17-1 The physics of induction

17-2 Decreasing in the "flux rule"

17-3 Particle acceleration by an induced electric field; the horizon

17-4 A paradox

17-5 Alternating-current generator

17-6 Mutual inductance

17-7 Self-inductance

17-8 Inductance and magnetic energy



Fig. 17-1. An emf ϵ is induced in a loop if the flux is changed by varying the area of the circuit.

field which drives the electrons around the wire, and so is responsible for the emf in a static circuit when there is a changing magnetic flux.

The general law for the electric field associated with a changing magnetic field is

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \quad (17.1)$$

We will call this Faraday's law. It was discovered by Faraday but was first written in differential form by Maxwell, as one of his equations. Let's see how this equation gives the "flux rule" for circuits.

Using Stokes' theorem, this law can be written in integral form as

$$\oint_C \mathbf{E} \cdot d\mathbf{s} = \int_S (\mathbf{E} \times \mathbf{B}) \cdot n \, dA = - \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot n \, dA, \quad (17.2)$$

where, as usual, C is any closed curve and S is any surface bounded by C . Here, remember, C is a *conventional* curve fixed in space and S is a flat surface. Right, the time derivative can be taken outside the integral and we have

$$\begin{aligned} \oint_C \mathbf{E} \cdot d\mathbf{s} &= - \frac{\partial}{\partial t} \int_S \mathbf{B} \cdot n \, dA \\ &= - \frac{\partial}{\partial t} (\text{Flux through } S). \end{aligned} \quad (17.3)$$

Applying this relation to a curve C that follows a fixed circuit of conductor, we get the "flux rule" once again. The integral on the left is the emf, and that on the right is the negative rate of change of the flux linked by the circuit C . So Eq. (17.1) applied to a fixed circuit is equivalent to the "flux rule."

So the "flux rule" that the emf in a circuit is equal to the rate of change of the magnetic flux through the circuit applies whether the flux changes because the field changes or because the circuit moves through it. The two possible lines "current changes" or "local charges" are not distinguishable in the statement of the rule. Yet again, application of the rule we have used two completely distinct laws for the two cases: $\nabla \times \mathbf{B}$ for "current changes" and $\nabla \times \mathbf{E} = -\partial \mathbf{B} / \partial t$ for "field changes."

We know of no other place in physics where such a simple and elegant general principle requires for its real understanding an analysis in terms of two different parameters. (Similarly such a beautiful generalization is found to stem from using a deeper underlying principle. Nevertheless, in this case there does not appear to be any such extended interpretation.) We have to understand the "rule" as the combination of two quite separate phenomena.

We must look at the "flux rule" in the following way. In general, the force on unit charge is $F = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$. In moving waves there is the force from the second term. Also, there is an E -field if there is somewhere a changing magnetic field. They are independent effects, but the end result for loops of wire is always equal to the sum of currents times magnetic flux through it.

17.2 Exceptions to the "flux rule"

We will now give some examples, due in part to Faraday, which show the importance of keeping clearly in mind the distinction between the two effects responsible for induced emfs. Our examples involve situations to which the "flux rule" cannot be applied, either because there is no wire at all or because the path taken by induced currents moves about within an extended volume of a conductor.

We begin by making an important point: The sum of currents that comes from the B -field does not depend on the existence of a physical wire (as does the $\mathbf{E} \times \mathbf{B}$ part). The A -field can exist in free space, and its line integral around any arbitrary line fixed in space is the rate of change of the flux of B through that line. (Note that this is quite unlike the A -field produced by static charges, for in that case the line integral of B around a closed loop is always zero.)

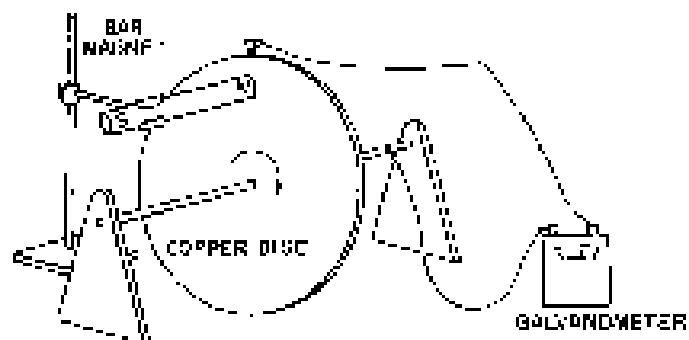


Fig. 17-2. When the disc rotates there is an emf, $\nabla \times B$, but with no change in the linked flux.

Now we will describe a situation in which the flux through a circuit does not change, but there is nevertheless an emf. Figure 17-2 shows a conducting disc which can be rotated on a fixed axis in the presence of a magnetic field. One contact is made to the shaft and another rests on the outer periphery of the disc. A circuit is completed through a galvanometer. As the disc rotates, the "current," in the sense of the place or source where the current starts, is always the same. But the part of the "circuit" in this case is an material which is moving. Although the joint through the "circuit" is constant, there is still an emf, as can be observed by the deflection of the galvanometer. Clearly, here is a case where $\nabla \times B$ force in the moving disc gives rise to an emf which cannot be equated to a change of flux.

Now we consider, as an opposite example, a somewhat unusual situation in which the flux through a "circuit" (even in the sense of the place where the current is) changes but where there is no emf. Imagine two metal plates with slightly curved edges, as shown in Fig. 17-3, placed in a uniform magnetic field perpendicular to their surfaces. Each plate is connected to one of the terminals of a galvanometer, as shown. The plates make contact at one point P , so there is a complete circuit if the plates are now rocked through a small angle; the point of contact will move to P' . If we now give the "circuit" to be completed (around the plates on the dotted line shown in the figure), the magnetic flux through this circuit changes by a large amount as the plates are rocked back and forth. Yet the reading can be done with small meters, so that $\nabla \times B$ is very small and there is practically no emf. The "flux rule" does not work in this case. It must be applied to circuits in which the material of the circuit remains the same. When the material of the circuit is changing, we must return to the basic law. The correct physics is always given by the flux rule!

$$F = \mu(E + v \times B).$$

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

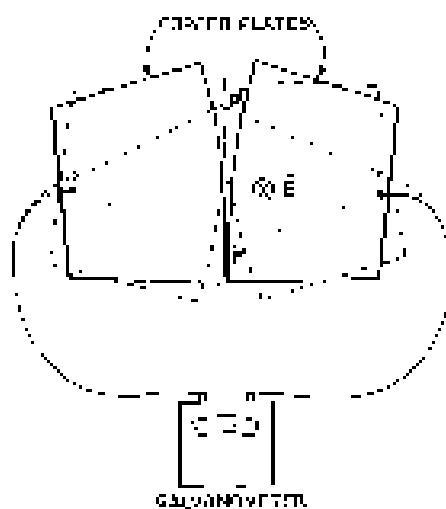


Fig. 17-3. When the plates are rocked in a uniform magnetic field, there can be a large change in the flux linkage without the generation of an emf.

17-3 Particle acceleration by an induced electric field; the betatron

We have seen that the electromagnetic fields produced by a changing magnetic field can exist even with no conductors. Just as $\nabla \times B$ can apply to an empty void with no wires, we may still have an electro-magnetic field existing in a "vacuum" (mathematically empty space). It is believed by just an educated guess that F integrated around the curve, Faraday's law says that the line integral is equal to the rate of change of the magnetic flux through the closed curve, Eq. (17-3)

As an example of the effect of such an unshielded field, we want to try to consider the motion of an electron going through a magnetic field. We imagine a magnetic field which, say, in a plane, points in a vertical direction, as shown in Fig. 17-4. The magnetic field is produced by a solenoid, etc., but we will not worry about the details. For our example we will assume that the magnetic field is uniform about some axis, i.e., that the strength of the magnetic field B depends only on the distance from the axis. The magnetic field is also varying with time. We can imagine an electron that is moving in the field on a path that is a circle of constant radius with the center at the axis of the field. (We will see later

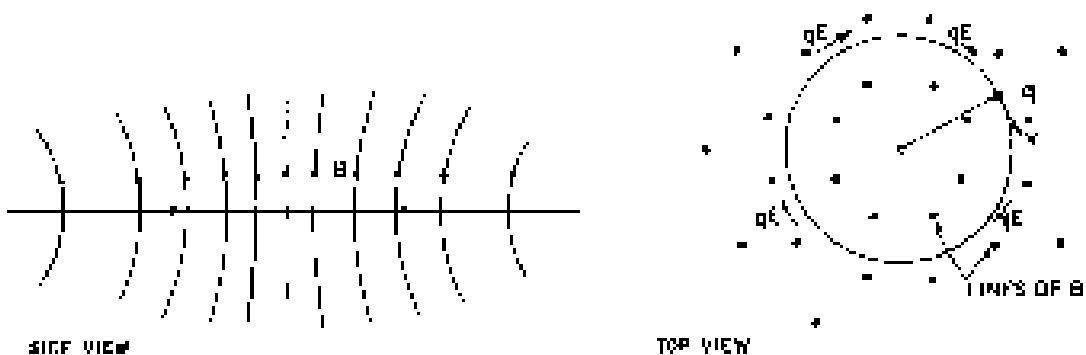


Fig. 17-4. An electron accelerating in an axially symmetric, time-varying magnetic field.

now this motion and be accelerated. Because of the changing magnetic field, there will be no electric field E tangential to the electron's orbit which will cause it to leave the circle. Because of the symmetry, this electric field will have the same value everywhere on the circle. If the electron's orbit has the radius r , the line integral of E around the orbit is equal to the rate of change of the magnetic flux through the circle. The line integral of E is just the component tangential to the electron's orbit of the electric field E . The magnetic flux must, in general, be obtained from an integral. For the moment, we let B_{av} represent the average magnetic field in the interior of the circle; then the flux is this average magnetic field times the area of the circle. We will have

$$2\pi r E = \frac{d}{dt} (B_{av} \cdot \pi r^2)$$

Since we are assuming r is constant, E is proportional to the time derivative of the average field:

$$E = \frac{1}{2} \frac{dB_{av}}{dt} \quad (17.3)$$

The electron will feel the electric force qE and will be accelerated by it. Remembering that the relativistically correct equation of motion is that the rate of change of the momentum is proportional to the force, we have

$$qE = \frac{dp}{dt} \quad (17.4)$$

For the circular orbit we have noticed, the electric force on the electron is always in the direction of its motion, so its total momentum will be increasing at the rate given by Eq. (17.4). Combining Eqs. (17.3) and (17.4), we may relate the rate of change of momentum to the change of the average magnetic field:

$$\frac{dp}{dt} = \frac{q}{2} \frac{d\theta_{av}}{dt} \quad (17.5)$$

Integrating with respect to t , we find for the electron's momentum

$$p = p_0 + \frac{q}{2} \Delta \theta_{av} \quad (17.6)$$

where p_0 is the momentum with which the electrons start out, and $\Delta \theta_{av}$ is the subsequent change in θ_{av} . The operation of a betatron - a machine for accelerating electrons to high energies - is based on this idea.

To see how the betatron operates in detail, we must now examine how the electron can be constrained to move on a circle. We have discussed in Chapter 15 of Vol. I the principle involved. If we arrange that there is a magnetic field B at the orbit of the electron, there will be a transverse force $qv \times B$ which, for a suitable

as by choice, B , can cause the electron to keep moving on its assigned orbit. In the derivation thus far, no force causes the electron to move in a circular orbit of constant radius. We can find out what the magnetic field at the orbit must be by using again the relativistic equation of motion, but this time for the transverse component of the force. In the betatron case (Fig. 17-4), B is straight, angular at $\pi/2$, so the force is $qvB \sin \theta = qvB$. Thus the force is equal to the rate of change of the transverse component p_θ of the momentum:

$$qvB = \frac{dp_\theta}{dt}. \quad (17.2)$$

When a particle e is moving in a circle, the rate of change of its transverse momentum is equal to the magnitude of the tangential velocity v , the angular velocity of rotation (following the arguments of Chapter 11, Vol. I),

$$\frac{dp_\theta}{dt} = vp_\theta, \quad (17.3)$$

where, since the motion is circular,

$$\omega = \frac{v}{r}. \quad (17.4)$$

Setting the magnetic force equal to the transverse acceleration, we have

$$qvB_{\text{ext}} = p_\theta \omega, \quad (17.5)$$

where B_{ext} is the field at the radius r .

As the motion increases, the momentum of the electron grows in proportion to B_{ext} , according to Eq. (17.3), and if the electron is to continue to move in its proper circle, Eq. (17.5) must continue to hold as the momentum of the electron increases. The value of B_{ext} must increase in proportion to the momentum p . Comparing Eq. (17.5) with Eq. (17.2), which determines p , we see that the following relation must hold between B_{ext} , the average magnetic field inside the orbit at the radius r , and the magnetic field B_{ext} at the orbit:

$$AB_{\text{ext}} = 3.2B_{\text{ext}}. \quad (17.6)$$

The correct operation of a betatron requires that the average magnetic field inside the orbit increase at twice the rate of the magnetic field at the orbit itself. In these circumstances, as the energy of the particle is increased by the induced electric field the magnetic field at the orbit increases at just the rate required to keep the particle moving in a circle.

The betatron is used to accelerate electrons to energies of tens of millions of volts, or even to hundreds of millions of volts. However, it becomes impractical for the acceleration of electrons to energies much higher than a few hundred million volts for several reasons. One of them is the practical difficulty of obtaining the required high average value for the magnetic field inside the orbit. Another is that Eq. (17.6) is no longer correct at very high energies because it does not include the loss of energy from the particle due to its radiation of electromagnetic energy (the so-called synchrotron radiation discussed in Chapter 36, Vol. I). For these reasons, the acceleration of electrons to the highest energies—“many billions of” electron volts—is accomplished by means of a different kind of machine, called a *synchrotron*.

17-4 A paradox

We would now like to describe for you an apparent paradox. A paradox is a situation which gives one answer when analyzed one way, and a different answer when analyzed another way, so that we are left in somewhat of a quandary as to actually what would happen. Of course, in physics there are never any real paradoxes because there is only one correct answer, at least we believe that now we will

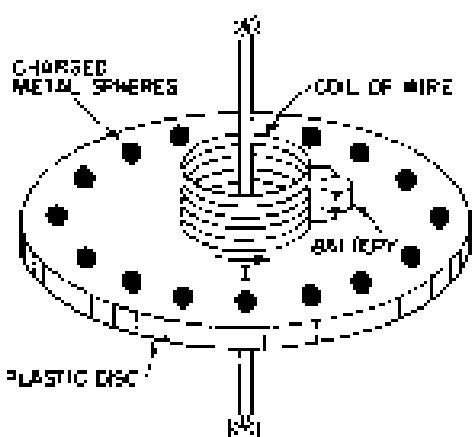


Fig. 17-5. Will the disc rotate if the current I is stopped?

in only one way (and that is the right way, naturally). So in physics it is always better to confine our own understanding. Here's our paradox.

Imagine that we construct a device like that shown in Fig. 17-5. There is a thin, circular plastic disc suspended on a vertical shaft with excellent bearings, so that it is quite free to rotate. In the center is a coil of wire in the form of a short solenoid concentric with the axis of rotation. This coil carries a steady current I provided by a small battery, as shown. Next, the edge of the disc and center (in fact, a small bar magnet) attract each other since all metal spheres scattered on a mesh grid just below the bottom of the disc. Each of these small conducting spheres is charged with the same negative charge Q . Everything is up to snuffing, and the disc is at rest. Suppose now that by some accident, or by pure happenstance, the current in the solenoid is interrupted, whether by intervention from the outside. Suddenly, when current ceases, there was no magnetic field through the solenoid, more or less parallel to the axis of rotation. When the current is interrupted, this has nothing to do. There will, therefore, be no electric field induced which will cause motion in either direction of the axis. The charged spheres at the perimeter of the disc will all experience an electric field tangential to the perimeter of the disc. This electric force is in the same sense toward the charges and so will result in a net torque on the disc. From these arguments we would expect that as the current in the solenoid ceases, the disc would begin to rotate. If we knew the magnitude of gravity (that is, the current in the solenoid, and the charges on, I usually think, we could not give the resulting angular velocity).

But we could also make a different argument. Using the principles of thermodynamics of a great civilization, we could say that the total moment of all the disc with all its equipment is nearly zero, and so the only disturbance in the system should stop it at once. We should be in rotation when the current is stopped. Which argument is correct? Well, the disc rotates or will it not? We will leave this question for you to think about.

We should warn you that the correct answer does not depend on any misconception of mine, such as the asymmetric position of a battery, for example. In fact, you can imagine several situations such as the following. The solenoid is made of a permanent magnet through which there is a current. After the current has been fully passed off, the temperature of the solenoid is allowed to rise slowly while the ferromagnetic of the core reaches the transition temperature between superconductivity and resistivity. Immediately, the current in the solenoid will be brought to zero by the resistance of the core. Finally, as before, let us generate there with an electric field through the axis. We should also warn you that the situation is not easy, or a simple trick. When you figure it out, you will have the crown of a laurel in the annals of electromagnetism.

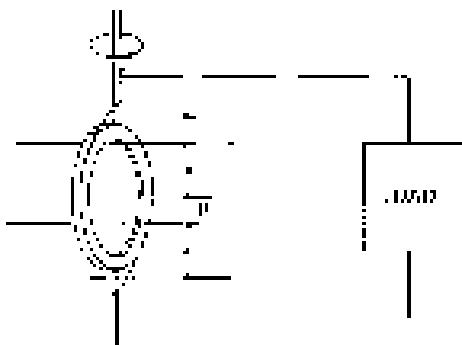


Fig. 17-6. A coil of wire rotating in a uniform magnetic field—the basic idea of the *dc* generator.

17-5 Alternating-current generator

In the remainder of this chapter we apply the principles of Section 17-4 to analyze numerous of the phenomena discussed in Chap. 16. We first look in more detail at the alternating-current generator. Such a generator consists basically of a coil of wire rotating in a uniform magnetic field. The same as drawn may be achieved by a fixed coil in a magnetic field whose direction rotates at the frequency described in the last chapter. We will consider only the latter case. Suppose we have a circular coil of wire which can be turned out at a constant rate of rotation. Let this coil be located in a uniform magnetic field perpendicular to the plane of rotation, as in Fig. 17-6. We also imagine that the two ends of the coil are brought to external connections through brushes kind of guy wires.

Due to the rotation of the coil, the magnetic flux through it will be changing. The current of the coil will therefore have an effect. Let θ be the angle of the coil and δ the angle between the magnetic field and the normal to the plane of the coil.⁴

⁴ Note that we are using the letter δ for the *W* in $\sin \delta$ instead of ϕ , θ , and ψ for a surface area.

The flux through the coil is then

$$\text{flux} = \int B \cdot d\mathbf{A} \quad (17.13)$$

If the coil is rotating at the uniform angular velocity ω , θ varies with time as $\theta = \omega t$. The curl \mathbf{E} in the coil is then

$$\begin{aligned} \mathbf{E} &= -\frac{d}{dt} (\text{flux}) = -\frac{d}{dt} (BS \cos \omega t) \\ &= BS \omega \sin \omega t. \end{aligned} \quad (17.14)$$

If we bring the wires from the generator to a point some distance from the rotating coil, where the magnetic field is zero, or at least is not varying with time, the curl of \mathbf{E} in this region will be zero and we can define an electric potential. In fact, if there is no current being drawn from the generator, the potential difference V between the two wires will be equal to the emf in the rotating coil. That is,

$$V = BS \omega \sin \omega t = E_0 \sin \omega t.$$

The potential difference between the wires varies sinusoidally. Such a varying potential difference is called an alternating voltage.

Since there is an electric field between the wires, they must be electrically charged. It is clear that the emf of the generator has passed some excess charge out of the wire until the electric field from E_0 is strong enough to exactly completely balance the induction force. Since far outside the generator, the two wires appear as though they had been electrically charged to the potential difference V , but as though the charge was being carried with time to give an alternating longitudinal voltage. There is also mutual induction between the two wires. If we require the generator to generate a constant potential difference V , we find that the self-losses are great if the wires try to be discharged but continue to provide charge to the wires as current is drawn from them, attempting to keep the wires always at the same potential difference. If, in fact, the generator is connected in a circuit whose total resistance is R , the current through the circuit will be proportional to the emf of the generator and inversely proportional to R . So the emf has sinusoidal time variation, so also does the current. There is an alternating current

$$I = \frac{E}{R} = \frac{V_0}{R} \sin \omega t.$$

The schematic diagram of such a circuit is shown in Fig. 17-7.

We can also see just the end determines how much energy is supplied by the generator. Each charge in the wire is receiving energy at the rate $F \cdot v$, where F is the force on the charge and v is its velocity. Now let q represent all moving charges per unit length of the wire be w . Then the power being delivered into any element ds of the wire is

$$F \cdot v \cdot ds.$$

For a wire, v is always along ds , so we can rewrite the power as

$$q e F \cdot ds.$$

The total power being delivered to the complete circuit is the integral of this expression around the complete loop:

$$\text{Power} = \oint q e F \cdot ds. \quad (17.15)$$

Now remember that $q ds$ is the current I , and that the emf is defined as the integral of $E \cdot ds$ around the circuit. We get the result:

$$\text{Power from a generator} = VI \quad (17.16)$$

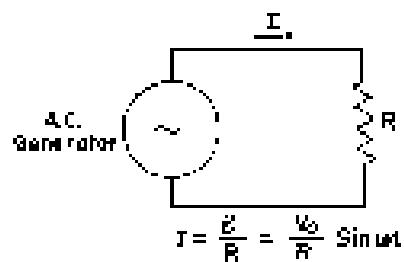


Fig. 17-7. A circuit with an ac generator and a resistor.

When there is a current in the coil of the generator, there will also be mechanical forces on it. In fact, we know that the torque on the coil is proportional to its magnetic moment, to the magnetic field strength B , and to the sine of the angle between. The magnetic moment is the current in the coil times its area. Therefore the torque is

$$\tau = 130 \sin \theta. \quad (17.17)$$

The rate at which mechanical work must be done to keep the coil at θ is the angular velocity ω times the torque

$$\frac{dE}{dt} = \omega \tau = \omega B D \sin \theta. \quad (17.18)$$

Comparing this equation with eq. (17.14), we see that the rate of mechanical work required to move the coil against the magnetic forces is just equal to E_t , the rate at which electrical energy is delivered by the emf of the generator. All of the mechanical energy used up in the generator appears as electrical energy in the circuit.

As a second example of the currents and forces due to an induced emf, let's analyze what happens in the set-up described in Section 12, and shown in Fig. 17.1. There are two parallel wires and a sliding crossbar located in a uniform magnetic field perpendicular to the plane of the parallel wires. Now let's assume that the "bottom" of the L (the left side in the figure) is made of wires of high resistance, while the "two side wires" are made of a good conductor like copper. Then we don't need to worry about the change of the circuit resistance as the crossbar is moved. As before, the emf in the circuit is

$$e = vBw. \quad (17.19)$$

The current in the circuit is proportional to this emf and inversely proportional to the resistance of the circuit:

$$I = \frac{e}{R} = \frac{vBw}{R}. \quad (17.20)$$

Because of this current there will be a magnetic force on the crossbar that is proportional to its length, to the current in it, and to the magnetic field, such that

$$F = BIp. \quad (17.21)$$

Taking s from eq. (17.06), we have for the force

$$F = \frac{B^2 w^2}{R} s. \quad (17.22)$$

We see that the force is proportional to the velocity of the crossbar. The direction of the force, as you can easily see, is opposite to its velocity. Such a "velocity-dependent" force, which is like the force of viscosity, is found whenever induced currents are produced by moving conductors in a magnetic field. The examples of eddy currents we gave in the last chapter also produced forces on the conductors proportional to the velocity of the conduction, even though such situations are general, give a complicated distribution of currents which is difficult to analyze.

It is often convenient in the design of mechanical systems to have damping forces which are proportional to the velocity. Eddy-current forces permit one of the most convenient ways of getting such a velocity-dependent force. An example of the application of such a force is found in the conventional domestic wattmeter. In the wattmeter there is a thin aluminum disc that rotates between the poles of a permanent magnet. This disc is driven by a small electric motor whose torque is proportional to the power being consumed in the electrical circuit, of the house. Because of the eddy-current forces in the disc, there is a resistive force proportional to the velocity. In equilibrium, the velocity is therefore proportional to the rate of consumption of electrical energy. By means of a counter attached to the rotating disc, a record is kept of the number of revolutions it makes. This counts is an indication of the total energy consumption, i.e., the number of "watt-hours" used.

We may also point out that Eq. (17.22) shows that the force from induced emf—that is, the eddy-current force—is inversely proportional to the resistance. The force will be larger the better the conductivity of the material. The series, of course, is that an emf produces more current if the resistance is low, and the stronger current represents greater mechanical forces.

We can also see from our formulas how mechanical energy is converted into electrical energy. As before, the electrical energy supplied to the resistance of the circuit is the product IR . The rate at which work is done in moving the conducting mass is the force on the bar times its velocity. Using Eq. (17.21) for the force, the rate of doing work is

$$\frac{dW}{dt} = \frac{\epsilon^2 R^2 v^2}{R}.$$

We see that this is indeed equal to the product of the work we would get from Eqs. (17.19) and (17.20). Again the mechanical work appears as electrical energy.

17-6 Mutual Inductance

We now want to consider a situation in which there are fixed coils of wire bar changing magnetic fields. When we described the production of magnetic fields by currents, we considered only the case of steady currents. But so long as the currents are changed slowly, the magnetic field will in each inductor be nearly the same as the magnetic field of a steady current. We will assume in the discussion of this section that the currents are always varying sufficiently slowly that this is true.

In FIG. 17.8 is shown an arrangement of two coils which demonstrates the basic effects responsible for the operation of a transformer. Coil 1 consists of a conductive wire wound in the form of a long solenoid. Around this coil 1 and insulated from it—is wound coil 2, consisting of a few turns of wire. If now a current is passed through coil 1, we know that a magnetic field will appear inside it. This magnetic field also passes through coil 2. As the current in coil 1 is varied, the magnetic flux will also vary, and there will be an induced emf in coil 2. We will now calculate this induced emf.

We have seen in Section 17-5 that the magnetic field inside a long solenoid is uniform and has the magnitude

$$B = \frac{1}{\epsilon_0 c^2} N_1 I_1, \quad (17.21)$$

where N_1 is the number of turns in coil 1, I_1 is the current through it, and c is its length. Let's say that the cross-sectional area of coil 1 is A_1 . Then the flux of B is its magnitude times A_1 . If coil 2 has N_2 turns, this flux links the coil N_2 times. Therefore the emf in coil 2 is given by

$$e_2 = -N_2 \frac{d\Phi}{dt}. \quad (17.22)$$

The only quantity in Eq. (17.22) which varies with time is I_1 . The emf is therefore given by

$$e_2 = -\frac{N_1 N_2 A_1}{\epsilon_0 c^2} \frac{dI_1}{dt}. \quad (17.23)$$

We see that the emf in coil 2 is proportional to the rate of change of the current in coil 1. The constant of proportionality, which is basically a measure factor of the two coils, is called the mutual inductance, and is usually designated M_{12} . Equation (17.23) is then written

$$e_2 = M_{12} \frac{dI_1}{dt}. \quad (17.24)$$

Suppose now that we were to pass a current through coil 2 and ask about the emf in coil 1. We would compute the magnetic field, which is everywhere

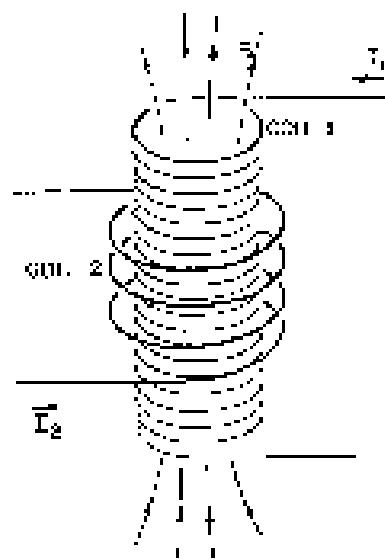


Fig. 17.8. A current in coil 1 produces a magnetic field through coil 2.

proportional to the current I_2 . The flux linkage through coil 1 would depend on the geometry, but would be proportional to the current I_2 . The emf in coil 1 would, therefore, again be proportional to all I_2 's! We can write

$$\mathcal{E}_1 = M_{12} \frac{dI_2}{dt}. \quad (17.27)$$

The computation of M_{12} would be more difficult than the computation we have just done for M_{21} , but we can do it through their computation now, because we will show later in this chapter that M_{12} is necessarily equal to M_{21} .

Since for any coil its field is proportional to its current, the same kind of result would be obtained for any two coils of wire. The equations (17.26) and (17.27) would have the same form; only the constants M_{21} and M_{12} would be different. Their values would depend on the shapes of the coils and their relative positions.

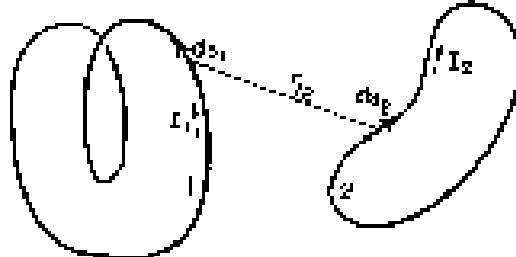


Fig. 17-9. Any two coils have a mutual inductance M_{12} proportional to the integral of $d\phi_1/ds_2 ds_2/ds_1$.

Suppose that we wish to find the mutual inductance between any two arbitrary coils—for example, those shown in Fig. 17-9. We know that the general expression for the emf in coil 1 can be written as

$$\mathcal{E}_1 = - \frac{d}{dt} \int_{C_1} \mathbf{B} \cdot d\mathbf{s}_1,$$

where \mathbf{B} is the magnetic field and the integral is to be taken over a surface bounded by circuit 1. We have seen in Section 14-3 that such a surface integral of \mathbf{B} can be related to a line integral of the vector potential. In particular,

$$\int_{C_1} \mathbf{B} \cdot d\mathbf{s}_1 = \oint_{C_1} \mathbf{A} \cdot d\mathbf{s}_1,$$

where \mathbf{A} represents the vector potential and ds_1 is an element of circuit 1. The last integral is to be taken around closed circuit 1. The emf in coil 1 can therefore be written as

$$\mathcal{E}_1 = - \frac{d}{dt} \oint_{C_1} \mathbf{A} \cdot d\mathbf{s}_1. \quad (17.28)$$

Now let's assume that the vector potential \mathbf{A} of coil 1 comes from currents in circuit 2. Then it can be written as a line integral around circuit 2:

$$\mathbf{A} = \frac{1}{4\pi\mu_0 c^2} \oint_{C_2} \frac{I_2 ds_2}{r_{21}}, \quad (17.29)$$

where I_2 is the current in circuit 2, and r_{21} is the distance from the element of the circuit ds_2 to the point 1 on circuit 1, at which we are evaluating the vector potential. (See Fig. 17-9.) Considering Eqs. (17.28) and (17.29), we can express the emf in circuit 1 as a double line integral:

$$\mathcal{E}_1 = - \frac{1}{4\pi\mu_0 c^2} \frac{d}{dt} \oint_{C_1} \oint_{C_2} \frac{I_2 ds_2}{r_{21}} \frac{\partial \mathbf{A}_2}{\partial s_1} \cdot d\mathbf{s}_1.$$

In this equation the integrals are all taken with respect to stationary circuits. The only variable quantity is the current I_2 , which does not depend on the variables of 17-10.

integration. We may therefore take it out of the integral. The emf can then be written as

$$\mathcal{E}_1 = M_{12} \frac{dI_2}{dt},$$

where the coefficient M_{12} is

$$M_{12} = -\frac{1}{4\pi\epsilon r^2} \oint_{C_1} \oint_{C_2} d\sigma_1 d\sigma_2, \quad (17.30)$$

We see from this integral that M_{12} depends only on the circuit geometry. It depends on a kind of average separation of the two circuits, with the average weighted over parallel segments of the two coils. Our equation can be used for calculating the mutual inductance of any two circuits of arbitrary shape. Also, it shows that the integral for M_{12} is identical to the integral for M_{21} . We have therefore shown that the two coefficients are identical. For a system with only two coils, the coefficients M_{12} and M_{21} are often represented by the symbol M without subscripts, called simply the *mutual inductance*:

$$\mathcal{E}_{12} = \mathcal{E}_{21} = M.$$

17-7 Self-inductance

In discussing the induced electromotive forces in the two coils of figs. 17-8 or 17-9, we have considered only the case in which there was a current in one coil or the other. If there are currents in the two coils simultaneously, the magnetomotive force in either coil will be the sum of the contributions which would exist separately because the law of superposition applies for magnetic fields. The total current in coil 1 will therefore be proportional not only to the change of the current in the other coil, but also to the change in the current of the coil itself. Thus the total emf in coil 2 should be written*

$$\mathcal{E}_2 = M_{22} \frac{df}{dt} + M_{12} \frac{dI_1}{dt}. \quad (17.31)$$

Similarly, the emf in coil 1 will depend not only on the changing current in coil 2, but also on the changing current in itself:

$$\mathcal{E}_1 = M_{12} \frac{dI_2}{dt} + M_{21} \frac{dI_2}{dt}. \quad (17.32)$$

The coefficients M_{22} and M_{11} are always negative numbers. It is usual to write

$$M_{11} = -S_1, \quad M_{22} = -S_2, \quad (17.33)$$

where S_1 and S_2 are called the *self-inductances* of the two coils.

The self-induced emf will, of course, exist even if we have only one coil. Any coil by itself will have a self-inductance S . The emf will be proportional to the rate of change of the current in it. For a single coil, it is usual to adopt the convention that the emf and the current are considered positive if they are in the same direction. With this convention, we may write for the emf of a single coil

$$\mathcal{E} = -S \frac{di}{dt}. \quad (17.34)$$

The negative sign indicates that the emf opposes the change in current. It is often called a "back emf".

Since any coil has a self-inductance which opposes the change in current, the emf in the coil has a kind of inertia. In fact, if we wish to change the current in

* The signs of " M_{11} " and " M_{22} " in Eqs. (17.31) and (17.32) depends on the arbitrary choices for the sense of a positive current in the two coils.

in coil we must overcome the inertia of connecting the coil to some external voltage source such as a battery or a generator, as shown in the schematic diagram of Fig. 17-10(a). In such a circuit, the current I depends on the voltage V according to the relation

$$V = L \frac{dI}{dt}. \quad (17.18)$$

This equation has the same form as Newton's law of motion for a particle in one dimension. We can therefore study it by the principle that "the same quantities have the same solutions." Thus, if we make the externally applied voltage V correspond to an externally applied force F , and the current I to a coil current i , to the velocity v of a particle, the inductance L of the coil corresponds to the mass m of the particle.* See Fig. 17-10(b). We can make the following table of corresponding quantities.

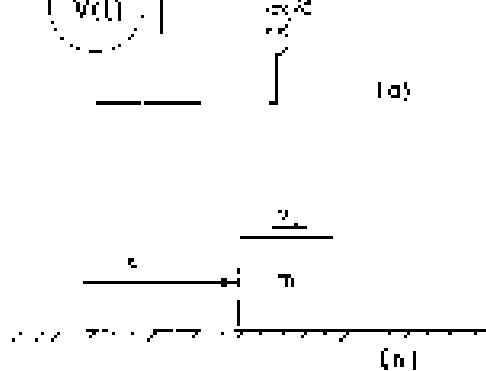


Fig. 17-10 (a) A circuit with a voltage source and an inductance. (b) An analogous mechanical system.

Particle	Coin
A (Force)	Gravitational potential energy
v (velocity)	Kinetic energy
x (displacement)	Work
$F = m \frac{dv}{dt}$	$\dot{x} = \frac{dx}{dt}$
net (kinetic energy)	$\frac{1}{2}m$
$\frac{1}{2}mv^2$ (kinetic energy)	$\frac{1}{2}x\dot{x}$ (magnetic energy)

17-8 Inductance and magnetic energy

Continuing with the analogy of the preceding section, we would expect that corresponding to the mechanical momentum $p = mv$, whose rate of change is the applied force, there would be an analogous quantity equal to Li , whose rate of change is V . We have no right, otherwise, to say that Li is the total momentum of the circuit, if i , in fact, is right. The whole circuit may be standing still and have no momentum. It is only that Li corresponds to the momentum in the sense of satisfying the corresponding equations. In the same way, to the kinetic energy mv^2 , there corresponds an analogous quantity Li^2 . But there will be a surprise. This Li^2 is really the energy in the circuit; however, this is because the rate of doing work in the inductance is Li , and in the mechanical system it is F , the expression of the quantity. Therefore, in the case of the energy, the quantity is not only correspondingly mathematically, but also has the same physical meaning as well.

We may see this in more terms as follows. As we found in Eq. (17.16), the rate of electrical work by an applied force is the product of the electromotive force and the current:

$$\frac{dW}{dt} = VI$$

Replacing E by its expression in terms of the current (from Eq. (17.14)), we have

$$\frac{dW}{dt} = -LI \frac{dI}{dt} \quad (17.19)$$

Integrating this equation, we find that the energy required from an external source to overcome the end in the self-inductance while building up the current (which must equal the energy stored, U) is

$$-W = U = \frac{1}{2}LI^2 \quad (17.20)$$

Therefore the energy stored in an inductor is $\frac{1}{2}LI^2$.

* This is, incidentally, not the only way a correspondence can be set up between mechanics and electromagnetism.

† We are neglecting any energy loss to heat from the current in the resistance of the coil. Such losses require additional energy from the source that is not changing the energy which goes into the inductor.

Applying the same arguments to a pair of coils such as those in Figs. 17-8 or 17-5, we can show that the total electrical energy of the system is given by

$$U = \frac{1}{2} L_1 I_1^2 + \frac{1}{2} L_2 I_2^2 + M I_1 I_2. \quad (17.38)$$

For, starting with $I = 0$ on both coils, we could increase the current I_1 in coil 1, with $I_2 = 0$. The work done is just $\frac{1}{2} L_1 I_1^2$. But now, on turning up I_2 , we not only do the work $\frac{1}{2} L_2 I_2^2$ against the field in circuit 2, but also an additional contribution $M I_2$, which is the integral $[M dI_2/dt]$ in circuit 1 times the new dI_2/dt due to I_1 in circuit 1.

Suppose we now wish to find the force between any two coils carrying the currents I_1 and I_2 . We might at first expect that we could use the principle of virtual work, by taking the derivative in the energy of Eq. (17.38). We must remember, of course, that as we change the relative position of the coils, the only quantity which varies is the ratio L_1/L_2 , whence M . We might then write the expression of virtual work as

$$\cdot F \Delta x = \Delta U = I_1 I_2 \Delta \theta \text{ (wrong).}$$

But this equation is wrong, because, as we have seen earlier, it reflects only the change in the energy of the two coils and not the change in the energy of the sources which are maintaining the currents I_1 and I_2 at their constant values. We can now understand that these sources must supply energy against the induced currents in the coils as they are moved. If we wish to apply the principle of virtual work correctly, we must also include these energies. As we have seen, however, we may take a short cut and use the principle of virtual work by remembering that the total energy is the negative of what we have called U , or the "natural energy". We can therefore write by the force

$$F \Delta x = -\Delta U_{\text{natural}} = -\Delta U. \quad (17.39)$$

The force between two coils is also given by

$$F \Delta x = I_1 I_2 \Delta \theta.$$

Equation (17.38) for the energy of a system of two coils can be used to show that an interesting inequality exists between mutual inductance M and the self-inductances L_1 and L_2 of the two coils. It is clear that the energy of two coils must be positive. If we begin with zero currents in the coils and increase these currents to static values, we have been adding energy to the system. If not, the currents would spontaneously increase with release of energy to the rest of the world—an unlikely thing, so happen! Now our energy equation, Eq. (17.38), can equally well be written in the following form:

$$U = \frac{1}{2} L_1 \left(I_1 - \frac{M}{L_2} I_2 \right)^2 + \frac{1}{2} \left(L_2 - \frac{M^2}{L_1} \right) I_2^2. \quad (17.40)$$

This is just an algebraic form of the first. This quantity must always be positive for any values of I_1 and I_2 . In particular, it must be positive if I_2 should happen to have the special value

$$I_2 = -\frac{L_1}{M} I_1. \quad (17.41)$$

But with this current for I_2 , the first term in Eq. (17.40) is zero. If the energy is to be positive, the last term in (17.40) must be greater than zero. We know the requirement that

$$L_2 > M^2/L_1.$$

We have thus proved the general result that the magnitude of the mutual inductance of any two coils is necessarily less than or equal to the geometric mean of the two L inductances. (This M may be positive or negative, depending on the sign

conventions for the currents i_1 and i_2)

$$V_{AB} = \sqrt{L_1 L_2} \quad (17.43)$$

The relationship between V_{AB} and the self-inductances is usually written as

$$V_{AB}^2 = L_1 L_2. \quad (17.44)$$

The constant C is called the coefficient of coupling. If most of the flux from one coil links the other coil, the coefficient of coupling is near one; we say the coils are "highly coupled." If the coils are far apart or otherwise arranged so that there is very little mutual flux linkage, the coefficient of coupling is near zero and the mutual inductance is very small.

For calculating the mutual inductance of two coils, we take profit in Eq. (17.40) a formula which is a double line integral around the two circuits. We might think that the same formula could be used to get the self-inductance of a single coil by carrying out both line integrals around the same coil. This, however, will not work, because in integrating around the two coils, the denominator r_{12} of the integrand will go to zero when the two line elements are at the same point. The self-inductance obtained from this formula is infinite. The reason is that this formula is an approximation that is valid only when the cross sections of the wires of the two circuits are small compared with the distance from one circuit to the other. Clearly, this approximation does not hold for a single coil. It is, in fact, true that the inductance of a single coil tends logarithmically to infinity as the diameter of its wire is made smaller and smaller.

We must, then, look for a different way of calculating the self-inductance of a single coil. It is necessary to take into account the distribution of the currents within the wires because the area of the wire is an important parameter. We should therefore ask not what is the inductance of a "circuit," but what is the inductance of a distribution of conductors. Perhaps the easiest way to find this inductance is to make use of the magnetic energy. We found earlier, in Section 15-5, an expression for the magnetic energy of a distribution of stationary currents:

$$U = \frac{1}{2} \int j \cdot A dV. \quad (17.45)$$

If we know the distribution of current density j , we can compute the vector potential A and then evaluate the integral of Eq. (17.45) to get the energy. This energy is equal to the inductive energy of the self-inductance, $\frac{1}{2}L^2$. Equating the two gives us a formula for the inductance:

$$L = \frac{1}{j^2} \int j \cdot A dV. \quad (17.46)$$

We expect, of course, that the inductance is a number depending only on the geometry of the circuit and not on the current I in the circuit. The formula of Eq. (17.46) will indeed give such a result, because the integral in this equation is proportional to the square of the current—the current appears once through j and again through the vector potential A . The integral divided by I^2 will depend on the geometry of the circuit, but not on the current I .

Equation (17.46) for the energy of a current distribution can be put in a quite different form which is sometimes more convenient for calculation. Also, as we will see later, it's a form that is important because it is more generally valid. In the energy equation, Eq. (17.45), both j and A can be related to B , so we can hope to express the energy in terms of the magnetic field. Just as we were able to relate the electromagnetic energy to the electric field, we begin by replacing j by $i_B r^2 \nabla \times B$. We cannot replace A so easily, since $B = -\nabla \times A$ cannot be inverted to give A in terms of B . Anyway, we can write

$$U = \frac{1}{2} \int (\nabla \times B) \cdot r^2 B dV. \quad (17.47)$$

The interesting thing is that, with some restrictions, this integral can be written as

$$U = \frac{c\epsilon_0^2}{2} \int \mathbf{B} \cdot (\nabla \times \mathbf{A}) dV. \quad (17.47)$$

To see this, we write out in detail a typical term. Suppose that we take the term $(\nabla \times \mathbf{B})_{\phi}$, which occurs in the integral of Eq. (17.46). Writing out the components, we get

$$\int \left(\frac{\partial B_x}{\partial x} - \frac{\partial B_z}{\partial y} \right) A_\phi dx dy dz.$$

(There are, of course, two more integrals of the same kind.) We now integrate the first term with respect to x , integrating by parts. That is, we can say

$$\int \frac{\partial B_x}{\partial x} A_\phi dx = B_x A_\phi - \int B_x \frac{\partial A_\phi}{\partial x} dx.$$

Now suppose that our system—measuring the sources and fields—is finite, so that as we go to large distances all fields go to zero. Then if the integrals are carried out over all space, evaluating the term $B_x A_\phi$ at the limits will give zero. We have left only the term with $B_x \partial A_\phi / \partial x$, which is evidently one part of $\mathbf{B}_x (\nabla \times \mathbf{A})_x$, and, therefore, of $\mathbf{B} \cdot (\nabla \times \mathbf{A})$. If you work out the other five terms, you will see that Eq. (17.47) is indeed equivalent to Eq. (17.46).

That shows we can replace $(\nabla \times \mathbf{A})$ by \mathbf{B} , to get

$$U = \frac{c\epsilon_0^2}{2} \int \mathbf{B} \cdot \mathbf{B} dV. \quad (17.48)$$

We have expressed the energy of a magnetostatic situation in terms of the magnetic field only. The expression corresponds directly to the formula we found for the electrostatic energy:

$$U = \frac{c\epsilon_0}{2} \int \mathbf{E} \cdot \mathbf{E} dV. \quad (17.49)$$

One reason for emphasizing these two energy formulas is that sometimes they are more convenient to use. More important, it turns out that for dynamic fields (when E and B are changing with time) the two expressions (17.48) and (17.49) remain true, whereas the other formulas we have given for electric or magnetic energies are no longer correct—they hold only for static fields.

If we know the magnetic field \mathbf{B} of a single coil, we can find the self-inductance by equating the energy expression (17.48) to $\frac{1}{2} L I^2$. Let's see how this works by finding the self-inductance of a long solenoid. We have seen earlier that the magnetic field inside a solenoid is uniform and B outside is zero. The magnitude of the field inside is $B = \mu_0 n I^2$, where n is the number of turns per cm. length in the winding and I is the current. If the radius of the coil is r and its length is L (we take L very long, so that we can neglect end effects, i.e., $L \gg r$), the volume inside is $\pi r^2 L$. The magnetic energy is therefore

$$U = \frac{c\epsilon_0^2}{2} B^2 \cdot (\nabla dV) = \frac{n^2 I^2}{2 \mu_0 c^2} \pi r^2 L,$$

which is equal to $\frac{1}{2} L I^2$. Or,

$$L = \frac{\pi r^2 n^2}{c \epsilon_0} L. \quad (17.50)$$

The Maxwell Equations

18-1 Maxwell's equations

In this chapter we come back to the complete set of the four Maxwell equations that we took as our starting point in Chapter 1. Until now, we have been studying Maxwell's equations in bits and pieces; it is time to add one final piece, and to put them all together. We will then have the complete and correct story for electro-magnetic fields that may be changing with time in any way. Anything said in this chapter that contradicts something said earlier is true and what was said earlier is false—because what was said earlier applied to such special situations as, for instance, steady currents or static charges. Although we have been very careful to point out the restrictions whenever we write an equation, it is easy to forget all of the qualifications and to learn lots well the wrong equations. Now we are ready to give the whole truth, with no qualifications (or almost none).

The complete Maxwell equations are written in Table 18-1, in words as well as in mathematical symbols. The fact that the words are equivalent to the equations should by this time be familiar—you should be able to translate back and forth from one form to the other.

The first equation—that the divergence of E is the charge density over ϵ_0 —is true in general. In dynamic as well as in static fields, Gauss' law is always valid. The flux of B through any closed surface is proportional to the charge inside. The third equation is the corresponding general law for magnetic fields. Since there are no magnetic charges, the flux of B through any closed surface is always zero. The second equation, that the curl of E is $-\partial B/\partial t$, is Faraday's law and was discussed in the last two chapters. It also is generally true. The last equation is something new. We have seen before only the part of it which holds for steady currents. In that case we said that the curl of B is $J/e\epsilon_0^2$, but the current general equation has a new part that was discovered by Maxwell.

Until Maxwell's work, the known laws of electricity and magnetism were those we have studied in Chapters 3 through 17. In particular, the equation for the magnetic field of steady currents was known only as

$$\nabla \times B = \frac{\rho}{\epsilon_0 c^2}. \quad (18.1)$$

Maxwell began by combining these known laws and expressing them as differential equations, as we have done here. (Although the ∇ notation was not yet invented, it is mainly due to Maxwell that the importance of the combinations of derivatives, which we today call the curl and the divergence, first became apparent.) He then noticed that there was something strange about Eq. (18.1). If one takes the divergence of this equation, the left-hand side will be zero, because the divergence of a curl is always zero. So this equation requires that the divergence of J also be zero. But if the divergence of J is zero, then the total flux of current out of any closed surface is also zero.

The flux of current through a closed surface is the decrease of the charge inside the surface. This quantity cannot in general be zero because we know that the charges can be moved from one place to another. The equation

$$\nabla \cdot j = -\frac{\delta \rho}{\delta t} \quad (18.2)$$

has, in fact, been almost our definition of j . This equation expresses the very funda-

18-1 Maxwell's equations

18-2 How the new term works

18-3 All of classical physics

18-4 A travelling field

18-5 The speed of light

18-6 Solving Maxwell's equations: the potentials and the wave equation

Table 18-1 Classical Physics

Maxwell's equations

$$\text{I. } \nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad (\text{Flux of } \mathbf{E} \text{ through a closed surface}) = (\text{Charge inside})/\epsilon_0$$

$$\text{II. } \nabla \times \mathbf{B} = -\frac{\partial \mathbf{E}}{\partial t} \quad (\text{Line integral of } \mathbf{B} \text{ around a loop}) = -\frac{d}{dt} (\text{Flux of } \mathbf{B} \text{ through the loop})$$

$$\text{III. } \nabla \cdot \mathbf{B} = 0 \quad (\text{Flux of } \mathbf{B} \text{ through a closed surface}) = 0$$

$$\text{IV. } \epsilon^2 \nabla \times \mathbf{B} = \frac{j}{\epsilon_0} + \frac{\partial \mathbf{E}}{\partial t} \quad \epsilon^2 (\text{Integral of } \mathbf{B} \text{ around a loop}) = (\text{Current through the loop})/\epsilon_0 \\ + \frac{1}{\epsilon_0} (\text{Flux of } \mathbf{E} \text{ through the loop})$$

Conservation of charge

$$\nabla \cdot \mathbf{j} = -\frac{\partial \rho}{\partial t} \quad (\text{Flux of current through a closed surface}) = -\frac{d}{dt} (\text{Charge inside})$$

Force law

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$$

Law of motion

$$\frac{d}{dt}(\mathbf{p}) = \mathbf{F}, \quad \text{where} \quad \mathbf{p} = \frac{mv}{\sqrt{1 - v^2/c^2}} \quad (\text{Newton's law, with Einstein's modification})$$

Gravitation

$$\mathbf{F} = -G \frac{m_1 m_2}{r^2} \mathbf{e}_r$$

initial law that electric charge is conserved—any law of charge must come from some supply. Maxwell appreciated this difficulty and proposed that it could be avoided by adding the term $\partial \mathbf{E}/\partial t$ to the right-hand side of Eq. (18.1); he then got the fourth equation in Table 18-1:

$$\text{IV. } \epsilon^2 \nabla \times \mathbf{B} = \frac{j}{\epsilon_0} + \frac{\partial \mathbf{E}}{\partial t}$$

It was far yet customary in Maxwell's time to think in terms of abstract fields. Maxwell discussed his ideas in terms of a model in which the vacuum was like an elastic medium. He also tried to explain the meaning of his new equations in terms of the mechanical model. There was much reluctance to accept his theory, first because of the model, and second because there was at first no experimental justification. Today, we understand better that what survives the equations themselves, and not the model used to get them. We may only question whether the equations are true or false. This is answered by doing experiments, and until numbers of experiments have confirmed Maxwell's equations, if we take away the scaffolding he used to build it, we find that Maxwell's beautiful edifice stands on its own. He brought together all of the laws of electricity and magnetism and made one complete and beautiful theory.

Let us show that the extra term is just what is necessary to strengthen our law differently Maxwell discovered. Taking the divergence of his equation (IV in Table 18-1), we must have that the divergence of the right-hand side is zero:

$$\nabla \cdot \frac{j}{\epsilon_0} + \nabla \cdot \frac{\partial \mathbf{E}}{\partial t} = 0. \quad (18.3)$$

In the second term, the order of the derivatives with respect to coordinates and time can be reversed, so the equation can be rewritten as

$$\nabla \cdot j - \epsilon_0 \frac{\partial}{\partial t} (\nabla \cdot E) = 0. \quad (18.4)$$

But the first of Maxwell's equations says that the divergence of E is ρ/ϵ_0 . Inserting this equality in Eq. (18.4), we get back Eq. (18.2), which we know is true. Conversely, if we accept Maxwell's equations—and we do because no one has ever found an experiment that disagrees with them—we must conclude that charge is always conserved.

The laws of physics have no answer to the question: "What happens if a charge is suddenly created at this point—what electromagnetic effects are produced?" No answer can be given because our equations say it doesn't happen. If it were to happen, we would need new laws, but we cannot say what they would be. We have not had the chance to observe how a world without charge conservation behaves. According to our equations, if you suddenly place a charge at some point, you had to carry it there from somewhere else. In that case, we can say what would happen.

When we added a new term to the equation for the curl of E , we found that a whole new class of phenomena was described. We shall see that Maxwell's little addition to the equation for $\nabla \times B$ also has far-reaching consequences. We can touch on only a few of them in this chapter.

18-2 How the new term works

As our first example we consider what happens with a spherically symmetric radial distribution of current j . Suppose we imagine a little sphere with radioactive material in it. This radioactive material is ejecting out some charged particles. (Or we could imagine a large block of jello with a small hole in the center into which some charge had been injected with a hypodermic needle and from which the charge is slowly leaking out.) In either case we would have a current that is everywhere radially outward. We will assume that it has the same magnitude in all directions.

Let the total charge inside any radius r be $Q(r)$. If the radial current density at the same radius is $j(r)$, then Eq. (18.2) requires that Q decreases at the rate

$$\frac{dQ(r)}{dr} = -4\pi r^2 j(r). \quad (18.5)$$

We now ask about the magnetic field produced by the currents in this situation. Suppose we draw some loop C on a sphere of radius r , as shown in Fig. 18-1. There is some current through this loop, so we might expect to find a magnetic field circulating in the direction shown.

But we are already in difficulty. How can the B have any particular direction on the sphere? A different choice of C would allow us to conclude that its direction is exactly opposite to that shown. So how can there be any circulation of B around the currents?

We are saved by Maxwell's equation. The circulation of B depends not only on the total current through C but also on the rate of change with time of the magnetic flux through it. It must be that these two parts just cancel. Let's see if that works out.

The electric field at the radius r must be $Q(r)/4\pi\epsilon_0 r^2$ —so long as the charge is spherically distributed, as we assume. It is radial, and its rate of change is taken

$$\frac{\partial E}{\partial r} = \frac{1}{4\pi\epsilon_0 r^2} \frac{dQ}{dt}. \quad (18.6)$$

Comparing this with Eq. (18.5), we see that at any radius

$$\frac{\partial E}{\partial r} = -\frac{j}{\epsilon_0}. \quad (18.7)$$

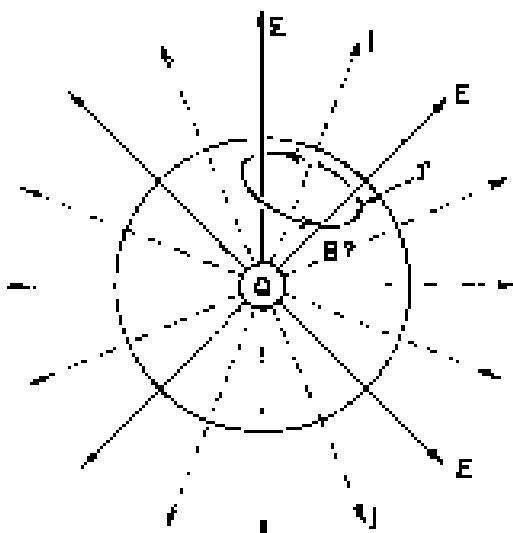


Fig. 18-1. What is the magnetic field of a spherically symmetric current?

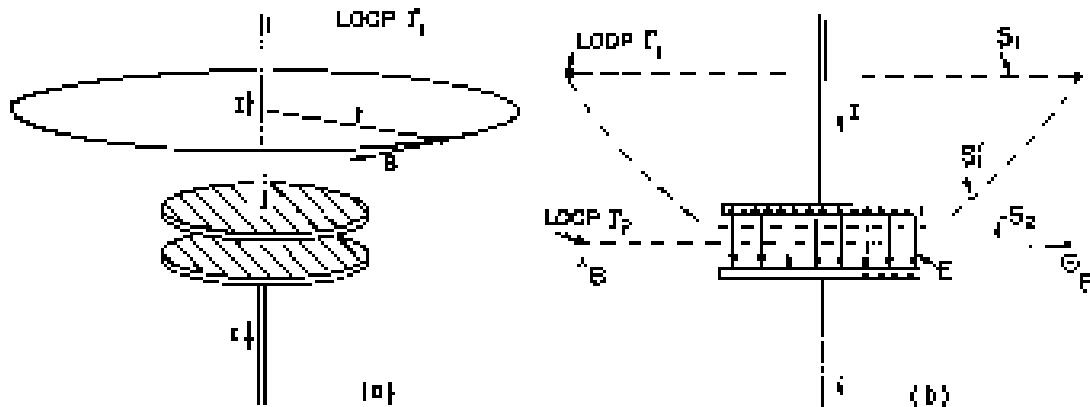


Fig. 18-2. The magnetic field near a charging capacitor.

In Eq. 14 the two source terms cancel and the end of \mathbf{B} is always zero. There is no magnetic field in our example.

As our second example, we consider the magnetic field of a wire used to charge a parallel-plate condenser (see Fig. 18-2). If the charge Q on the plates is changing with time (but not too fast), the current in the wires is equal to dQ/dt . We would expect that this current will produce a magnetic field that encircles the wire. Surely, the current close to the wire must produce the normal magnetic field—it cannot depend on where the current is going.

Suppose we take a loop Γ_1 , which is a circle with radius r , as shown in part (a) of the figure. The line integral of the magnetic field should be equal to the current I divided by $c_0 r^2$. We have

$$\oint_{\Gamma_1} \mathbf{B} \cdot d\mathbf{l} = \frac{I}{c_0 r^2}. \quad (18.8)$$

This is what we would get for a steady current, but it is also correct with Maxwell's addition, because if we consider the plane surface S_1 inside the circle, there are no electric fields on it (assuming the wire to be a very good conductor). The surface integral of $\mathbf{E} \cdot d\mathbf{l}_S$ is zero.

Suppose, however, that we now slowly move the curve Γ downward. We get always the same result until we draw even with the plates of the condenser. Then the current I goes to zero. Does the magnetic field disappear? That would be quite strange. Let's see what Maxwell's equation says for the curve Γ_2 , which is a circle of radius r whose plane passes between the condenser plates [Fig. 18-2(b)]. The line integral of \mathbf{B} around Γ_2 is $2\pi r B$. This must equal the time derivative of the flux of \mathbf{E} through the plane circular surface S_2 . This flux of \mathbf{E} , we know from Gauss' law, must be equal to $1/c_0$ times the charge Q on one of the condenser plates. We have

$$r^2 2\pi r B = \frac{d}{dt} \left(\frac{Q}{c_0} \right). \quad (18.9)$$

That is very convenient. It is the same result we found in Eq. (18.6). Integrating over the changing electric field gives the same magnetic field as does integrating over the current in the wire. Of course, that is just what Maxwell's equation says. It is easy to see that this must always be so by applying our same arguments to the two surfaces S_1 and S_2 that are bounded by the same circle Γ_1 in Fig. 18-2(b). Through S_1 there is the current I , but no electric flux. Through S_2 there is no current, but an electric flux changing at the rate I/c_0 . The same B is obtained if we use Eq. 14 with either surface.

From our discussion so far of Maxwell's new term, you may have the impression that it doesn't add much—that it just fixes up the equations to agree with what we already expect. It is true that if we just consider Eq. 14 by itself, nothing particularly new comes out. The words "by itself" are, however, all-important. Maxwell's small change in Eq. 14, when combined with the other equations, does indeed produce much that is new and important. Before we take up these matters, however, we want to speak more about Table 18-1.

18-3 All of classical physics

In Table 18-1 we have all that was known of fundamental classical physics, that is, the physics that was known by 1905. Here it all is, in one table. With these equations we can understand the complete realm of classical physics.

First we have the Maxwell equations—written in both the expanded form and the short mathematical form. Then there is the conservation of charge, which is even written in parentheses, because the moment we have the complete Maxwell equations, we can deduce from them the conservation of charge. So the table is even a little redundant. Next, we have written the force law, because having all the electric and magnetic fields doesn't tell us anything until we know where they do to charges. Knowing E and B , however, we can find the force on an object with the charge q moving with velocity v . Finally, having the force doesn't tell us anything until we know what happens when a force pushes on something; we need the law of motion, which is that the force is equal to the rate of change of the momentum. (Remember? We had that in Volume I.) We even include relativity effects by writing the momentum as $p = m_0 v / \sqrt{1 - v^2/c^2}$.

If we really want to be complete, we should add one more law—Newton's law of gravitation—but we put that at the end.

Therefore in one small table we have all the fundamental laws of classical physics—even with room to write them out in words and with some redundancy. This is a great moment. We have climbed a great peak. We are on the top of K-2—we are nearly ready for Mount Everest, which is quantum mechanics. We have climbed the peak of a "Great Divide," and now we can go down the other side.

We have mainly been trying to learn how to understand the equations. Now that we have the whole thing put together, we are going to study what the equations mean—what new things they say that we haven't already seen. We've been working hard to get up to this point. It has been a great effort, but now we are going to have nice coasting downhill as we see all the consequences of our accomplishment.

18-4 A travelling field

Now for the new consequences. They come from putting together all of Maxwell's equations. First, let's see what would happen in a circumstance which we pick to be particularly simple. By assuming that all the quantities vary only in one coordinate, we will have a one-dimensional problem. The situation is shown in Fig. 18-1. We have a sheet of charge located on the yz -plane. The sheet is first at rest. Then instantaneously given a velocity v in the x -direction, and kept moving with this constant velocity. You might worry about having such an "infinite" acceleration, but it doesn't really matter; just imagine that the velocity is brought to a very quickly. So we have suddenly a surface current J (J is the current per unit

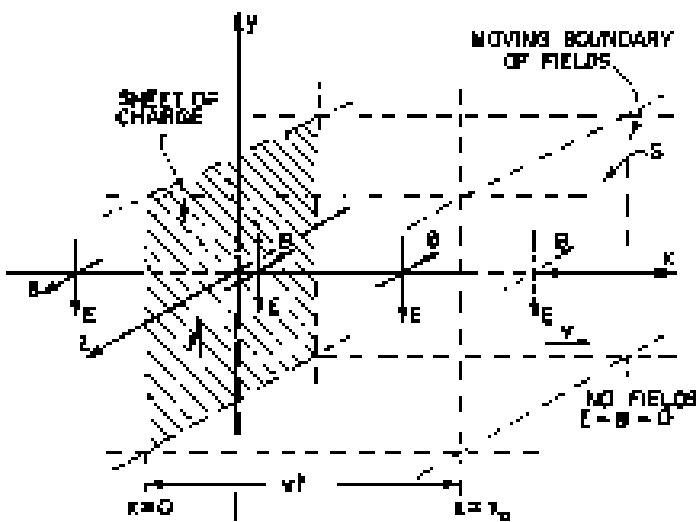


Fig. 18-3. An infinite sheet of charge is suddenly set into motion parallel to itself. There are magnetic and electric fields that propagate out from the sheet at a constant speed.

width in the z -direction). To keep the problem simple, we suppose that there is also a stationary sheet of charge of opposite sign superposed to the y -plane, so that there are no electrostatic effects. Also, although in the figure we show only what is happening in a finite region, we imagine that the sheet extends to infinity in $-y$ and $+z$. In other words, we have a situation where there is no current, and then suddenly there is a uniform sheet of current. What will happen?

Well, when there is a sheet of current in the plus y -direction, there is, as we know, a magnetic field generated which will be in the minus z -direction for $x > 0$ and in the opposite direction for $x < 0$. We could find the magnitude of B by using the fact that the line integral of the magnetic field will be equal to the current over πw^2 . We would get that $B = J/2\pi w^2$ (since the current J in a strip of width w is Jw and the line integral of B is $2B\pi$).

This gives us the field next to the sheet—for small x —but since we are imagining an infinite sheet, we would expect the same argument to give the magnetic field further out for larger values of x . However, that would mean that the moment we turn on the current, the magnetic field is suddenly changed from zero to a finite value everywhere. But wait! If the magnetic field is suddenly changed, it will produce tremendous electrical effects. (If it changes in any way, there are electrical effects.) So because we turned the sheet of charge, we make a changing magnetic field, and therefore electric fields must be generated. If there are electric fields generated, they had to start from zero and change to something else. There will be some dE/dt that will make a contribution, together with the current J , to the production of the magnetic field. So through the various equations there is a big interaction, and we have to try to solve for all the fields at once.

By looking at the Maxwell equations alone, it is not easy to see directly how to get the solution. So we will first show you what the answer is and then verify that it does indeed satisfy the equations. The answer is the following: The field B that we computed is, in fact, generated right next to the current sheet (for small x) It must be so, because if we make a tiny loop around the sheet, there is no result for any electric flux to go through it. But the field B out farther—for large x —is, at first, zero. It stays zero for awhile, and then suddenly turns on. In short, we turn on the current and the magnetic field immediately next to it turns on to a constant value B ; then the turning on of B spreads out from the source region. After a certain time, there is a uniform magnetic field everywhere out to some value x , and then zero beyond. Because of the symmetry, it spreads in both the y and z directions.

The E -field does the same thing. Before $t = 0$ (when we turn on the current), the field is zero everywhere. Then after the time t , both E and B are uniform out to the distance $x \approx w$, and zero beyond. The fields make their way forward like a tidal wave, with a front moving at a uniform velocity which turns out to be c , but for a while we will just call it v . A graph of the magnitude of E or B versus x , as they appear at the time t , is shown in Fig. 18-4(a). Looking again at Fig. 18-3, at the time t , the region between $x = -w$ is “filled” with the fields, but they have not yet reached beyond. We emphasize again that we are assuming that the current sheet and, therefore, the fields E and B , extend infinitely far in both the y - and z -directions. (We cannot draw an infinite sheet, so we have shown only what happens in a finite area.)

We want now to analyze qualitatively what is happening. To do that, we want to look at two cross-sectional views, a top view looking down along the y -axis, as shown in Fig. 18-5, and a side view looking back along the x -axis, as shown in Fig. 18-6. Suppose we start with the side view. We see the charged sheet moving up; the magnetic field points into the page for $+x$, and out of the page for $-x$, and the electric field is downward everywhere—out to $x = \pm w$.

Let's see if these fields are consistent with Maxwell's equations. Let's first draw one of those loops that we will be able to calculate a line integral, say the rectangle R shown in Fig. 18-6. You notice that one side of the rectangle is in the region where there are fields, but one side is in the region the fields have still not reached. There is some capacitive flux through this loop. If E is changing, there should be an emf around it. If the wavelength is w , we will have a changing magnetic

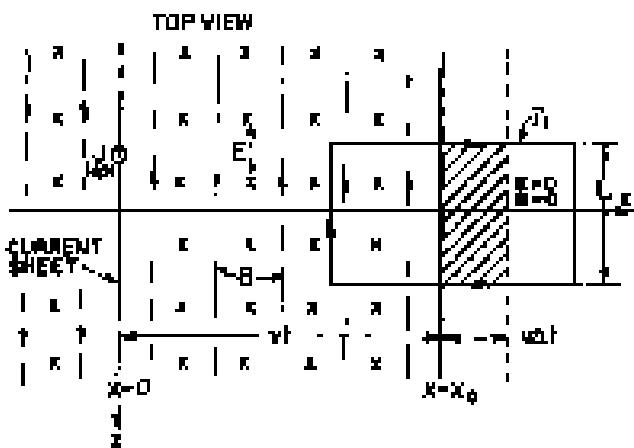


Fig. 18-5. Top view of Fig. 18-3.

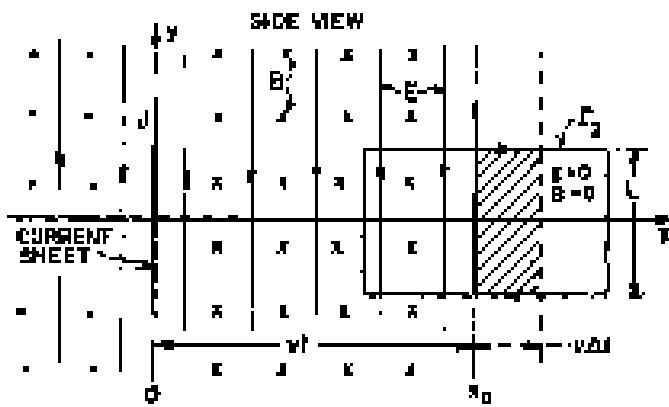


Fig. 18-6. Side view of Fig. 18-3.

flux, because the area in which B exists is progressively increasing at the velocity v . The flux inside I_2 is B times the part of the area inside I_2 which has a magnetic field. The rate of change of the flux, since the magnitude of B is constant, is the magnitude times the rate of change of the area. The rate of change of the area is easy. If the width of the rectangle I_2 is L , the area in which B exists changes by $L \cdot \Delta x$ in the time Δt . (See Fig. 18-6.) The rate of change of flux is then $B L v$. According to Faraday's law, this should equal the line integral of E around I_2 , which is just $E L$. We have the equation

$$E = vB. \quad (18.10)$$

So if the ratio of E to B is v , the fields we have assumed will satisfy Faraday's equation.

But that is not the only equation, we have the other equation relating E and B :

$$c^2 \nabla \times B = \frac{d}{dt} + \frac{\partial E}{\partial t}. \quad (18.11)$$

To apply this equation, we look at the top view in Fig. 18-5. We have seen that this equation will give us the value of B next to the current sheet. Also, for any loop drawn outside the sheet but behind the wavefront, there is no curl of B nor any j or changing E , so the equation is correct there. Now let's look at what happens for the curve I_1 that intersects the wavefront, as shown in Fig. 18-5. There there are no currents, so Eq. (18.11) can be written—in integral form—as

$$c^2 \oint_{I_1} B \cdot d\mathbf{s} = \frac{d}{dt} \int_{\text{inside } I_1} \mathbf{E} \cdot d\mathbf{s}. \quad (18.12)$$

The line integral of B is just B times L . The rate of change of the flux of E is due only to the advancing wavefront. The area inside I_1 , where E is not zero, is increasing at the rate vL . The right-hand side of Eq. (18.12) is then vLE . That equation becomes

$$c^2 B = Lv. \quad (18.13)$$

We have a solution in which we have a constant B and a constant E behind the front, both at right angles to the direction in which the front is moving and at right angles to each other. Maxwell's equations specify the ratio of E to B . From Eqs. (18.10) and (18.13),

$$E = vB, \quad \text{and} \quad E = \frac{c^2}{v} B.$$

But one moment! We have found two different conditions on the ratio E/B . Can such a field as we describe really exist? There is, of course, only one velocity v for which both of these equations can hold, namely $v = c$. The wavefront must travel with the velocity c . We have no example in which the electrical influence from a current propagates at a certain finite velocity v .

Now let's ask what happens if we suddenly stop the motion of the charged sheet after it has been on for a short time T . We can see what will happen by the principle of superposition. We had a current that was zero and then was suddenly turned on. We knew the solution for that case. Now we are going to add another set of fields. We take another charged sheet and suddenly start it moving, in the opposite direction with the same speed, only at the time T after we started the first current. The total current of the two added together is first zero, then on for a time T , then off again—because the two currents cancel. We have a square "pulse" of current.

The new negative current produces the same fields as the positive one, only with all the signs reversed and, of course, delayed in time by T . A wave front again travels out at the velocity c . At the time t it has reached the distance $x = c(t - T)$, as shown in Fig. 18-4(b). So we have two "blocks" of field traveling out at the speed c , as in parts (a) and (b) of Fig. 18-4. The combined fields are as shown in part (c) of the figure. The fields are zero for $x > ct$; they are constant (with the values we found above) between $x = ct - T$ and $x = ct$, and again zero for $x < c(t - T)$.

In short, we have a little piece of field—a block of thickness cT —which has left the current source and is travelling through space all by itself. The fields have "taken off"; they are propagating freely through space, no longer connected in any way with the source. The caterpillar has turned into a butterfly!

How can this bundle of electric and magnetic fields sustain itself? The answer is: by the combined effects of the Faraday law, $\nabla \times E = -\partial B/\partial t$, and the new term of Maxwell, $c^2 \nabla \times B = \partial E/\partial t$. They cannot help maintaining themselves. Suppose the magnetic field were to disappear. There would be a changing magnetic field which would produce an electric field. If this electric field tried to go away, the changing electric field would create a magnetic field back again. So by a perpetual interplay—by the switching back and forth from one field to the other—they must go on forever. It is impossible for them to disappear.* They sustain themselves in a kind of a dance—one making the other, the second making the first—propagating outward through space.

18-5 The speed of light

We have a wave which leaves the spherical source and goes outward at the velocity c , which is the speed of light. But let's go back a moment. From a historical point of view, it wasn't known that the coefficient c in Maxwell's equations was also the speed of light propagation. There was just a constant in the equations. We have called it c from the beginning, because we knew what it would turn out to be. We didn't think it would be sensible to make you learn the formulas with a different constant and then go back to substitute c wherever it belonged. From the point of view of electricity and magnetism, however, we just start out with two constants, ϵ_0 and c^2 , that appear in the equations of electrostatics and magnetostatics;

$$\nabla \cdot B = \frac{\rho}{\epsilon_0} \quad (18.14)$$

and

$$\nabla \times B = \frac{j}{\epsilon_0 c^2}. \quad (18.15)$$

If we take any arbitrary definition of a unit of charge, we can determine experimentally the constant ϵ_0 required in Eq. (18.14)—say by measuring the force between two unit charges at rest, using Coulomb's law. We must also determine experimentally the constant c, c^2 that appears in Eq. (18.15), which we can do, say, by measuring the force between two unit currents. (A unit current means one unit of charge per second.) The ratio of these two experimental constants is c^2 , just another "electromagnetic constant!"

* Well, not quite. They can be "absorbed" if they get to a region where there are charges. By which we mean that other fields can be produced somewhere which superpose on these fields and "cancel" them by destructive interference (see Chapter 21, Vol. I).

Notice now that this constant c^2 is the same no matter what we choose for our unit of charge. If we put twice as much "charge"—say twice as many proton charges—in our "unit" of charge, ϵ_0 would need to be one-fourth as large. When we pass two of these "unit" currents through two wires, there will be twice as much "charge" per second in each wire, so the force between two wires is four times larger. The constant $\epsilon_0 c^2$ must be reduced by one-fourth. But the ratio $\epsilon_0 c^2 / \mu_0$ is unchanged!

So just by experiments with charges and currents we find a number c^2 which turns out to be the square of the velocity of propagation of electromagnetic influences. From static measurements—by measuring the forces between two unit charges and between two unit currents—we find that $c = 3.00 \times 10^8$ meters/sec. When Maxwell first made this calculation with his equations, he said that bundles of electric and magnetic fields should be propagated at this speed. He also remarked on the mysterious coincidence that this was the same as the speed of light. "We can scarcely avoid the inference," said Maxwell, "that light consists in the transverse vibrations of the same medium which is the cause of electric and magnetic phenomena!"

Maxwell had made one of the great unifications of physics. Before his time, there was light, and there was electricity and magnetism. The latter two had been unified by the experimental work of Faraday, Oersted, and Ampere. Then, all of a sudden, light was no longer "something else," but was only electricity and magnetism in the new form—little pieces of electric and magnetic fields which propagate through space on their own.

We have called your attention to some characteristics of this special solution, which turn out to be true, however, for any electromagnetic wave: that the magnetic field is perpendicular to the direction of motion of the wavefront; that the electric field is likewise perpendicular to the direction of motion of the wavefront; and that the two vectors E and B are perpendicular to each other. Furthermore, the magnitude of the electric field E is equal to c times the magnitude of the magnetic field B . These three facts—that the two fields are transverse to the direction of propagation, that B is perpendicular to E , and that $E = cB$ —are generally true for any electromagnetic wave. (One special case is a good one—it shows all the main features of electromagnetic waves.)

18-6 Solving Maxwell's equations; the potentials and the wave equation

Now we would like to do something mathematical; we want to write Maxwell's equations in a simpler form. You may consider that we are complicating them, but if you will be patient a little bit, they will suddenly become simpler. Although by this time you are thoroughly used to each of the Maxwell equations, there are many places that must all be put together. That's what we want to do.

We begin with $\nabla \cdot B = 0$ —the simplest of the equations. We know that it implies that B is the curl of something. So, if we write

$$\mathbf{B} = \nabla \times \mathbf{A}, \quad (18.16)$$

we have already solved one of Maxwell's equations. (Incidentally, you hypothesize that it remains true that another vector A' would be just as good if $A' = A - \nabla\phi$ —where ϕ is any scalar field—because the curl of $\nabla\phi$ is zero, and B is still the same. We have talked about that before.)

We take next the Faraday law, $\nabla \times E = -\partial B / \partial t$, because it doesn't involve any currents or charges. If we write B as $\nabla \times A$ and differentiate with respect to t , we can write Faraday's law in the form

$$\nabla \times E = -\frac{\partial}{\partial t} \nabla \times A.$$

Since we can differentiate either with respect to time or to space first, we can also write this equation as

$$\nabla \times \left(E + \frac{\partial A}{\partial t} \right) = 0. \quad (18.17)$$

We see that $\mathbf{E} = \partial \mathbf{A} / \partial t$ is a vector whose curl is equal to zero. Therefore that vector is the gradient of something. When we worked on electrostatics, we had $\nabla \times \mathbf{B} = 0$, and then we decided that \mathbf{B} itself was the gradient of something. We took it to be the gradient of $-\phi$ (the minus for technical convenience). We do the same thing for $\mathbf{E} + \partial \mathbf{A} / \partial t$: we set

$$\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = -\nabla \phi. \quad (18.18)$$

We use the same symbol ϕ so that, in the electrostaticic case where nothing changes with time and the $\partial \mathbf{A} / \partial t$ term disappears, \mathbf{E} will be our old $-\nabla \phi$. So Faraday's equation can be put in the form

$$\mathbf{E} = -\nabla \phi - \frac{\partial \mathbf{A}}{\partial t}. \quad (18.19)$$

We have solved two of Maxwell's equations already, and we have found that to describe the electromagnetic fields \mathbf{E} and \mathbf{B} , we need four potential functions: a scalar potential ϕ and a vector potential \mathbf{A} , which is, of course, three functions.

Now that \mathbf{A} determines part of \mathbf{E} , as well as \mathbf{B} , what happens when we change \mathbf{A} to $\mathbf{A}' = \mathbf{A} + \nabla \psi$? In general \mathbf{E} would change if we didn't take some special precautions. We can, however, still allow \mathbf{A} to be changed in this way without affecting the fields \mathbf{E} and \mathbf{B} : that is, without changing the physics—if we always change \mathbf{A} and ϕ together by the rules

$$\mathbf{A}' = \mathbf{A} + \nabla \psi, \quad \phi' = \phi - \frac{\partial \psi}{\partial t}. \quad (18.20)$$

Then neither \mathbf{B} nor \mathbf{E} , obtained from Eq. (18.19), is changed.

Previously, we chose to make $\nabla \cdot \mathbf{A} = 0$, to make the equations of statics somewhat simpler. We are not going to do that now; we are going to make a different choice. But we'll wait a bit before saying what the choice is, because later it will be clear why the choice is made.

Now we return to the two remaining Maxwell equations which will give us relations between the potentials and the sources ρ and \mathbf{j} . Once we can determine \mathbf{A} and ϕ from the currents and charges, we can always get \mathbf{E} and \mathbf{B} from Eqs. (18.16) and (18.19), so we will have another form of Maxwell's equations.

We begin by substituting Eq. (18.19) into $\nabla \cdot \mathbf{E} = \rho / \epsilon_0$; we get

$$\nabla \cdot \left(-\nabla \phi - \frac{\partial \mathbf{A}}{\partial t} \right) = \frac{\rho}{\epsilon_0},$$

which we can write also as

$$-\nabla^2 \phi - \frac{\partial}{\partial t} \nabla \cdot \mathbf{A} = \frac{\rho}{\epsilon_0}. \quad (18.21)$$

This is one equation relating ϕ and \mathbf{A} to the sources.

Our final equation will be the most complicated. We start by rewriting the fourth Maxwell equation as

$$c^2 \nabla \times \mathbf{B} - \frac{\partial \mathbf{E}}{\partial t} = \frac{\mathbf{j}}{\epsilon_0},$$

and then substitute for \mathbf{B} and \mathbf{E} in terms of the potentials, using Eqs. (18.16) and (18.19):

$$c^2 \nabla \times (\nabla \times \mathbf{A}) - \frac{\partial}{\partial t} \left(-\nabla \phi - \frac{\partial \mathbf{A}}{\partial t} \right) = \frac{\mathbf{j}}{\epsilon_0}.$$

The first term can be rewritten using the algebraic identity: $\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$; we get

$$-c^2 \nabla^2 \mathbf{A} + c^2 \nabla(\nabla \cdot \mathbf{A}) + \frac{\partial}{\partial t} \nabla \phi - \frac{\partial^2 \mathbf{A}}{\partial t^2} = \frac{\mathbf{j}}{\epsilon_0}. \quad (18.22)$$

It's not very simple!

Fortunately, we can now make use of our freedom to choose arbitrarily the divergence of A . What we are going to do is to use our choice to fix things so that the equations for A and for ϕ are separated but have the same form. We can do this by taking*

$$\nabla \cdot A = - \frac{1}{c^2} \frac{\partial \phi}{\partial t}. \quad (18.23)$$

When we do this, the two middle terms in A and ϕ in Eq. (18.22) cancel, and that equation becomes much simpler:

$$\nabla^2 A - \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} = - \frac{j}{\epsilon_0 c^2}. \quad (18.24)$$

And our equation for ϕ —Eq. (18.21)—takes on the same form:

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = - \frac{\rho}{\epsilon_0}. \quad (18.25)$$

What a beautiful set of equations! They are beautiful, first, because they are nicely separated—with the charge density, plus ϕ , with the current, plus A . Furthermore, although the left side looks a little funny—a Laplacian together with a $(\partial/\partial t)^2$ —when we unfold it we see

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = - \frac{\rho}{\epsilon_0}. \quad (18.26)$$

It has a nice symmetry in x , y , z —the $-1/c^2$ is necessary because, of course, time and space are different; they have different units.

Maxwell's equations have led us to a new kind of equation for the potentials ϕ and A but to the same mathematical form for all four functions ϕ , A_x , A_y , and A_z . Once we learn how to solve these equations, we can get B and E from $\nabla \times A$ and $-\nabla \phi - \partial A/\partial t$. We have another form of the electromagnetic laws exactly equivalent to Maxwell's equations, and in many situations they are much simpler to handle.

We have, in fact, already solved an equation much like Eq. (18.26). When we studied sound in Chapter 17 of Vol. I, we had an equation of the form

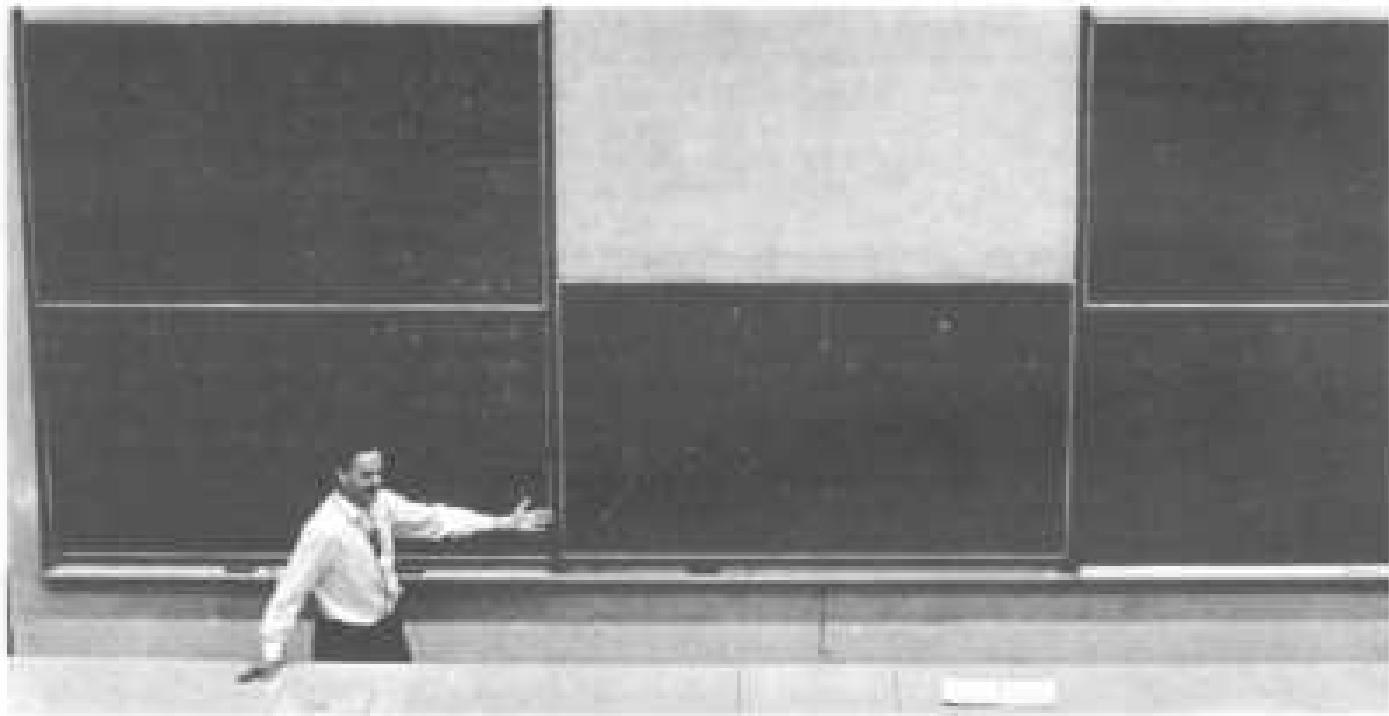
$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2},$$

and we saw that it described the propagation of waves in the x -direction at the speed c . Equation (18.26) is the corresponding wave equation for three dimensions. So in regions where there are no longer any charges and currents, the solution of these equations is not that ϕ and A are zero. (Although that is indeed one possible solution.) There are solutions in which there is some set of ϕ and A which are changing in time but always moving out at the speed c . The fields travel outward through free space, as in our example at the beginning of the chapter.

With Maxwell's new term in Eq. 1V, we have been able to write the field equations in terms of A and ϕ in a form that is simple and thus makes immediately apparent that there are electromagnetic waves. For many practical purposes, it will still be convenient to use the original equations in terms of E and B . But they are on the other side of the mountain we have already climbed. Now we are ready to cross over to the other side of the peak. Things will look different—we are ready for some new and beautiful views.

* Choosing the $\nabla \cdot A$ is called "choosing a gauge." Changing A by adding $\nabla \phi$ is called a "gauge transformation." Equation (18.23) is called "the Lorenz gauge."

The Principle of Least Action



A special lecture—almost verbatim*

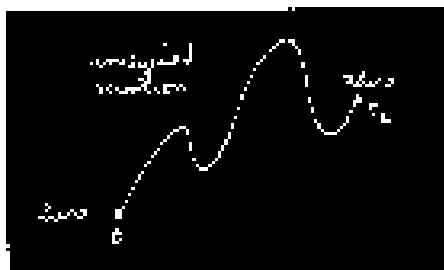
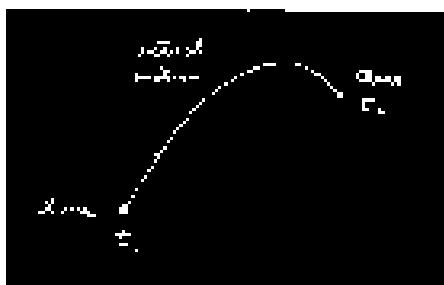
"When I was in high school, my physics teacher—whose name was Mr. Bader—called me down one day after physics class and said, 'You look bored; I want to tell you something interesting.' Then he told me something which I found absolutely fascinating, and have, since then, always found fascinating. Every time the subject comes up, I work on it. In fact, when I began to prepare this lecture I found myself making more analyses on the thing. Instead of worrying about the lecture, I got involved in a new problem. The subject is this—the principle of least action.

"Mr. Bader told me the following: Suppose you have a particle (in a gravitational field, for instance) which starts somewhere and moves to some other point by free motion—you draw it, and it goes up and comes down.

It goes from the original place to the final place in a certain amount of time. Now, you try a different curve. Suppose that to get from here to there, it would take

but put there is just the same amount of time. There is only this: If you calculate the kinetic energy at every instant on the path, take away the potential energy, and integrate it over the time during the whole path, you'll find that the number you'll get is bigger than that for the actual motion.

* Later chapters depend largely on the material of this special lecture—which is intended to be for "entertainment."



"In other words, the laws of Newton could be stated not in the form $\ddot{x} = F$, but in the form: the average kinetic energy less the average potential energy is as little as possible for the path of an object going from one point to another.

"Let me illustrate a little bit better what I mean. If you have an idea of the particle form's field, then if the particle has the path $x(t)$ let's take the difference for a moment: we take a trajectory like goes up and down and just sideways), where x is the height above the ground, the kinetic energy is $\frac{1}{2}m(\dot{x})^2$, and the potential energy at any time is mgh . Now I take the kinetic energy minus the potential energy at every moment along the path and integrate that with respect to time from the initial time to the final time. Let's suppose that at the original time t_0 we start at some height and at the end of the time t_f we are definitely ending at some other place.

"Then the integral is

$$\int_{t_0}^{t_f} \left[\frac{1}{2}m\left(\frac{dx}{dt}\right)^2 - mgh \right] dt$$

The whole motion is some kind of a curve—it's a parabola if we plot against the time—and gives a certain value for the integral. But we could imagine some other motion that went very high and came up and down in some peculiar way.

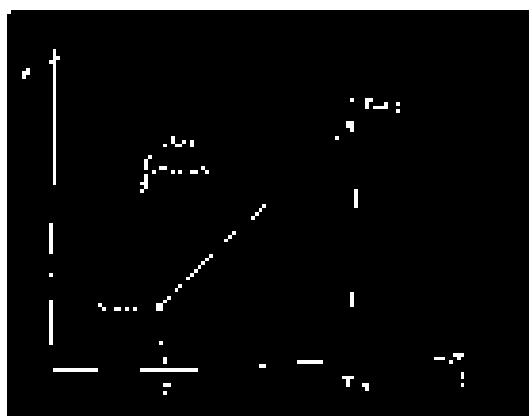
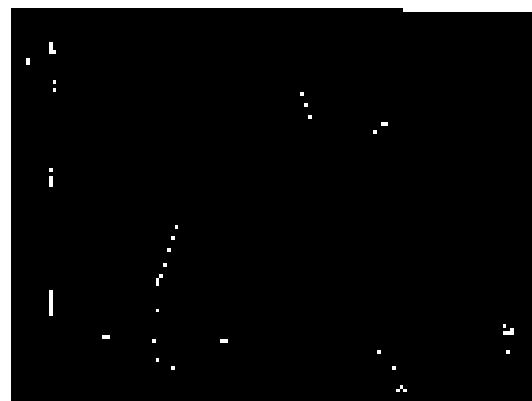
We can calculate the kinetic energy minus the potential energy and integrate for such a path... or for any other path we want. The remark is that the true path is the one for which this integral is least.

"Let's try it out. First, suppose we take the case of a free particle for which there is no potential energy at all. Then the rule says that in going from one point to another in a given amount of time, the kinetic energy integral is least, so it must go at uniform speed. (We know that's the right answer—because it's a uniform speed.) Why is that? Because if the particle were to go any other way, the velocities would be sometimes higher and sometimes lower than the average. The average velocity is the same for every case because it has to get from 'here' to 'there' in a given amount of time.

"As an example, say your job is to start from home and get to school in a given length of time with this car. You can do it several ways: You can加速 a little mad in the beginning and slow down with the brakes near the end, or you can go at a uniform speed, or you can go backwards for a while and then go forward, and so on. The thing is that average speed has got to be, of course, the total distance that you have gone over the time. But if you do anything but go at a uniform speed, then sometimes you are going too fast and sometimes you are going too slow. Now the mean square of something that deviates around an average, as you know, is always greater than the square of the mean: so the kinetic energy integral would always be higher than with your velocity than if you went the uniform velocity. So we see that the integral is a minimum, if the velocity is a constant (when there are no forces). The curve path is like this.

"Now, an object thrown up in a gravitational field does rise faster first and then slows down. That is because there is also the potential energy, and we must have the least difference of kinetic and potential energy on the average. However the potential energy does as we go to a space, we will get a little different. So we can get as much as possible up to where there is a high potential energy. Then we can take that potential away from the kinetic energy and get a lower average. So it is better to take a path which goes up and gets a lot of negative stuff from the potential energy.

"On the other hand, you can't go up too fast, or you fall because you will then have too much kinetic energy involved: you have to go very low to get very up and come down again in the fixed amount of time available. So you don't want to go too far up, but you want to go up some. So it turns out that the solution is some kind of balance between trying to get more potential energy with the least amount of extra kinetic energy. Try to find the differences, to take minus the potential, as small as possible.



"That is all my teacher told me, because he was a very good teacher and knew when to stop talking. But I don't know when to stop talking. So instead of leaving it as an interesting remark, I am going to horrify and disgust you with the complexities of life by proving that it is so. The kind of mathematical problem we will have is very difficult and a new kind. We have a certain quantity which is called the action, S . It is the kinetic energy, minus the potential energy, integrated over time.

$$\text{Action} = S = \int_{t_1}^{t_2} (KE - PE) dt.$$

Remember that the PE and KE are both functions of time. For each different possible path you get a different number for this action. Our mathematical problem is to find out for what curve that number is the least.

"You say—Oh, that's just the ordinary calculus of maxima and minima. You calculate the action and just differentiate to find the minimum.

"But watch out. Ordinarily we just have a function of some variable, and we have to find the value of that variable where the function is least or most. For instance, we have a rod which has been heated in the middle and the heat is spread around. For each point on the rod we have a temperature, and we want find the point at which that temperature is largest. But now for each path in space we have a number—quite a different thing—and we have to find the path in space for which the number is the minimum. That is a completely different branch of mathematics. It is not the ordinary calculus. In fact, it is called the calculus of variations.

"There are many problems in this kind of mathematics. For example, the circle is usually defined as the locus of all points at a constant distance from a fixed point, but another way of defining a circle is this: a circle is that curve of given length which encloses the biggest area. Any other curve encloses less area for a given perimeter than the circle does. So if we give the problem: find that curve which encloses the greatest area for a given perimeter, we would have a problem of the calculus of variations—a different kind of calculus than you're used to.

"So we make the calculation for the path of an object. Here is the way we are going to do it. The idea is that we imagine that there is a true path and that any other curve we draw is a false path, so that if we calculate the action for the false path we will get a value that is bigger than if we calculate the action for the true path.

"Problem: Find the true path. Where is it? One way, of course, is to calculate the action for millions and millions of paths and look at which one is lowest. When you find the lowest one, that's the true path.

"That's a possible way. But we can do it better than that. When we have a quantity which has a minimum—for instance, in an ordinary function like the temperature—one of the properties of the minimum is that if we go away from the minimum in the first order, the derivative of the function from its minimum value is only second order. At any place else on the curve, if we move a small distance the value of the function changes also in the first order. But at a minimum, a tiny motion away makes, in the first approximation, no difference.

"That is what we are going to use to calculate the true path. If we have the true path, a curve which differs only a little bit from it will, in the first approximation, make no difference in the action. Any difference will be in the second approximation, if we really have a minimum.

"That is easy to prove. If there is a change in the first order when I deviate the curve a certain way, there is a change in the action that is proportional to the deviation. The change presumably makes the action greater; otherwise we haven't got a minimum. But then if the change is proportional to the deviation, reversing the sign of the deviation will make the action less. We would get the action to increase one way and to decrease the other way. The only way that it could really be a minimum is that in the first approximation it doesn't make any change, that the changes are proportional to the square of the deviations from the true path.



"So we work in this way: We call $x(t)$ (without underbrace) the true path—the one we are trying to find. We take some trial path $\underline{x}(t)$ that differs from the true path by a small amount which we will call $\eta(t)$ (eta of t)."

"Now the idea is that if we calculate the action S for the path $x(t)$, then the difference between that S and the action that we calculated for the path $\underline{x}(t)$ —to simplify the writing we can call it \underline{S} —the difference of S and \underline{S} must be zero in the first-order approximation of small η . It can differ in the second order, but in the first order the difference must be zero."

"And that cannot be true for any η at all. Well, not quite. The method doesn't mean anything unless you consider paths which all begin and end at the same two points—each path begins at a certain point at t_1 and ends at a certain other point at t_2 , and those points and times are kept fixed. So the deviations in our η have to be zero at each end, $\eta(t_1) = 0$ and $\eta(t_2) = 0$. With that condition, we have specified our mathematical problem."

"If you didn't know any calculus, you might do the same kind of thing to find the minimum of an ordinary function $f(x)$. You could discuss what happens if you take $f(x)$ and add a small amount δ to x and argue that the correction to $f(x)$ in the first order in δ must be zero at the minimum. You would substitute $x + \delta$ for x and expand out to the first order in δ . . . just as we are going to do with η .

"The idea is then that we substitute $\underline{x}(t) = x(t) + \eta(t)$ in the formula for the action:

$$S = \int \left[\frac{m}{2} \left(\frac{dx}{dt} \right)^2 - V(x) \right] dt,$$

where I call the potential energy $V(x)$. The derivative dx/dt is, of course, the derivative of $\underline{x}(t)$ plus the derivative of $\eta(t)$, so for the action S get this expression:

$$S = \int_{t_1}^{t_2} \left[\frac{m}{2} \left(\frac{dx}{dt} + \frac{d\eta}{dt} \right)^2 - V(x + \eta) \right] dt.$$

"Now I must write this out in more detail. For the squared term I get

$$\left(\frac{dx}{dt} \right)^2 + 2 \frac{dx}{dt} \frac{d\eta}{dt} + \left(\frac{d\eta}{dt} \right)^2.$$

But wait. I'm not worrying about higher than the first order, so I will take all the terms which involve η^2 and higher powers and put them in a little box called 'second and higher order.' From this term I get only second order, but there will be more from something else. So the kinetic energy part is

$$\frac{m}{2} \left(\frac{dx}{dt} \right)^2 = m \frac{dx}{dt} \frac{d\eta}{dt} + (\text{second and higher order}).$$

"Now we need the potential V at $x + \eta$. I consider η small, so I can write $V(x)$ as a Taylor series. It is approximately $V(x)$: in the next approximation (from the ordinary nature of derivatives) the correction is η times the rate of change of V with respect to x , and so on:

$$V(x + \eta) = V(x) + \eta V'(x) + \frac{\eta^2}{2} V''(x) + \dots$$

I have written V' for the derivative of V with respect to x in order to save writing. The term in η^2 and the ones beyond fall into the 'second and higher order' category and we don't have to worry about them. Putting it all together,

$$S = \int_{t_1}^{t_2} \left[\frac{m}{2} \left(\frac{dx}{dt} \right)^2 - V(x) + m \frac{dx}{dt} \frac{d\eta}{dt} - \eta V'(x) + (\text{second and higher order}) \right] dt.$$



Now if we look carefully at the thing, we see that the first two terms which I have arranged here correspond to the action of that I would have calculated with the true path x . The thing I want to concentrate on is the change in S —the difference between the S and the \bar{S} that we would get for the right path. This difference we will write as δS , called the variation in S . Leaving out the 'second and higher order' terms, I have for δS

$$\delta S = \int_{x_1}^{x_2} \left[m \frac{dx}{dt} \frac{d\eta}{dx} + \eta V(x) \right] dt.$$

"Now the problem is this: Here is a certain integral. I don't know what the x is yet, but I do know that no matter what η is, this integral must be zero. Well, you think, the only way that that can happen is that what multiplies η must be zero. But what about the first term with $d\eta/dx$? Well, after all, if η can be anything at all, its derivative is anything also, so you conclude that the coefficient of $d\eta/dx$ must also be zero. That isn't quite right. It isn't quite right because there is a connection between η and its derivative; they are not absolutely independent, because $\eta(x)$ must be zero at both x_1 and x_2 .

"The method of solving all problems in the calculus of variations always uses the same general principle. You make the shift in the thing you want to vary (as we did by adding η); you look at the first-order terms, then you always arrange things in such a form that you get an integral of the form 'some kind of stuff times the shift (η)', but with no other derivatives (no $d\eta/dt$). It must be rearranged so it is always 'something' times η . You will see the great value of that in a minute. (There are formulas that tell you how to do this in some cases without actually calculating, but they are not general enough to be worth bothering about: the best way is to calculate it just this way.)

"How can I rearrange the term in $d\eta/dt$ to make it have an η ? I can do that by integrating by parts. It turns out that the whole trick of the calculus of variations consists of writing down the variation of S and then integrating by parts so that the derivatives of η disappear. It is always the same in every problem in which derivatives appear.

"You remember the general principle for integrating by parts. If you have any function f times $d\eta/dt$ integrated with respect to t , you write down the derivative of tf :

$$\frac{d}{dt}(tf) = \eta \frac{df}{dt} + f \frac{d\eta}{dt}.$$

The integral you want is over the last term, so

$$\int f \frac{d\eta}{dt} dt = -tf + \int \eta \frac{df}{dt} dt.$$

"In our formula for δS , the function f is m times dx/dt ; therefore, I have the following formula for δS :

$$\delta S = m \frac{dx}{dt} \eta(t) \Big|_{x_1}^{x_2} - \int_{x_1}^{x_2} \frac{d}{dt} \left(m \frac{dx}{dt} \right) \eta(t) dt - \int_{x_1}^{x_2} V(x) \eta(t) dt.$$

The first term must be evaluated at the two limits x_1 and x_2 . Then I must have the integral from the rest of the integration by parts. The last term is brought down without change.

"Now comes something which always happens—the integrated part disappears. (In fact, if the integrated part does not disappear, you restate the principle, adding conditions to make sure it does!) We have already said that η must be zero at both ends of the path, because the principle is that the action is a minimum provided that the varied curve begins and ends at the chosen points. The condition is that

$\eta(t_1) = 0$ and $\eta(t_2) = 0$. So the integral term is zero. We collect the other terms together and obtain this:

$$S = \int_{t_1}^{t_2} \left[-m \frac{d^2x}{dt^2} - F'(x) \right] \eta(t) dt.$$

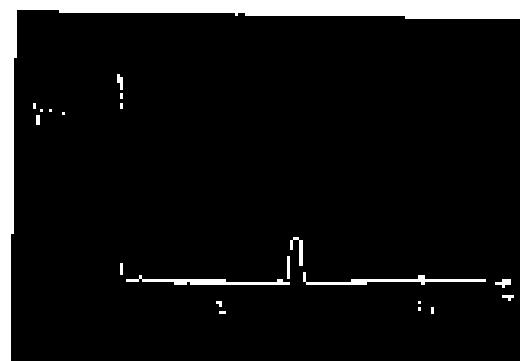
The variation in S is now the way we wanted it—there is the stuff in brackets, say F , all multiplied by $\eta(t)$ and integrated from t_1 to t_2 .

"We have that an integral of something or other times $\eta(t)$ is always zero:

$$\int F(t) \eta(t) dt = 0.$$

I have some function of t ; I multiply it by $\eta(t)$; and I integrate it from one end to the other. And no matter what the η is, I get zero. That means that the function $F(x)$ is zero. That's obvious, but anyway I'll show you one kind of proof.

"Suppose that for $\eta(t)$ I took something which was zero for all t except right near one particular value. It stays zero until it gets to this t ,



then it blips up for a moment and blips right back down. When we do the integral of this η times any function F , the only place that you get anything other than zero was where $\eta(t)$ was blipping, and then you get the value of F at that place times the integral over the blip. The integral over the blip alone isn't zero, but when multiplied by F it has to be; so the function F has to be zero where the blip was. But the blip was anywhere I wanted to put it, so F must be zero everywhere.

"We see that if our integral is zero for any η , then the coefficient of η must be zero. The action integral will be a minimum for the path that satisfies this complicated differential equation:

$$\left[-m \frac{d^2x}{dt^2} - F'(x) \right] = 0.$$

It's not really so complicated; you have seen it before. It's just $F = -ma$. The first term is the mass times acceleration, and the second is the derivative of the potential energy, which is the force.

"So, for a conservative system at least, we have demonstrated that the principle of least action gives the right answer; it says that the path that has the *minimum* action is the one satisfying Newton's law.

"One remark: I did not prove it was a minimum—maybe it's a maximum. In fact, it doesn't really have to be a minimum. It is quite analogous to what we found for the 'principle of least time' which we discussed in optics. There also, we said at first it was 'least' time. It turned out, however, that there were situations in which it wasn't the least time. The fundamental principle was that for any first-order variation away from the optical path, the change in time was zero; it is the same story. What we really mean by 'least' is that the first-order change in the value of S , when you change the path is zero. It is not necessarily a minimum."

"Next, I remark on some generalizations. In the first place, the thing can be done in three dimensions. Instead of just x , I would have x , y , and z as functions of t ; the action is then complicated. For three-dimensional motion, you have to use the complete kinetic energy $(m/2)$ times the whole velocity squared. That is,

$$KE = \frac{m}{2} \left| \left(\frac{dx}{dt} \right)^2 + \left(\frac{dy}{dt} \right)^2 + \left(\frac{dz}{dt} \right)^2 \right|.$$

Also, the potential energy is a function of x , y , and z . And what about the path? The path is some general curve in space, which is not so easily drawn, but the idea is the same. And what about the η ? Well, η can have three components. You could shift the path in x , or in y , or in z —or you could shift in all three directions simultaneously. So η would be a vector. This doesn't really complicate things too much, though. Since only the first-order variation has to be zero, we can do the calculation by three successive shifts. We can shift x only in the x -direction and

say that coefficient never be zero. We get one equation. Then we shift it in the y -direction and get another. And in the z -direction and get another. Or, of course, in any order that you want. Anyway, you get three equations. And, of course, Newton's law is really three equations in the three dimensions—one for each component. I think that you can practically see that it is bound to work, but we will leave you to show for yourself that it will work for three dimensions. Indeed, you might use any coordinate system you want, polar or otherwise, and get Newton's laws appropriate in that system right off by seeing what happens if you have the shift τ in radius, or in angle, etc.

"Similarly, the method can be generalized to any number of particles. If you have, say, two particles with a force between them, so that there is a mutual potential energy, then you just add the kinetic energy of both particles and take the potential energy of the mutual interaction. And what do you vary? You vary the paths of both particles. Then, for two particles moving in three dimensions, there are six equations. You can vary the position of particle 1 in the x -direction, in the y -direction, and in the z -direction, and similarly for particle 2; so there are six equations. And that's as it should be. There are the three equations that determine the acceleration of particle 1 in terms of the force on it and three for the acceleration of particle 2, from the force on it. You follow the same game through, and you get Newton's law in three dimensions for any number of particles.

"I have been saying that we get Newton's law. That is not quite true, because Newton's law includes nonconservative forces like friction. Newton said that $m\ddot{x}$ is equal to say F . But the principle of least action only works for conservative systems—where all forces can be gotten from a potential function. You know, however, that at a microscopic level—on the deepest level of physics—there are no nonconservative forces. Nonconservative forces, like friction, appear only because we neglect microscopic complications—there are just too many particles to analyze. But the fundamental laws can be put in the form of a principle of least action.

"Let me generalize still further. Suppose we ask what happens if the particle moves relativistically. We did not get the right relativistic equations of motion; $F = ma$ is only right nonrelativistically. The question is: Is there a corresponding principle of least action for the relativistic case? There is. The formula in the case of relativity is the following:

$$S = -mc^2 \int_{t_1}^{t_2} \sqrt{1 - v^2/c^2} dt - q \int_{r_1}^{r_2} [\phi(r, \dot{r}, z, t) - \mathbf{A} \cdot \mathbf{dr}] dt.$$

The first part of the action integral is the rest mass m , times c^2 times the integral of a function of velocity, $\sqrt{1 - v^2/c^2}$. Then instead of just the potential energy, we have an integral over the scalar potential ϕ and over v times the vector potential A . Of course, we are then including only electromagnetic forces. All electric and magnetic fields are given in terms of ϕ and A . This action function gives the complete theory of relativistic motion of a single particle in an electromagnetic field.

"Of course, whenever I have written v , you understand that before you try to figure anything out, you must substitute dx/dt for v_x and so on for the other components. Also, you put the path along the path at time t , $x(t), y(t), z(t)$ where I wrote simply x, y, z . Properly, it is only after you have made these replacements for the v 's that you have the formula for the action for a relativistic particle. I will leave to the more advanced of you the problem to demonstrate that this action formula does, in fact, give the correct equations of motion for relativity. May I suggest you do it first without the A , that is, for no magnetic field? Then you should get the components of the equation of motion, $dv/dt = -q \nabla \phi$, where, you remember, $\rho = mv/\sqrt{1 - v^2/c^2}$.

"It is much more difficult to include even the case with a vector potential. The variations get much more complicated. But in the end, the force term does turn out equal to $q(E + \mathbf{v} \times \mathbf{B})$, as it should. But I will leave that for you to play with.

"I would like to emphasize that in the general case, for instance in the relativistic formula, the action integral no longer has the form of the kinetic energy

minus the potential energy. That's only true in the non-relativistic approximation. For example, the term $m\epsilon c^2 \sqrt{1 - v^2/c^2}$ is not what we have called the kinetic energy. The question of what the action should be for any particular case must be determined by some kind of trial and error. It is just the same problem as determining what are the laws of motion in the first place. You just have to fiddle around with the equations that you know and see if you can get them into the form of the principle of least action.

"One other point on terminology. The function that is integrated over time to get the action S is called the Lagrangian, L , which is a function only of the velocities and positions of particles. So the principle of least action is also written

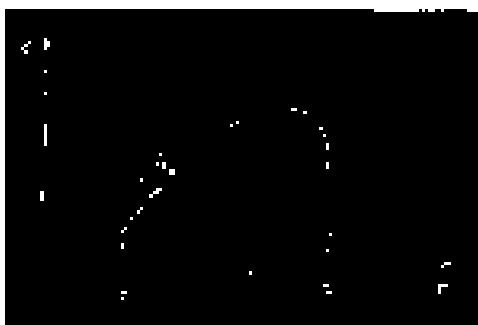
$$S = \int_{t_1}^{t_2} L(x_i, \dot{x}_i) dt,$$

where by x_i and \dot{x}_i are meant all the components of the positions and velocities. So if you hear someone talking about the 'Lagrangian,' you know they are talking about the function that is used to find S . For relativistic motion in an electromagnetic field

$$\mathcal{L} = m\epsilon c^2 \sqrt{1 - \frac{\dot{x}^2}{c^2}} - q(\phi - \mathbf{A} \cdot \mathbf{x}).$$

"Also, I should say that S is not really called the 'action' by the most precise and pedantic people. It is called 'Hamilton's first principal function.' Now I have to give a lecture on 'the-principle-of-least-Hamilton's-first-principal-function.' So I call it 'the action.' Also, more and more people are calling it the action. You see, historically something else which is not quite as useful was called the action, but I think it's more sensible to change to a newer definition. So now you too will call the new function the action, and pretty soon everybody will call it by that simple name.

"Now I want to say some things on this subject which are similar to the discussions I gave about the principle of least time. There is quite a difference in the characteristic of a law which says a certain integral from one place to another is a minimum—which tells something about the whole path—versus a law which says that as you go along, there is a force that makes it accelerate. The second way tells how you took your way along the path, and the other is a stated statement about the whole path. In the case of light, we talked about the connection of these two. Now, I would like to explain why it is true that there are differential laws when there is a least action principle of this kind. The reason is the following: Consider the usual path in space and time. As before, let's take only one dimension, so we can plot the graph of x as a function of t . Along the true path, S is a minimum. Let's suppose that we have the true path and that it goes through some point a in space and time, and also through another nearby point b .



Now if the entire integral from t_1 to t_2 is a minimum, it is also necessary that the integral along the little section from a to b is also a minimum. It can't be that the path from a to b is a little bit more. Otherwise you could just jingle with just that piece of the path and make the whole integral a little lower.

"So every subsection of the path must also be a minimum. And this is true no matter how short the subsection. Therefore the principle that the whole path gives a minimum can be stated also by saying that an infinitesimal section of path along has a curve such that it has a minimum action. Now if we take a short enough section of path—between two points a and b very close together—how the potential varies from one place to another far away is not the important thing, because you are staying almost in the same place over the whole little piece of the path. The only thing that you have to discuss is the first-order change in the potential. The answer can only depend on the derivative of the potential and not on the potential everywhere. So the statement about the gross property of the whole path becomes a statement of what happens for a short section of the path—a differential statement. And this differential statement only involves the derivatives of the potential, that is, the force at a point. That's the qualitative explanation of the relation between the gross law and the differential law.

"In the case of light we also discussed the question: How does the particle find the right path? From the differential point of view, it is easy to understand. Every moment it gets an acceleration, and knows only what to do at that instant. But all your instincts or values and effect go haywire when you say that the particle decides to take the path that is going to give the minimum action. Does it 'smell' the neighboring paths to find out whether or not they have more action? In the case of light, when we put blocks in the way so that the photons could not test all the paths, we found that they couldn't figure out which way to go, and we had the phenomenon of diffraction.

"Is the same thing true in mechanics? Is it true that the particle doesn't just 'take the right path' but that it looks at all the other possible trajectories? And if by having blocks in the way, we don't let it look, that we will get an analog of diffraction? The miracle of it all is, of course, that it does just that. That's what the laws of quantum mechanics say. So our principle of least action is incompletely stated. It isn't that a particle takes the path of least action but that it smells all the paths in the neighborhood and chooses the one that has the least action by a method analogous to the one by which light chose the shortest time. You remember that the way light chose the shortest time was this: If it went on a path that took a different amount of time, it would arrive at a different phase. And the total amplitude at some point is the sum of contributions of amplitude for all the different ways the light can arrive. All the paths that give wildly different phases don't add up to anything. But if you can find a whole sequence of paths which have phases almost all the same, then the little contributions will add up and you get a reasonable total amplitude to arrive. The important path becomes the one for which there are many nearby paths which give the same phase.

"It is just exactly the same thing for quantum mechanics. The complete quantum mechanics (for the nonrelativistic case and neglecting electron spin) works as follows: The probability that a particle starting at point 1 at the time t_1 will arrive at point 2 at the time t_2 is the square of a probability amplitude. The total amplitude can be written as the sum of the amplitudes for each possible path—for each way of arrival. For every path that we could have—for every possible trajectory—we have to calculate an amplitude. Then we add them all together. What do we take for the amplitude for each path? Our action integral tells us what the amplitude for a single path ought to be. The amplitude is proportional to some constant times $e^{iS/\hbar}$, where S is the action for that path. That is, if we represent the phase of the amplitude by a complex number, the phase angle is S/\hbar . The action S has dimensions of energy times time, and Planck's constant \hbar has the same dimensions. It is the constant that determines when quantum mechanics is important.

"Here is how it works: Suppose that for all paths, S is very large compared to \hbar . One path contributes a certain amplitude. For a nearby path, the phase is quite different, because with an enormous S even a small change in S means a completely different phase—because \hbar is so tiny. So nearby paths will normally cancel their effects out in taking the sum—except for one region, and that is where a path and a nearby path will give the same phase in the first approximation (more precisely, the same action within \hbar). Only those paths will be the important ones. So in the limiting case in which Planck's constant \hbar goes to zero, the correct quantum-mechanical laws can be summarized by simply saying: 'Forget about all these probability amplitudes. The particle does go on a special path, namely, that one for which S does not vary in the first approximation.' That's the relation between the principle of least action and quantum mechanics. The fact that quantum mechanics can be formulated in this way was discovered in 1942 by a student of that same teacher, Radler, I spoke of at the beginning of this lecture. [Quantum mechanics was originally formulated by giving a differential equation for the amplitude (Schrödinger) and also by some other people (mathematicians (Heisenberg).)]

"Now I want to talk about other minimum principles in physics. There are many very interesting ones. I will not try to list them all now but will only describe one more. Later on, when we come to a physical phenomenon which has a nice minimum principle, I will tell about it then. I want now to show that we can de-

scribe electrostatics, not by giving a differential equation for the field, but by saying that a certain integral is a maximum or a minimum. First, let's take the case where the charge density is known everywhere, and the problem is to find the potential ϕ everywhere in space. You know that the answer should be

$$\nabla^2\phi = -\rho/\epsilon_0.$$

But another way of stating the same thing is this: Calculate the integral U^* , where

$$U^* = \frac{\epsilon_0}{2} \int (\nabla\phi)^2 dV - \int \rho\phi dV,$$

which is a volume integral to be taken over all space. This thing is a minimum for the correct potential distribution $\phi(x, y, z)$.

We can show that the two statements about electrostatics are equivalent. Let's suppose that we pick any function ϕ . We want to show that when we take for ϕ the correct potential ϕ , plus a small deviation f , then in the first order, the change in U^* is zero. So we write

$$\phi = \phi_0 + f.$$

The ϕ is what we are looking for, but we are making a variation of it to find what it has to be so that the variation of U^* is zero to first order. For the first part of U^* , we need

$$(\nabla\phi)^2 = (\nabla\phi_0)^2 + 2\nabla\phi \cdot \nabla f + (\nabla f)^2.$$

The only first-order term that will vary is

$$2\nabla\phi \cdot \nabla f.$$

In the second term of the quantity U^* , the integrand is

$$\rho\phi = \rho\phi_0 + \rho f,$$

whose variable part is ρf . So, keeping only the variable parts, we need the integral

$$\Delta U^* = \int (\epsilon_0 \nabla\phi_0 \cdot \nabla f - \rho f) dV.$$

Now, following the old general rule, we have to get the damn thing all clear of derivatives of f . Let's look at what the derivatives are. The dot product is

$$\frac{\partial \phi}{\partial x} \frac{\partial f}{\partial x} + \frac{\partial \phi}{\partial y} \frac{\partial f}{\partial y} + \frac{\partial \phi}{\partial z} \frac{\partial f}{\partial z},$$

which we have to integrate with respect to x , to y , and to z . Now here is the trick: to get rid of $\partial f / \partial x$ we integrate by parts with respect to x . That will carry the derivative over onto the ϕ . It's the same general idea we used to get rid of derivatives with respect to z . We use the equality

$$\int \frac{\partial \phi}{\partial x} \frac{\partial f}{\partial x} dx = f \frac{\partial \phi}{\partial x} - \int f \frac{\partial^2 \phi}{\partial x^2} dx.$$

The integrated term is zero, since we have to make f zero at infinity. (That corresponds to making ϕ zero at x_1 and x_2 . So our principle should be more accurately stated: U^* is best for the true ϕ than for any other $\phi(x, y, z)$ having the same values at infinity.) Then we do the same thing for y and z . So our integral ΔU^* is

$$\Delta U^* = \int (-\epsilon_0 \nabla^2 \phi_0 - \rho) f dV.$$

In order for this variation to be zero for any f , no matter what, the coefficient of f must be zero and, therefore,

$$\nabla^2 \phi = -\rho/\epsilon_0.$$

We get back our old equation. So our "minimum" proposition is correct.

"We can generalize our proposition if we do our algebra in a little different way. Let's go back and do our integration by parts without taking components. We start by looking at the following equality:

$$\nabla \cdot (f \nabla \phi) = \nabla f \cdot \nabla \phi + f \nabla^2 \phi.$$

If I differentiate out the left-hand side, I can show that it is just equal to the right-hand side. Now we can use this equation to integrate by parts. In our integral ΔU^* , we replace $-\nabla \phi \cdot \nabla f$ by $\nabla^2 \phi \cdot \nabla f$, which gets integrated over volume. The divergence term integrated over volume can be replaced by a surface integral:

$$\int \nabla \cdot (f \nabla \phi) dV = \int f \nabla \phi \cdot n d\sigma.$$

Since we are integrating over all space, the surface over which we are integrating is at infinity. There, f is zero and we get the same answer as before.

"Only now we see how to solve a problem when we don't know where all the charges are. Suppose that we have conductors with charges spread out on them, in some way. We can still use our minimum principle if the potentials of all the conductors are fixed. We carry out the integral for ΔU^* only in the space outside of all conductors. Then, since we can't vary ϕ on the conductor, f is zero on all those surfaces, and the surface integral

$$\int f \nabla \phi \cdot n d\sigma$$

is still zero. The remaining volume integral

$$\Delta U^* = \int (-\epsilon_0 \nabla^2 \phi - \rho \phi) dV$$

is only to be carried out in the spaces between conductors. Of course, we get Poisson's equation again,

$$\nabla^2 \phi = -\rho/\epsilon_0$$

So we have shown that our original integral U^* is also a minimum if we evaluate it over the space outside of conductors all at fixed potentials (that is, such that any trial $\phi(x, y, z)$ must equal the given potential of the conductor, which x, y, z is a point on the surface of a conductor).

"There is an interesting case when the only charges are on conductors. Then

$$U^* = \frac{\epsilon_0}{2} \int (\nabla \phi)^2 dV.$$

Our minimum principle says that in the case where there are conductors set at certain given potentials, the potential between them adjusts itself so that integral U^* is least. What is this integral? The term $\nabla \phi$ is the electric field, so the integral is the electrostatic energy. The true field is the sum of all those coming from the gradient of a potential, with the minimum total energy.

"I would like to use this result to calculate something particular to show you that these things are really quite practical. Suppose I take two conductors in the form of a cylindrical condenser.

The inside conductor has the potential V , and the outside is at the potential zero. Let the radius of the inside conductor be a and that of the outside, b . Now we can suppose any distribution of potential between the two. If we use the correct ϕ , and calculate $\epsilon_0/2 \int (\nabla \phi)^2 dV$, it should be the energy of the system, $\frac{1}{2}CV^2$.



So we can also calculate C by our principle. But if we use a wrong distribution of potential and try to calculate the capacity C by this method, we will get a capacity that is too big, since V is specified. Any assigned potential ϕ that is not the exactly correct one will give a fake C that is larger than the correct value. But if my false ϕ is any rough approximation, the C will be a good approximation, because the error in C is second order to the error in ϕ .

"Suppose I don't know the capacity of a cylindrical condenser. I can use this principle to find it. I just guess at the potential function ϕ until I get the lowest C . Suppose, for instance, I pick a potential that corresponds to a constant field. (You know, of course, that the field isn't really constant here; it varies as $1/r$.) A field which is constant means a potential which goes linearly with distance. To fit the conditions at the two conductors, it must be

$$\phi = V \left(1 - \frac{r-a}{b-a} \right).$$

This function is V at $r = a$, zero at $r = b$, and in between has a constant slope equal to $-V/(b-a)$. So what one does is find the integral $\int d^3r$ is multiply the square of this gradient by $\epsilon_0/2$ and integrate over all volume. Let's do this calculation for a cylinder of unit length. A volume element at the radius r is $2\pi r dr$. Doing the integral, I find that my first try at the capacity gives

$$\frac{1}{2} C V^2 (\text{first try}) = \frac{\epsilon_0}{2} \int_a^b \frac{V^2}{(b-a)r} 2\pi r dr.$$

The integral is easy; it is just

$$\pi V^2 \left(\frac{b+a}{b-a} \right).$$

So I have a formula for the capacity which is not the true one but is an approximate job:

$$\frac{C}{2\pi\epsilon_0} = \frac{b+a}{2(b-a)}.$$

It is, naturally, different from the correct answer $C = 2\pi a \sqrt{b/a}$, but it's not too bad. Let's compare it with the right answer for several values of b/a . I have computed out the answers in this table:

a	$\frac{C_{\text{true}}}{2\pi\epsilon_0}$	$\frac{C(\text{first approx})}{2\pi\epsilon_0}$
2	1.4423	1.900
4	0.721	0.833
10	0.454	0.612
100	0.267	0.31
1.5	2.4662	2.50
1.1	10.492070	10.500000

Even when b/a is as big as 2—which gives a pretty big variation in the field compared with a linearly varying field—I get a pretty fair approximation. The answer is, of course, a little too high, as expected. The thing gets much worse if you have a tiny wire inside a big cylinder. Then the field has enormous variations and if you represent it by a constant, you're not doing very well. With $b/a = 100$, we're off by nearly a factor of two. Things are much better for small b/a . To take the opposite extreme, when the conductors are not very far apart—say $b/a = 1.1$ —then the constant field is a pretty good approximation, and we get the correct value for C to within a tenth of a percent.

"Now I would like to tell you how to improve such a calculation. (Of course, you know the right answer for the cylinder, but the method is the same for some other field shapes, where you may not know the right answer.) The next step is to try a better approximation to the unknown true ϕ . For example, we might try a

constant plus an exponential ϕ , etc. But how do you know when you have a better approximation unless you know the true ϕ ? Answer: You calculate C ; the lowest C is the value nearest the truth. Let us try this idea out. Suppose that the potential is not linear but say quadratic in r —that the electric field is not constant but linear. The most general quadratic form that fits $\phi = 0$ at $r = b$ and $\phi = V$ at $r = a$ is

$$\phi = V \left[1 + \alpha \left(\frac{r-a}{b-a} \right) - (1+\alpha) \left(\frac{r-a}{b-a} \right)^2 \right],$$

where α is any constant number. This formula is a little more complicated. It involves a quadratic term in the potential as well as a linear term. It is very easy to get the field out of it. The field is just

$$E = -\frac{d\phi}{dr} = -\frac{\alpha V}{b-a} + 2(1+\alpha) \frac{(r-a)V}{(b-a)^2}.$$

Now we have to square this and integrate over volume. But wait a moment. What should I take for α ? I can take a parabola for the ϕ ; but what parabola? Here's what I do: Calculate the capacity with an arbitrary α . When I get to

$$\frac{C}{2\pi\epsilon_0} = \frac{\alpha}{b-a} \left[\frac{b}{a} \left(\frac{a^3}{b^2} + \frac{2a}{b} + 1 \right) + \frac{1}{6} \alpha^2 + 1 \right].$$

It looks a little complicated, but it occurs out of integrating the square of the field. Now I can pick any α . I know that the truth lies somewhere between anything that I am going to calculate, so whatever I put in for α is going to give me an answer too big. But if I keep playing with α and get the lowest possible value (say, that lowest value is nearer to the truth than any other value). So what I do next is to pick the α that gives the minimum value for C . Working it out by ordinary calculus, I get that the minimum C occurs for $\alpha = -2b/(b+a)$. Substituting that value into the formula, I obtain for the minimum capacity

$$\frac{C}{2\pi\epsilon_0} = \frac{b^2 + 4ab + a^2}{3(b^2 - a^2)}.$$

"I've worked out what this formula gives for C for various values of b/a . I call these numbers $C_{\text{quadratic}}$. Here is a table that compares $C_{\text{quadratic}}$ with the true C .

$\frac{b}{a}$	$\frac{C_{\text{true}}}{2\pi\epsilon_0}$	$\frac{C_{\text{quadratic}}}{2\pi\epsilon_0}$
2	1.4423	1.444
4	0.721	0.733
10	0.434	0.475
100	0.267	0.346
1.5	2.4662	2.4667
1.1	10.492070	10.492065

"For example, when the ratio of the radii is 2 to 1, I have 1.444, which is a very good approximation to the true answer, 1.4423. Even for larger b/a , it stays pretty good—it is much, much better than the first approximation. It is even fairly good, only off by 10 percent, when b/a is 10 to 1. But when it gets to be 100 to 1—well, things begin to go wild. I get that C is 0.346 instead of 0.267. On the other hand, for a ratio of radii of 1.5, the answer is excellent; and for a b/a of 1.1, the answer comes out 10.492065 instead of 10.492070. Where the answer should be good, it is very, very good."

"I have given these examples, first, to show the theoretical value of the principles of minimum action and minimum principles in general and, second, to show their practical utility—not just to calculate a capacity when we already know the answer. For any other shape, you can guess an approximate field with some unknown parameters like α and adjust them to get a minimum. You will get excellent numerical results for otherwise intractable problems."

A note added after the lecture

"I should like to add something that I didn't have time for in the lecture. (I always seem to prepare more than I have time to tell about.) As I mentioned earlier, I got interested in a problem while working on this lecture. I want to tell you what that problem is. Among the minimum principles that I could mention, I noticed that most of them sprang in one way or another from the least action principle of mechanics and electrodynamics. But there is also a class that does not. As an example, if currents are made to go through a piece of material obeying Ohm's law, the currents distribute themselves inside the piece so that the rate at which heat is generated is as little as possible. Also we can say (if things are kept Euclidean) that the rate at which energy is generated is a minimum. Now, this principle also holds, according to classical theory, in determining even the distribution of velocities of the electrons inside a metal which is carrying a current. The distribution of velocities is not exactly the equilibrium distribution [Chapter 40, Vol. I; Eq. (40.6)] because they are drifting sideways. The new distribution can be found from the principle that it is the distribution for a given current for which the energy developed per second by collisions is as small as possible. The true description of the electrons' behavior ought to be by quantum mechanics, however. The question is: Does the same principle of minimum entropy generation also hold when the situation is described quantum-mechanically? I haven't found that yet.

"The question is interesting academically, of course. Such principles are fascinating, and it is always worth while to try to see how general they are. But also from a more practical point of view, I want to know. I, with some colleagues, have published a paper in which we calculated by quantum mechanics approximately the electrical resistance left by an electron moving through an ionic crystal like NaCl. [Feynman, Hellworth, Hedges, and Platman, "Mobility of Slow Electrons in a Polar Crystal," *Phys. Rev.* 127, 1004 (1962).] But if a minimum principle existed, we could use it to make the results much more accurate, just as the minimum principle for the capacity of a condenser permitted us to get such accuracy for that capacity even though we had only a rough knowledge of the electric field."

Solutions of Maxwell's Equations in Free Space

20-1 Waves in free space; plane waves

In Chapter 18 we had reached the point where we had 1 to Maxwell's equations in complete form. All that is left now is to know what the classical theory of the electric and magnetic fields can tell us of the joint equations:

$$\text{I. } \nabla \cdot E = \frac{\rho}{\epsilon_0}$$

$$\text{II. } \nabla \times E = -\frac{\partial B}{\partial t}$$

$$\text{III. } \nabla \cdot B = 0$$

$$\text{IV. } c\nabla \times B = \frac{j}{\epsilon_0} + \frac{\partial E}{\partial t} \quad (20.1)$$

When we put all these equations together, a remarkable new phenomenon occurs: fields generated by moving charges can leave the sources and travel alone through space. We considered a special example in which an infinite current sheet is suddenly turned on. After the current has been on for the time t , there are uniform electric and magnetic fields extending out the distance $c(t)$ from the source. Suppose that the current sheet lies in the xy -plane with a surface current density j going toward positive x . The electric field will have only a y -component, and the magnetic field, only a z -component. The magnitude of the field components is given by

$$E_y = cB_z = \frac{j}{2\epsilon_0 c} \quad (20.2)$$

for positive values of x less than $c(t)$. For larger x the fields are zero. There are, of course, similar fields extending the same distance from the current sheet in the negative x direction. In Fig. 20-1 we show a graph of the magnitude of the fields as a function of x at the instant t . As time goes on, the "wavefront" at $c(t)$ moves outward in x at the constant velocity c .

Now consider the following sequence of events. We turn on a current of unit strength for a while, then suddenly increase the current strength to three units, and hold it constant at this value. What do the fields look like then? We can see what the fields will look like in the following way. First, we imagine a current of unit strength that is turned on at $x = 0$ and left constant forever. The fields for positive x are then given by the graph in part (a) of Fig. 20-2. Next, we ask what would happen if we turn on a steady current of two units at the time t_1 .

The fields in this case will be twice as high as before, but will extend out in x only the distance $c(t - t_1)$, as shown in part (b) of the figure. When we add these two solutions, using the principle of superposition, we find that the sum of the two sources is a current of one unit for the time from zero to t_1 , and a current of three units for times greater than t_1 . At the time t the fields will vary with x as shown in part (c) of Fig. 20-2.

Now let's take a more complicated problem. Consider a current which is turned on to one unit for a while, then turned up to three units, and later turned off to zero. What are the fields for such a current? We can find the solution in the same way - by adding the solutions of three separate problems. First, we find the fields for a step current of unit strength. (We have solved that problem already.) Next, we find the fields produced by a step current of two units. Finally, we solve for the fields of a step current of seven three units. When we add the three solutions, we will have a current which is one unit strong from $x = 0$ to some later time, say t_1 , then three units strong until a still later time t_2 , and then turned off—(or

20-1 Waves in free space; plane waves

20-2 Three-dimensional waves

20-3 Scientific imagination

20-4 Spherical waves

References: Chapter 49, Vol. I: *Sound*; *The Wave Approach*
Chapter 50, Vol. I: *Electromagnetic Radiation*

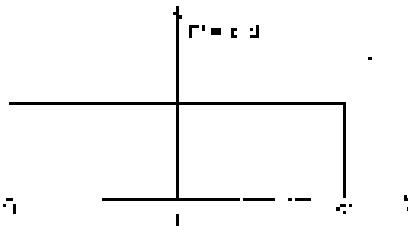


Fig. 20-1. The electric and magnetic field as a function of x at the time t after the current sheet is turned on.

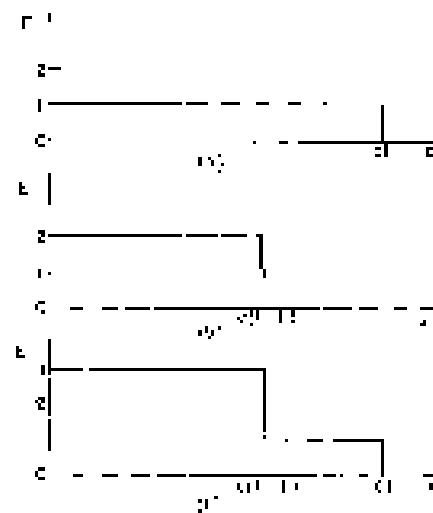


Fig. 20-2. The electric field of a current sheet. (a) One unit of current turned on at $t = 0$; (b) Two units of current turned on at $t = t_1$; (c) Superposition of (a) and (b).

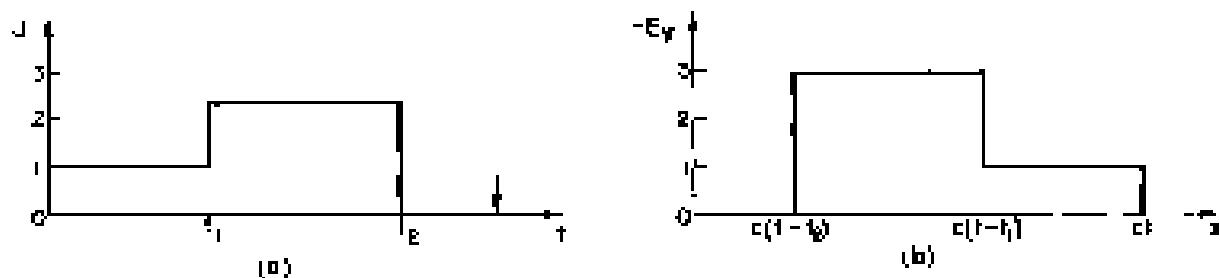


Fig. 20-3. If the current source strength varies as shown in (a), then at the time t shown by the arrow the electric field as a function of x is as shown in (b).

is to zero. A graph of the current as a function of time is shown in Fig. 20-3(a). When we find the three solutions for the electric field, we find that its variation with x , at a given instant t , is as shown in Fig. 20-3(b). The field is no valid representation of the current. The field distribution in space is a nice graph of current variations with time, only drawn backwards. As time goes on, the whole picture moves outward at the speed c , so there is a little blob of field travelling toward positive x , which contains a completely detailed memory of the history of all the current variations. If we were to stand miles away, we could tell from the variation of the electric or magnetic field exactly how the current had varied at the source.

You will also notice that long after all activity at the source has completely stopped and all charges and currents are zero, the blob of field continues to travel through space. We have a distribution of electric and magnetic fields that exist independently of any charges or currents. That is the new effect that comes from the complete set of Maxwell's equations. If we want, we can give a complete mathematical representation of the analysis we have just done by writing that the electric field at a given place and a given time is proportional to the current at the source, only not at the same time, but at the earlier time $t - x/c$. We can write

$$E_y(t) = \frac{J(t-x/c)}{2\epsilon_0 c} \quad (20.5)$$

We have, believe it or not, already derived this same equation from another point of view in Vol. I, when we were dealing with the theory of the index of refraction. Then, we had to figure out what fields were produced by a thin layer of oscillating dipoles in a sheet of dielectric material, with the dipoles set in motion by the electric field of an incoming electromagnetic wave. Our problem was to calculate the combined fields of the original wave and the waves emitted by the oscillating dipoles. How could we have calculated the fields produced by moving charges when we didn't have Maxwell's equations? At that time we took as our starting point (without any derivation) a formula for the radiation fields produced at large distances from an accelerating point charge. If you will look in Chapter 41 of Vol. I, you will see that Eq. (21.10) there is just the same as the Eq. (20.5) that we have just written down. Although our earlier derivation was correct only at large distances from the source, we see now that the same result continues to be correct even right up to the source.

We want now to look in a general way at the behavior of electric and magnetic fields in empty space far away from the sources, i.e., from the currents and charges. Very near the sources—near enough so that during the delay in transmission, the source has not had time to change much—the fields are very much the same as we have found in what we called the electrostatic or magnetostatic cases. If we go out to distances large enough so that the delays become important, however, the nature of the fields can be radically different from the solutions we have found. In a sense, the fields begin to lose some character of their own when they have gone a long way from all the sources. So we can begin by discussing the behavior of the fields in a region where there are no currents or charges.

Suppose we ask: What kind of fields can there be in regions where ρ and j are both zero? In Chapter 15 we saw that the physics of Maxwell's equations could also be expressed in terms of differential equations for the scalar and vector potentials:

$$\nabla^2\phi - \frac{1}{c^2} \frac{\partial^2\phi}{\partial t^2} = -\frac{\rho}{\epsilon_0}, \quad (20.4)$$

$$\nabla^2A - \frac{1}{c^2} \frac{\partial^2A}{\partial t^2} = -\frac{j}{\epsilon_0 c^2}. \quad (20.5)$$

If ρ and j are zero, these equations take on the simpler form

$$\nabla^2\phi - \frac{1}{c^2} \frac{\partial^2\phi}{\partial t^2} = 0, \quad (20.6)$$

$$\nabla^2A - \frac{1}{c^2} \frac{\partial^2A}{\partial t^2} = 0. \quad (20.7)$$

In free space the scalar potential ϕ and each component of the vector potential A all satisfy the same mathematical equation. Suppose we let ψ (phi) stand for any one of the four quantities ϕ, A_x, A_y, A_z ; then we want to investigate the general solutions of the following equation:

$$\nabla^2\psi - \frac{1}{c^2} \frac{\partial^2\psi}{\partial t^2} = 0. \quad (20.8)$$

This equation is called the three-dimensional wave equation (three-dimensions), because the function ψ may depend in general on x, y , and z , and we need to worry about variations in all three coordinates. This is made clear if we write out explicitly the three terms of the Laplacian operator:

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2\psi}{\partial t^2} = 0. \quad (20.9)$$

In free space, the electric fields E and B also satisfy the same equation. For example, since $B = \nabla \times A$, we can get a differential equation for B by taking the curl of Eq. (20.7). Since the Laplacian is a scalar operator, the order of the Laplacian and curl operations can be interchanged:

$$\nabla \times (\nabla^2A) = \nabla^2(\nabla \times A) = \nabla^2B.$$

Similarly, the order of the operations curl and $\partial/\partial t$ can be interchanged:

$$\nabla \times \frac{1}{c^2} \frac{\partial^2A}{\partial t^2} = \frac{1}{c^2} \frac{\partial^2}{\partial t^2}(\nabla \times A) = \frac{1}{c^2} \frac{\partial^2B}{\partial t^2}.$$

Using these results, we get the following differential equation for B :

$$\nabla^2B - \frac{1}{c^2} \frac{\partial^2B}{\partial t^2} = 0. \quad (20.10)$$

So each component of the magnetic field B satisfies the three-dimensional wave equation. Similarly, using the fact that $E = -\nabla\phi - dA/dt$, it follows that the electric field E in free space also satisfies the three-dimensional wave equation:

$$\nabla^2E - \frac{1}{c^2} \frac{\partial^2E}{\partial t^2} = 0. \quad (20.11)$$

All of our electromagnetic fields satisfy the same wave equation, Eq. (20.8). We might well ask: What is the most general solution to this equation? However, rather than tackling that difficult question right away, we will look first at what can be said in general about those solutions in which nothing varies in y and z . (Always do an easy case first so that you can see what is going to happen, and then you can go to the more complicated cases.) Let's suppose that the magnitudes

of \mathbf{B} fields depend only upon x —but there are no restrictions of the fields with y and z . We are, of course, considering plane waves along x . We should expect to get results something like those in the previous section. In fact, we will find precisely the same answers. You may ask: "Why do that over again?" It is important to do it again, first, because we did not show that the waves we found were the most general solutions for plane waves, and second, because we found the fields only from a very particular kind of current source. We would like to ask now: What is the most general kind of one-dimensional wave there can be in free space? We cannot find that by seeing what happens for the one-dimensional waves you must work with greater generality. Also we are going to work this time with differential equations instead of with integral forms. Although we will get the same results, it is a way of practicing back and forth to show that it doesn't make any difference which way you go. You should know how to do things every which way, because when you get a hard problem, you will often find that only one of the various ways is tractable.

We could consider directly the solution of the wave equation for some electromagnetic quantity. Instead, we want to start right from the beginning with Maxwell's equations in free space so that you can see their close relationship to the electromagnetic waves. So we start with the equations in (20.1), setting the charges and currents equal to zero. They become

$$\begin{aligned} \text{I. } \nabla \cdot \mathbf{E} &= 0 \\ \text{II. } \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \text{III. } \nabla \cdot \mathbf{B} &= 0 \\ \text{IV. } \epsilon' \nabla \times \mathbf{B} &= \frac{\partial \mathbf{E}}{\partial t} \end{aligned} \quad (20.12)$$

We write the first equation out in components:

$$\nabla \cdot \mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0. \quad (20.13)$$

We are assuming that there are no variations with y and z , so the last two terms are zero. This equation then tells us that

$$\frac{\partial E_x}{\partial x} = 0. \quad (20.14)$$

The solution is that E_x , the component of the electric field in the x direction, is a constant in space. If you look at IV in (20.12), suppose no B -variation in y and z either, you can see that E is also constant in time. Such a field would be the steady x field from some charged condenser plates a long distance away. We are not interested now in such an uninteresting static field; we are at the moment interested only in dynamically varying fields. For $dynamic$ fields, $E_x = 0$.

We have then the important result that for the propagation of plane waves in any direction, the electric field must be at right angles to the direction of propagation. It can, of course, still vary in an complicated way with the coordinate x .

The transverse E -field can always be resolved into two components, say the y -component and the z -component. So let's first work out a case in which the electric field has only one transverse component. We'll take first an electric field that is always in the y direction, with zero z -component. Equivalently, if we solve this problem we can also solve for the case where the electric field is always in the z -direction. The general solution can always be expressed as the superposition of two such fields.

How easy our equations now get. The only component of the electric field that is non-zero is E_y , and all derivatives—except those with respect to x —are zero. The rest of Maxwell's equations then become quite simple.

Let's look next at the second of Maxwell's equations [11 of Eq. (20.12)]. Writing out the components of the curl \mathbf{E} , we have

$$(\nabla \times \mathbf{E})_y = \frac{\partial E_x}{\partial y} - \frac{\partial E_y}{\partial z} = 0,$$

$$(\nabla \times \mathbf{E})_y = \frac{\partial E_x}{\partial z} - \frac{\partial E_x}{\partial y} = 0,$$

$$(\nabla \times \mathbf{E})_x = \frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}.$$

The y -component of $\nabla \times \mathbf{E}$ is zero because the derivatives with respect to y and z are zero. The x -component is also zero; the first term is zero because the derivative with respect to z is zero, and the second term is also because E_x is zero. The only non-zero component of the curl of \mathbf{E} that is not zero is the z -component, which is equal to $\partial E_z / \partial x$. Setting the three components of $\nabla \times \mathbf{E}$ equal to the corresponding components of $-\partial \mathbf{B} / \partial t$, we can conclude the following:

$$\frac{\partial B_z}{\partial t} = 0, \quad \frac{\partial B_z}{\partial z} = 0. \quad (20.15)$$

$$\frac{\partial B_z}{\partial t} = -\frac{\partial E_z}{\partial x}. \quad (20.16)$$

Since the x -component of the magnetic field and the y -component of the magnetic field both have zero time derivatives, these two components are just constant fields and correspond to the magnetostatic solutions we found earlier. Somebody may have left some permanent magnets near where the waves are propagating. We will ignore these constant fields and set B_x and B_y equal to zero.

Incidentally, we would already have concluded that the x -component of \mathbf{B} should be zero for a different reason. Since the divergence of \mathbf{B} is zero (from the third Maxwell equation), applying the same arguments we used above for the electric field, we would conclude that the longitudinal component of the magnetic field can have no variation with x . Since we are ignoring such uniform fields in our wave solutions, we would have set B_x equal to zero. In plane electromagnetic waves the B -field, as well as the E -field, must be directed at right angles to the direction of propagation.

Equation (20.16) gives us the additional propagation that if the electric field has only a y -component, the magnetic field will have only a z -component. So \mathbf{E} and \mathbf{B} are at right angles to each other. This is exactly what happened in the special wave we have already considered.

We are now ready to use the last of Maxwell's equations (or box space [IV of Eq. (20.12)]). Writing out the components, we have

$$c^2 (\nabla \times \mathbf{B})_y = c^2 \frac{\partial B_z}{\partial y} + c^2 \frac{\partial B_y}{\partial z} = \frac{\partial E_x}{\partial t},$$

$$c^2 (\nabla \times \mathbf{B})_y = c^2 \frac{\partial B_z}{\partial z} + c^2 \frac{\partial B_z}{\partial x} = \frac{\partial E_y}{\partial t}, \quad (20.17)$$

$$c^2 (\nabla \times \mathbf{B})_x = c^2 \frac{\partial B_y}{\partial x} - c^2 \frac{\partial B_x}{\partial y} = \frac{\partial E_z}{\partial z}.$$

Of the six derivatives of the components of \mathbf{B} , only the term $\partial B_z / \partial x$ is not equal to zero. So the three equations give us simply

$$-c^2 \frac{\partial B_z}{\partial x} = \frac{\partial E_z}{\partial z}. \quad (20.18)$$

The result of all our work is that only one component each of the electric and magnetic fields is not zero, and that these components linearly satisfy Eqs. (20.16) and (20.18). The two equations can be exchanged into one if we differentiate the first with respect to x and the second with respect to t ; the left-hand sides of the

two equations will then be the same (except for the factor c^2). So we find that ψ satisfies the equation

$$\frac{\partial^2 \psi}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} = 0. \quad (20.19)$$

We have seen the same differential equation before, when we studied the propagation of sound. It is the wave equation for one-dimensional waves.

You should note that in the process of our derivation we have found something more than is contained in Eq. (20.11). Maxwell's equations have given us the fundamental fact that electromagnetic waves have field components only at right angles to the direction of the wave propagation.

Let's review what we know about the solutions of the one-dimensional wave equation. If any quantity ψ satisfies the one-dimensional wave equation,

$$\frac{\partial^2 \psi}{\partial x^2} - \frac{1}{c^2} \frac{\partial^2 \psi}{\partial t^2} = 0, \quad (20.20)$$

then one possible solution is a function $\psi(x, t)$ of the form

$$\psi(x, t) \sim R(x - ct), \quad (20.21)$$

that is, some function of the single variable $(x - ct)$. The function $f(x - ct)$ represents a "rigid" pattern in x which travels toward positive x at the speed c (see Fig. 20-4). For example, if the function f has a maximum when its argument is zero, then for $t = 0$ the maximum of ψ will occur at $x = 0$. At some later time, say $t = 10$, ψ will have its maximum at $x = 10c$. As time goes on, the maximum moves toward positive x at the speed c .

Sometimes it is more convenient to say that a solution of the one-dimensional wave equation is a function of $(x - x/c)$. However, this is saying the same thing, because any function of $(x - x/c)$ is also a function of $(x - ct)$:

$$F(t - x/c) = F\left[1 - \frac{x - ct}{c}\right] = f(x - ct).$$

Let's show that $f(x - ct)$ is indeed a solution of the wave equation. Since ψ is a function of only one variable—the variable $(x - ct)$ —we shall let f' represent the first derivative of f with respect to its variable; and f'' represent the second derivative of f . Differentiating Eq. (20.21) with respect to x , we get

$$\frac{\partial \psi}{\partial x} = f'(x - ct),$$

since the derivative of $(x - ct)$ with respect to x is 1. The second derivative of ψ with respect to x is clearly

$$\frac{\partial^2 \psi}{\partial x^2} = f''(x - ct). \quad (20.22)$$

Taking derivatives of ψ with respect to t , we find

$$\begin{aligned} \frac{\partial \psi}{\partial t} &= f'(x - ct)f'(-1), \\ \frac{\partial^2 \psi}{\partial t^2} &= -c^2 f''(x - ct). \end{aligned} \quad (20.23)$$

We see that ψ does indeed satisfy the one-dimensional wave equation.

You may be wondering: "If I have the wave equation, how do I know that I should take $f(x - ct)$ as a solution? I don't like this backward picture. Isn't there some forward way to find the solution?" Well, one good forward way is to know the solution. It is possible to "cook up" an apparently forward mathematical argument, especially because we know what the solution is supposed to be, but with an equation as simple as this we don't have to play games. Soon you will get so that when you see Eq. (20.20), you nearly simultaneously see 20-6



Fig. 20-4. The function $\psi(x - ct)$ represents a constant "shape" that travels toward positive x with the speed c .

$\psi = f(x - ct)$ as a solution. (Just as you see the integral of $x^2 dx$, you know right away that the answer is $x^{1/2}$.)

Actually you should also see a little more. Not only is any function of $(x - ct)$ a solution, but any function of $(x + ct)$ is also a solution. Since the wave equation contains only c^2 , changing the sign of c makes no difference. In fact, the most general solution of the one-dimensional wave equation is the sum of two arbitrary functions, one of $(x - ct)$ and the other of $(x + ct)$:

$$\psi = f(x - ct) + g(x + ct). \quad (20.24)$$

The first term represents a wave travelling toward positive x , and the second term an a chirring wave travelling toward negative x . The general solution is the superposition of two such waves both existing at the same time.

We will leave the following interesting question for you to think about. Take a function ψ of the following form:

$$\psi = \sin kx \cos \omega t.$$

This equation isn't in the form of a function of $(x - ct)$ or of $(x + ct)$. Yet you can easily show that this function is a solution of the wave equation by direct substitution into Eq. (20.23). How can we then say that the general solution is of the form of Eq. (20.24)?

Applying our conclusions about the solution of the wave equation to the y -component of the electric field, E_y , we conclude that E_y can vary with x in any arbitrary fashion. However, the fields which do exist can always be represented as the sum of two patterns. One wave is sailing through space in one direction with speed c , with an associated magnetic field perpendicular to the electric field; another wave is travelling in the opposite direction with the same speed. Such waves correspond to the electromagnetic waves that we know about—light, radio-waves, infrared radiation, ultraviolet radiation, X rays, and so on. We have already discussed the emission of light in great detail in Vol. I. Since everything we learned there applies to any electromagnetic wave, we don't need to consider it in great detail here—the behavior of these waves.

We should perhaps make a few further remarks on the question of the polarization of the electromagnetic waves. In our solution we chose to consider the special case in which the electric field has only a y -component. There is clearly another solution for waves travelling in the plus or minus x -direction, with an electric field which has only a x -component. Since Maxwell's equations are linear, the general solution for one-dimensional waves propagating in the x -direction is the sum of waves of E_y and waves of E_x . This general solution is summarized in the following equations:

$$\begin{aligned} E &= (0, E_y, E_z) \\ E_y &= f(x - ct) + g(x + ct) \\ E_z &= f(x - ct) + g(x + ct) \\ B &= (0, B_y, B_z) \\ cB_z &= f(x - ct) - g(x + ct) \\ cB_y &= -f(x - ct) + g(x + ct). \end{aligned} \quad (20.25)$$

Such electromagnetic waves have an E -vector whose direction is not constant but which gyrates around in some arbitrary way in the yz -plane. At every point the magnetic field is always perpendicular to the electric field and to the direction of propagation.

If there are only waves travelling in one direction, say the positive x -direction, there is a simple rule which tells the relative orientation of the electric and magnetic fields. The rule is that the cross product $E \times B$ which is, of course, a vector at right angles to both E and B , points in the direction in which the wave is travelling. If E is rotated into B by a right-hand screw, the sense of going in the direction of the wave velocity. (We shall see later that the vector $E \times B$ has a special physical significance; it is a vector which describes the flow of energy in an electromagnetic field.)

20-2 Three-dimensional waves

We want now to turn to the subject of three-dimensional waves. We have already seen that the vector E satisfies the wave equation. It is also easy to arrive at the same conclusion by arguing directly from Maxwell's equations. Suppose we start with the equation

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

and take the curl of both sides:

$$\nabla \times (\nabla \times E) = -\frac{\partial}{\partial t}(\nabla \times B). \quad (20.26)$$

You will remember that the curl of the curl of any vector can be written as the sum of two terms, one involving the divergence and the other the Laplacian,

$$\nabla \times (\nabla \times E) = \nabla(\nabla \cdot E) - \nabla^2 E.$$

In free space, however, the divergence of B is zero, so only the Laplacian term remains. Also, from the fourth of Maxwell's equations in free space [Eq. (20.12)] the time derivative of $\epsilon^{-1} \nabla \times B$ is the second derivative of B with respect to t ,

$$-\epsilon^{-1} \frac{\partial}{\partial t}(\nabla \times B) = \frac{\partial^2 B}{\partial t^2}.$$

Equation (20.26) then becomes

$$\nabla^2 E = \frac{1}{\epsilon \mu} \frac{\partial^2 B}{\partial t^2},$$

which is the three-dimensional wave equation. Written out in all its glory, this equation is, of course,

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} + \frac{\partial^2 E}{\partial z^2} = \frac{1}{\epsilon \mu} \frac{\partial^2 B}{\partial t^2} = 0. \quad (20.27)$$

How shall we find the general wave solution? The answer is that all the solutions of the three-dimensional wave equation can be represented as a superposition of the one-dimensional solutions we have already found. We obtained the equation for waves which move in the x -direction by supposing that the field did not depend on y and z . Obviously, there are other solutions in which the fields do not depend on y and z , representing waves going in the y -direction. Then there are solutions which do not depend on x and y , representing waves travelling in the z -direction. Or, in general, since we have written our equations in vector form, the three-dimensional wave equation can have solutions which are plane waves moving in any direction at all. Again, since the equations are linear, we may have simultaneously as many plane waves as we wish, travelling in as many different directions. Thus the most general solution of the three-dimensional wave equation is a superposition of all sorts of plane waves moving in all sorts of directions.

Try to imagine what the electric and magnetic fields look like at present in the space in this lecture room. First of all, there is a steady magnetic field; it comes from the currents in the wires in the earth—that is, the earth's steady magnetic field. Then there are some irregular, nearly static electric fields produced perhaps by electric charges generated by friction as various people move about in their

charge and only their ends sheared against the chain atoms. Then there are other magnetic fields produced by oscillating currents in the electrified wires—fields which vary in frequency, ν , cycles per second, in synchronism with the generator or Hertzian (Maxwell) field. But more interesting are the electric and mag. field varying at much higher frequencies. For instance, as light leaves from window to floor and wall to wall, there are little wiggles of the electric and magnetic fields moving along at 186,000 miles per second. Then there are also infrared waves travelling from the waterbedheads to the cold blackboard. And we have forgotten the ultraviolet light, λ -rays, and the radiowaves travelling through the room.

Flying across the room are electromagnetic waves which carry movies of a jacobean. There are waves oscillated by a series of impulses representing pictures of events going on in other parts of the world, or of inspiration aspirine dissolving in imaginary stomachs. To demonstrate the reality of these waves it is only necessary to turn on electronic equipment that converts these waves into pictures and sounds.

If we go into further detail to analyze even the smallest wiggles, there are tiny electromagnetic waves that have traveled to the room from enormous distances. There are now tiny oscillations of the electric field, whose crests are separated by a distance of one foot, that have come from millions of miles away. (Remember to the earth from the Mariner 11 space craft which has just passed Venus. Its signals carry summaries of information it has picked up about the planets (information obtained from electromagnetic waves that travelled from the planet to the space craft).

There are very tiny wiggles of the electric and magnetic fields, but see waves which originated billions of light years away—from galaxies in the remotest corners of the universe. That this is true has been found by “filling the room with wires” by building antennas as large as this room. Such radiowaves have been detected from places in space beyond the range of the greatest optical telescopes. Even they, the optical telescopes, are simply detectors of electromagnetic waves. What we call the stars are only inferences, inferences drawn from the only planet readily we have yet visited, Earth. From them—from a careful study of the exceedingly complex undulations of the electric and magnetic fields reaching us on Earth.

There is, of course, more: the fields produced by lightning bolts, rays, the fields of the charged cosmic ray particles as they zip through the room, and more, and more. What a complicated thing is the electric field in the space around you! Yet it always satisfies the three dimensional wave equation.

20-3 Scientific imagination

I now appeal you to imagine these electric and magnetic fields. What do you do? Do you know how? How do I imagine the electric and magnetic field? What do I actually see? What are the demands of scientific imagination? Is it any different from trying to imagine that the world is full of invisible angels? No, it is not like imagining invisible angels. It requires a yet a higher degree of imagination to understand the electromagnetic field than to understand invisible angels. Why? Because to make invisible angels understandable, all I have to do is to alter their invisibility a little bit—I make them slightly visible, and then I can see the shapes of their wings, and bodies, and faces. Once I succeed in imagining a visible angel, the achievement required—which is to take almost invisible angels and imagine them completely invisible—is relatively easy. So you say, “Professor, please give me an approximate description of the electromagnetic waves, even though it may be slightly inaccurate, so that I too can see them, as well as I can see almost invisible angels. Then I will modify the picture to the necessary characteristic.”

I’m sorry I wait, the first for you. I don’t know how I have no picture of this electromagnetic field that is in any sense accurate. I have known about the electromagnetic field a long time—I was in the same position 20 years ago that you are now, and I have had 20 years more of experience thinking about these wiggling waves. When I start describing the magnetic field moving through space, I speak of the E - and B fields and wave my arms and you sit by listening that I can see them.

I'll tell you what I see: I see some kind of vague shadowy, wiggling lines—here and there is an E and B written on them somehow, and perhaps some of the lines have arrows on them—an arrow here or there which disappears when I look too closely at it. When I talk about the fields swishing through space, I have a visible connection between the symbols I use to describe the objects and the objects themselves. I cannot really make a picture that is even nearly like the true waves. So if you have some difficulty in making such a picture, you should not be worried that your difficulty is unusual.

Our science makes terrible demands on the imagination. The degree of imagination that is required is much more extreme than that required for some of the ancient ideas. The modern ideas are much harder to imagine. We use a lot of tools, though. We use mathematical equations and rules, and make a lot of pictures. What I usually do is that when I talk about the electric and the field in space, I see some kind of a superposition of all of the diagrams which I've ever seen drawn about them. I don't see little bundles of field lines running about, because it worries me that if I run at a different speed, the bundles would disperse. I can't even always see the electric and magnetic fields because sometimes I think I should have made a picture with the vector potential and the scalar potential, for those were perhaps the more physically significant things that were wiggling.

Perhaps the only hope, you say, is to take a mathematical view. Now, what is a mathematical view? From a mathematical view, there is an electric field vector and a magnetic field vector at every point in space; that is, there are six numbers associated with every point. Can you imagine six numbers associated with each point in space? That's too hard. Can you imagine even one number associated with every point? I cannot! I can imagine such a thing as the temperature to a, every point in space. That seems to be understandable. There is a business and evidence that varies from place to place. But I honestly do not understand the idea of a number at every point.

So perhaps we should put the question: Can we represent the electric field by something more like a temperature, say like the displacement of a piece of jelly? Suppose that we were to begin by imagining that the world was filled with this jelly and that the fields represented some distortion, say a stretching or twisting—of the jelly. Then we could visualize the field. After we "see" what it is like we could extract the jelly away. For many years that's what people tried to do. Maxwell, Faraday, Faraday, and others tried to understand electromagnetism this way. (Sometimes they called the abstract jelly "ether.") But it turned out that the attempt to imagine the electromagnetic field in that way was really standing in the way of progress. We are unfortunately limited to abstractions, to using instruments to detect the field, to using mathematical symbols to describe the field, etc. But nevertheless, in some sense the fields are real, because after we are all finished fiddling around with mathematical equations—with or without making pictures and drawings and trying to visualize the thing—we can still make the instruments detect the signals from Mars or from the sun about galaxies a billion miles away, and so on.

The whole question of imagination in science is often misunderstood by people in other disciplines. They try to test our imagination in the following way. They say, "Here is a picture of some people in a situation. What do you imagine will happen next?" When we say, "I can't imagine," they may think we have a weak imagination. They overlook the fact that whatever we are allowed to imagine in science must be consistent with everything else we know; that the electric fields and the waves we talk about are not just some happy thoughts which we are free to state as we wish, but ideas which must be consistent with all the laws of physics we know. We can't allow ourselves to seriously imagine things which are obviously inconsistent with the known facts of nature. And so our kind of imagination is quite a difficult game. One has to have the imagination to think of something that has never been seen before, never been heard of before. At the same time the thoughts are restricted in a straitjacket, so to speak, limited by the conditions that come from our knowledge of the way nature really is. The problem of creating

something which is new, but which is consistent with everything which has been seen before, is one of extreme difficulty.

While I'm on this subject I would talk about whether it will ever be possible to imagine beauty that we can't see. It is an interesting question. When we look at a rainbow, it looks beautiful to us. Everybody says, "Ooh, a rainbow." If you see how scientific I am, I am afraid to say something is beautiful unless I have an experimental way of defining it. But how would we describe a rainbow if we were blind? We would find when we measure the infrared reflection coefficient of sodium chloride, or when we talk about the frequency of the waves that are coming from some galaxy that we can't see - we make a diagram; we make a plot. For instance, for the rainbow, what it probably is the intensity of the light vs. wavelength measured with a spectrophotometer for each direction in the sky. Eventually, such measurements would give a curve that was rather flat. Then some day, someone would discover that for certain directions of the weather, and at certain angles in the sky, the spectrum of intensity as a function of wavelength would behave strangely; it would have a hump. As the angle of the instrument was varied only a little, the maximum of the hump would move from one wavelength to another. Then one day the physical review of the blind men might publish a technical article with the title "The Intensity of Radiation as a Function of Angle under Various Conditions of the Weather." In this article there might appear a graph such as the one in Fig. 20-5. The author would perhaps remark that at the larger angles there was a large variation of wavelength, whereas at the smaller angles the maximum of the radiation moved at shorter wavelengths. From our point of view, we would say that the light at 43° is predominantly green and the light at 42° is predominantly red.

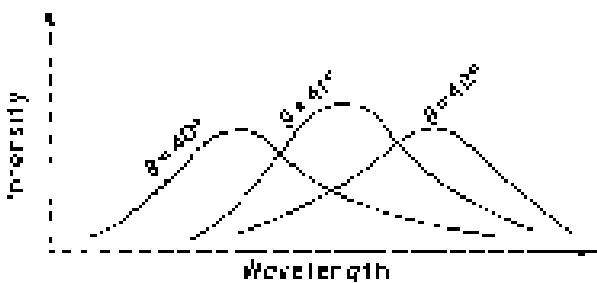


Fig. 20-5. The intensity of electromagnetic waves as a function of wavelength for three angles (measured from the direction opposite the sun, observed only with certain meteorological conditions).

Now if we find the graph of Fig. 20-5 beautiful, I don't know exactly what detail that we appreciate when we look at a rainbow, because our eyes cannot see the exact details in the shape of a spectrum. The eye, however, finds the rainbow beautiful. Do we have enough imagination to see in the spectrum of 1938's sun the beauty we see when we look directly at the sun here? I don't know.

If I suppose I have a graph of the reflection coefficient of a sodium chloride crystal as a function of wavelength in the infrared, and also as a function of angle, I would have a representation of how it would look to my eyes if they could see in the infrared - perhaps some glowing, shiny "Frodo," indeed with reflections from the surfaces of the galaxies! That would be a beautiful thing. But I don't know whether I can see such a graph of the reflection coefficient of NaCl measured with axes I can't see, and say just it has the same beauty.

On the other hand, even if we compare two to an particular microwave source, we can already claim to see a certain beauty in the equations which describe general physical laws. For example, in the wave equation, (20.6), there's a constant, proportional to the reciprocal of the square root of the x , the vector r , and $i\omega/c$. And this constant, being proportional to the x , y , z , and t , suggests to me that such a great beauty must lie in doing the four dimensions, the possibility that space has time, or simultaneous motion, the possibility of analyzing that and the developments of the special theory of relativity. So there is plenty of intellectual beauty associated with the equations.

20-4 Spherical waves

We have seen that there are solutions of the wave equation which correspond to plane waves, and that any electromagnetic wave can be described as a superposition of many plane waves. In certain specific cases, however, it is more convenient to describe the wave field in a different mathematical form. We would like to discuss how the theory of spherical waves, waves which correspond to spherical surfaces that are spreading out from some center. When you drop a stone into a lake, the ripples spread out in regular waves on the surface. They are two-dimensional waves. A spherically wave is a similar thing except that it spreads out in three dimensions.

Before we start describing spherical waves, we need a little mathematics. Suppose we have a function that depends only on the radial distance r from a certain origin—in other words, a function that is spherically symmetric. Let's call the function $\psi(r)$, r being ρ we mean

$$r = \sqrt{x^2 + y^2 + z^2},$$

the radial distance from the origin. In order to find out what functions $\psi(r)$ satisfy the wave equation, we will need an expression for the Laplacian of ψ . So we want to find the sum of the second derivatives of ψ with respect to x , y , and z . We will use the notation that $\psi'(r)$ represents the derivative of ψ with respect to r and $\psi''(r)$ represents the second derivative of ψ with respect to r .

First, we find the derivatives with respect to x . The first derivative is

$$\frac{\partial \psi(r)}{\partial x} = \psi'(r) \frac{\partial}{\partial x}.$$

The second derivative of ψ with respect to x is

$$\frac{\partial^2 \psi}{\partial x^2} = \psi''(r) \left(\frac{\partial r}{\partial x} \right)^2 + \psi' \frac{\partial^2 r}{\partial x^2}.$$

We can evaluate the partial derivatives of r with respect to x from

$$\frac{\partial r}{\partial x} = \frac{x}{r}, \quad \frac{\partial^2 r}{\partial x^2} = \frac{1}{r} \left(1 - \frac{x^2}{r^2} \right).$$

So the second derivative of ψ with respect to x is

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{x^2}{r^2} \psi''(r) + \frac{1}{r} \left(1 - \frac{x^2}{r^2} \right) \psi'. \quad (20.29)$$

Similarly,

$$\frac{\partial^2 \psi}{\partial y^2} = \frac{y^2}{r^2} \psi''(r) + \frac{1}{r} \left(1 - \frac{y^2}{r^2} \right) \psi', \quad (20.30)$$

$$\frac{\partial^2 \psi}{\partial z^2} = \frac{z^2}{r^2} \psi''(r) + \frac{1}{r} \left(1 - \frac{z^2}{r^2} \right) \psi'. \quad (20.31)$$

The Laplacian is the sum of these three derivatives. Remembering that $x^2 + y^2 + z^2 = r^2$, we get

$$r^2 \psi''(r) = \psi''(r) + \frac{2}{r} \psi'(r). \quad (20.32)$$

It is often more convenient to write this equation in the following form:

$$r^2 \psi'' = \frac{1}{r} \frac{d^2}{dr^2} (r \psi). \quad (20.33)$$

If you carry on the differentiation indicated in Eq. (20.33), you will see that the right-hand side is the same as in Eq. (20.31).

If we wish to consider spherically symmetric fields which can propagate as spherical waves, our field quantity must be a function of both r and θ . Suppose

we ask, then, what functions $\psi(r, t)$ are solutions of the three-dimensional wave equation

$$\nabla^2 \psi(r, t) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \psi(r, t) = 0. \quad (20.33)$$

Since $\psi(r, t)$ depends only on the spatial coordinates, letting r_1 denote the separation from the \hat{x} -axis we find (above, Eq. (20.32)) that the separation between r_1 and r_2 , since ψ is also a function of r , we could write the derivatives with respect to r as partial derivatives. Then the wave equation becomes

$$\frac{1}{r} \frac{\partial^2}{\partial r^2} (\psi r) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} (\psi r) = 0.$$

We must now solve this equation, which appears to be much more complicated than the plane wave case. But notice that, if we multiply this equation by r , we get

$$\frac{\partial^2}{\partial r^2} (r\psi) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} (r\psi) = 0. \quad (20.34)$$

This equation tells us that the function $r\psi$ satisfies the one-dimensional wave equation in the variable r . Using the general principle which we have emphasized so often, that the same equations always have the same solutions, we know that if ϕ is a function only of $(r + ct)$ the ϕ will be a solution of Eq. (20.34). So we know that spherical waves must have the form

$$r\psi(r, t) = f(r - ct).$$

Or, as we have seen before, we can equally well say that $r\psi$ can have the form

$$rk + f(t - r/c).$$

Dividing by r , we find that the field quantity ψ (whatever it may be) has the following form:

$$\psi = \frac{f(t - r/c)}{r}. \quad (20.35)$$

Such a function represents a general spherical wave travelling outward from the origin at the speed c . If we forget about the r in the denominator for a moment, the amplitude of the wave as a function of the distance from the origin at a given time has a certain shape that travels outward at the speed c . The factor r in the denominator, however, says that the amplitude of the wave decreases in proportion to $1/r$ as the wave propagates. In other words, unlike a plane wave in which the amplitude remains constant as the wave runs along, in a spherical wave the amplitude steadily decreases, as shown in Fig. 20-6. This effect is easy to understand from a simple physical argument.

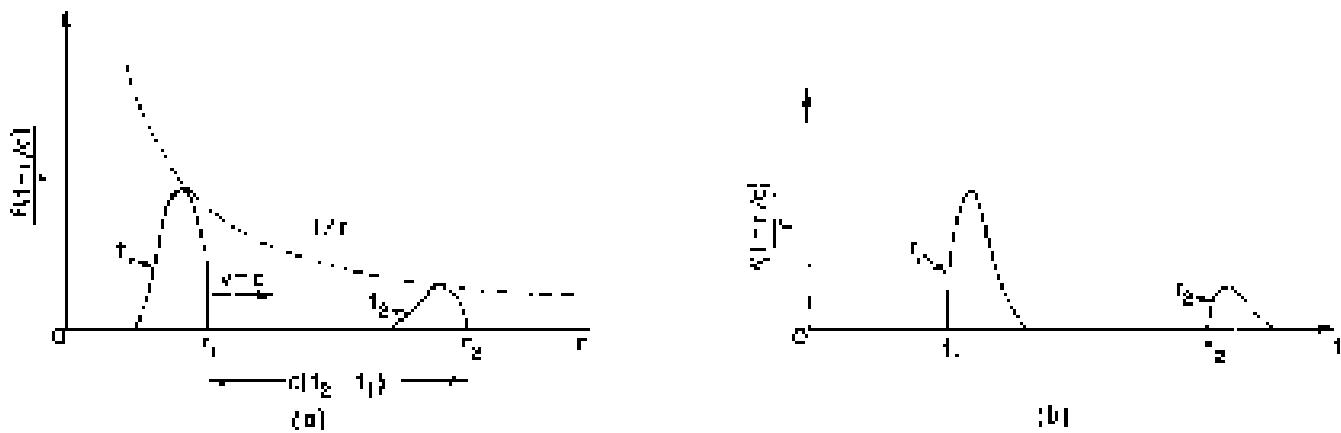


Fig. 20-6. A spherical wave $\psi = f(t - r/c)/r$: (a) ψ as a function of r for $r > r_1$ and the same wave for two later times t_2 ; (b) ψ as a function of t for $r = r_1$ and the same wave after t_1 .

We know that total energy density in a wave depends on the square of the wave amplitude a . As the wave spreads, its energy is spread over larger and larger areas proportional to the radial distance squared. If the total energy is conserved, the energy density must fall as $1/r^2$, and the amplitude of the wave must decrease as $1/r$. So Eq. (20.35) is the "reasonable" form for a spherical wave.

We have discussed the second possible solution to the one-dimensional wave equation:

$$\psi \sim g(r) + h(r),$$

or

$$\psi \sim g(r) + \frac{h(r)}{r}.$$

This also represents a spherical wave, but one which travels toward from large r toward the origin.

We are now going to make a special assumption. We say, without any demonstration whatsoever, that the waves generated by a source are only the waves which go outward. Since we know that waves are caused by the motion of charges, we would like think that the source projects nothing from the charges. It would be quite strange to imagine that before charges were set in motion, a spherical wave emitted out from infinity and arrived at the charges just at the time they began to move. This is a possible solution, but experience shows that when charges are accelerated the waves travel outward from the charges. Although Maxwell's equations would allow either possibility, we will put in an additional fact based on experience: that only the outgoing wave solution makes "physical sense."

We should realize, however, that there is an interesting consequence to this additional assumption: we are changing the symmetry with respect to time that exists in Maxwell's equations. The original equations for E and B , and also the wave equations we derived from them, have the property that if we change the sign of t , the equation is unchanged. These equations say that for every solution corresponding to a wave going in one direction there is an equally valid solution for a wave traveling in the opposite direction. Our statement that we will consider only the outgoing, spherical waves is an important additional assumption. (A formulation of electrodynamics in which this additional assumption is avoided has been recently studied. Surprisingly, in many circumstances it does not lead to physically absurd conclusions, but it would take us too far away to discuss these ideas, not now. We will talk about them a little more in Chapter 28.)

We are going to encounter a important point. In our solution for an outgoing wave, Eq. (20.35), the function ψ is infinite at the origin. That is not really perfect. We would like to find a wave solution which is smooth everywhere. Our solution must represent physically a situation in which there is some source at low energy. In other words, we have inadvertently made a mistake. We have not solved the free wave equation (20.35) everywhere; we have solved Eq. (20.35) with zero on the right everywhere, except at the origin. Our mistake crept in because some of the steps in our derivation are not "legal" when $r = 0$.

Let's show that it is easy to make the same kind of mistake in an electrostatics problem. Suppose we want a solution of the equation for an electrostatic potential in free space $\nabla^2 \phi = 0$. The Lagrangian is equal to zero, because we are assuming that there are no charges anywhere. But what about a spherically symmetric solution to this equation—that is, some function ϕ that depends only on r . Using the formula of Eq. (22.12) for the Laplacian, we have

$$\frac{1}{r} \frac{d^2}{dr^2} (\phi r) = 0.$$

Multiplying this equation by r , we have an equation which is readily integrated:

$$\frac{d^2}{dr^2} (\phi r) = 0.$$

If we integrate once with respect to r , we find that the first derivative of ϕr is a constant. If

constant, which we may call a :

$$\frac{d}{dr}(\alpha) = 0.$$

Integrating again, we find that α is of the form

$$r\phi = ar + b,$$

where b is another constant of integration. So we have found that the following ϕ is a solution for the electrostatic potential in free space:

$$\phi = a + \frac{b}{r}.$$

Somewhat is evidently wrong. In the region where there are no electric charges, we know the solution for the electrostatic potential: the potential is everywhere a constant. That corresponds to the first term in our solution. But we also have the second term, which says that there is a contribution to the potential that varies as one over the distance from the origin. We know, however, that such a potential corresponds to a point charge at the origin. So, although we thought we were solving for the potential in free space, our solution also gives the field for a point source at the origin. Do you see the similarity between what happened now and what happened when we solved for a spherically symmetric solution to the wave equation? If there were really no charges or currents at the origin, there would not be spherical outgoing waves. The spherical waves must, of course, be produced by sources at the origin. In the next chapter we will investigate the connection between the outgoing & incoming electromagnetic waves and the currents and voltages which produce them.

Solutions of Maxwell's Equations with Currents and Charges

21-1 Light and electromagnetic waves

We saw in the last chapter that solving their equations, Maxwell's equations have waves of electric and magnetic. These waves correspond to the phenomena of radio, light, x-rays, and so on, depending on the wavelength. We have already said a little in great detail in Vol. I, but one step we want to take toward the two subjects we want to show that Maxwell's equations can indeed form the basis for our further treatment of the phenomena of light.

When we studied light, we began by writing down an equation for the electric field produced by a charge which moves in any arbitrary way. That equation was

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2} \left[\hat{\mathbf{e}}_{\perp} + \frac{c}{r} \frac{d}{dt} \left(\hat{\mathbf{e}}_{\perp} \right) - \frac{1}{c^2} \frac{d^2}{dr^2} \hat{\mathbf{e}}_{\perp} \right] \quad (21.1)$$

$$\mathbf{B} = \hat{\mathbf{e}}_{\perp} \times \mathbf{E}.$$

[See Eq. (28.3), Vol. 1.]

If a charge moves in an arbitrary way, the electric field \mathbf{E} , would find out at some point depends only on the position and motion of the charge at now, but at an earlier time— t' , an instant which is earlier by the time it would take light, going at the speed c , to travel the distance r' from the charge to the field point. In other words, if we want the electric field at point (1) at the time t , we must calculate the location (2') of the charge and its motion at the time $(t - r'/c)$, where r' is the distance to the point (1) from the position of the charge (2') at the time $(t - r'/c)$. The prime is to remind you that r' is the "separation" (distance) from the point (2) to the point (1), and not the actual distance between point (2'), the position of the charge at the time t' , and the field point (1) [see Fig. 21.1]. Note that we are using a different convention here for the direction of the unit vector $\hat{\mathbf{e}}$. In Chapters 28 and 36 of Vol. I it was convenient to take $\hat{\mathbf{e}}$ round, hence $\hat{\mathbf{e}}$ pointing toward the source. Now we are following the definition we took for Coulomb's law, in which $\hat{\mathbf{e}}$ is directed from the charge, at (2'), toward the field point at (1). The only difference, of course, is that our new $\hat{\mathbf{e}}$ (and \mathbf{e}) are the negatives of the old ones.

We have also seen that if the velocity v of a charge is always much less than c , and if we consider only points at large distances from the charge, so that only the last term of Eq. (21.1) is important, the fields can also be written as

$$\mathbf{E} = -\frac{q}{4\pi\epsilon_0 c r'^2} \left[\text{acceleration of the charge at } (t - r'/c) \right] \hat{\mathbf{e}}_{\perp} \quad (21.1')$$

and

$$\mathbf{B} = \hat{\mathbf{e}}_{\perp} \times \mathbf{E}$$

Let's look at what the complete equation, Eq. (21.1), says in a little more detail. The vector $\hat{\mathbf{e}}_{\perp}$ is the unit vector to point (1) from the "earlier" position (2'). The first term, then, is what we would expect for the Coulomb field of the charge at its retarded position—we may call this "the retarded Coulomb field." The electric field depends inversely on the square of the distance r' and directly from the retarded position of the charge (the $\hat{\mathbf{e}}_{\perp}$ is in the direction of \mathbf{E}).

But let's look at the first term. The $d/\frac{dt}{dt}$ term tells us that the laws of electricity do not say that all the fields stay the same as the static ones, but instead change (which is what people sometimes like to say). To the "retarded Coulomb field" we must

21-1 Light and electromagnetic waves

21-2 Spherical waves from a point source

21-3 The general solution of Maxwell's equations

21-4 The fields of an oscillating dipole

21-5 The potentials of a moving charge; the general solution of Lienard and Wiechert

21-6 The potentials for a charge moving with constant velocity; the Lorentz formula

Review of Chapter 28, Vol. I, Electromagnetic Radiation
 Chapter 31, Vol. I, The Origin of the Reference Axis
 Chapter 36, Vol. I, Relativistic Effects on Radiation

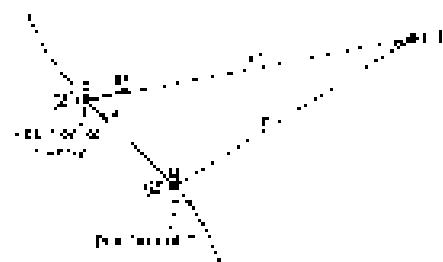


Fig. 21.1. The fields at (1) at the time t depend on the position (2') occupied by the charge q at the time $(t - r'/c)$.

add the other two terms. The second term says that there is a "correction" to the retarded Coulomb field which is the rest of charge of the retarded Coulomb field multiplied by c/ϵ_0 , the retardation delay. In a way of speaking, this term tends to compensate for the retardation in the first term. The first two terms correspond to computing the "retarded Coulomb field" and then extrapolating it forward in time by the amount c/ϵ_0 , that is, "up to the date." The extrapolation is linear, as if we were to assume that the "retarded Coulomb field" would continue to change at the same rate as for the charge at the point (t) . If the field is changing slowly, the effect of the retardation is almost completely removed by the correction term, and the two terms together give us an electric field that is the "instantaneous Coulomb field," that is, the Coulomb field of the charge at the point (t) to a very good approximation.

Finally, there is a third term in Eq. (21.1) which is the second derivative of the unit vector \hat{r} . For our study of the phenomena of light, we made use of the fact that far away from the charge the first two terms went inversely as the square of the distance and, for large distances, became very weak in comparison to the last term, which was $\propto 1/r^3$. So we concentrated initially on the last term, and we know now that it is typical for large distances (proportional to the separation of the source and of the charge). Light happens to be linear of sight. (Again, for most of our work in Vol. I we took the case in which the charges were moving; see earlier stability.) We considered the relativistic effects in only one chapter, Chapter 36.

Now we should try to connect the two things together. We have just Maxwell's equations, and we have Eq. (21.1) for the field of a point source. We should certainly ask whether they are equivalent. If we can deduce Eq. (21.1) from Maxwell's equations, we will really understand the connection between light and wave propagation. To make this connection is the main purpose of this chapter.

It turns out that we won't quite make it—that the mathematical details get too complicated for us to carry through in all their gory details. But, we will get as close as enough so that you should easily see how the connection could be made. The missing pieces will only be in the mathematical details. Some of you may find the mathematics in this chapter rather complicated, and you may not wish to follow the argument very closely. We think it is important, however, to make the connection between what you have learned earlier and what you are learning now, or at least to indicate how such a connection can be made. You will notice, if you look over the earlier chapters, that wherever we have taken a statement as a starting point for a discussion, we have carefully explained whether it is a new "assumption" that is a "basic law," or whether it can ultimately be deduced from some other laws. We hope to give you in the spirit of these lectures to make the connection between light and Maxwell's equations. If it gets difficult in places, well, that's life. There is no other way.

21-2 Spherical waves from a point source

In Chapter 18 we found that Maxwell's equations could be solved by letting

$$E = -\nabla \phi + \frac{\partial A}{\partial t} \quad (21.7)$$

$$\text{and} \quad B = \nabla \times A, \quad (21.8)$$

where ϕ and A must then be solutions of the equations

$$\nabla^2 \phi = \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = -\frac{\rho}{\epsilon_0} \quad (21.4)$$

$$\text{and} \quad \nabla^2 A = \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} = -\frac{j}{\mu_0 c^2}, \quad (21.5)$$

and must also satisfy the condition that

$$\nabla \cdot A = -\frac{1}{c^2} \frac{\partial \phi}{\partial t}. \quad (21.6)$$

Now we will find the solution of Eqs. (21.4) and (21.5). To do that we have to find the solution ψ of the equation

$$\nabla^2\psi + \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} = -\epsilon, \quad (21.7)$$

where ϵ , which we can take to be negative, is known. Of course, a consequence of $\epsilon < 0$ in Eq. (21.4) is $\epsilon = -\epsilon_0 r^2$ if ψ is to be finite, but we want to solve Eq. (21.7) as a most general problem and consider what ψ and ϵ are physically.

In place of ψ where a and b are zero—in what we have called “free” space—the potentials ϕ and ψ , and the fields E and B , all satisfy the three-dimensional wave equation without sources, whose “einsteinian” form is

$$\nabla^2\psi = \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2} = 0. \quad (21.8)$$

In Chapter 20 we saw that solutions of this equation can represent waves of various kinds: plane waves in the \hat{z} direction, $\psi = f(z) = e^{ikz}$; plane waves in the \hat{x} or \hat{y} direction, or in any other direction; or spherical waves of the form

$$\psi(r, \theta, \phi) = \frac{f(r - r/c)}{r}. \quad (21.9)$$

(The solutions can be written in still other ways, for example cylindrical waves that spread out from an axis.)

We also remarked that, physically, Eq. (21.9) does not represent a wave in free-space—that there must be charges at the origin to get the outgoing wave started. In other words, Eq. (21.9) is a solution of Eq. (21.8) everywhere except right near $r = 0$, where it cannot be a solution of the complete equation (21.7), including sources. Let's see how that works. What kind of a source ϵ in Eq. (21.7) would give rise to a wave like Eq. (21.9)?

Suppose we have the spherical wave of Eq. (21.9) and look at what's happening for very small r . Then the retardation $-r/c$ in $f(r - r/c)$ can be neglected—provided f is a smooth function—and ψ becomes

$$\psi \sim \frac{f(0)}{r} \quad (r \rightarrow 0). \quad (21.10)$$

So ψ is just like a Coulomb field for a charge at the origin that varies with $1/r$. That is, if we had a truly lumpy bit of charge, bounded in a very small region near the origin, with a density σ , we know that

$$\phi = \frac{Q/4\pi\epsilon_0}{r},$$

where $Q = \int \rho dV$. Now we know that such a ϕ satisfies the equation

$$\nabla^2\phi = -\frac{\rho}{\epsilon_0},$$

Following the same mathematics, we would say that the ψ of Eq. (21.10) satisfies

$$\nabla^2\psi = -\epsilon \quad (r \rightarrow 0), \quad (21.11)$$

where ϵ is related to f by

$$f = \frac{S}{4\pi},$$

with

$$S = \int \sigma dV.$$

The only difference is that in the general case, ψ and, therefore, S , can be a function of time.

Now the important thing is that if ψ satisfies Eq. (21.11) for small r , it also satisfies Eq. (21.7). As we go very close to the origin, the $1/r$ dependence of ψ

causes the space derivatives to become very large. But the time derivative is zero there since sources (1) are just point charges at rest. So we get to zero, the term $\epsilon_0 \nabla^2 \phi_0$ in Eq. (21.7) as the right-hand integral in Eq. (21.5), and Eq. (21.7) becomes equivalent to Eq. (21.11).

To summarize, then, if the source function ϕ_0 of Eq. (21.7) is located at the origin and has the value ϕ_0 , then

$$\Phi(t) = \int \phi(t') dV, \quad (21.12)$$

the solution of Eq. (21.7) is

$$\phi(x_1, y_1, z_1, t) = \frac{1}{4\pi} \frac{\delta(t - t')}{r}, \quad (21.13)$$

The only effect of the term $\epsilon_0 \nabla^2 \phi_0$ in Eq. (21.7) is to introduce the retardation $t - t'$ even in the Coulomb-like potential.

21-5 The general solution of Maxwell's equations

We have found the solution of Eq. (21.2) for a "point" source—the last question is: What is the solution for a piecewise source? That is, why we can think of any source $\phi(x_1, y_1, z_1, t)$ as made up of the sum of many "point" sources, each with a finite element of dV , and each with the source strength $\phi_0(x_1, y_1, z_1)$? Since Eq. (21.2) is linear, the resultant field is the superposition of the fields from all of such source elements.

Using the results of the preceding section (Eq. (21.12)) we know that the field ϕ_0 at the point (x_2, y_2, z_2) at (1) for short r at the time t from a source element with a dipole $\phi_0(x_1, y_1, z_1)$ at (2) for short r' is given by

$$d\phi(1, t) = \frac{\phi_0(x_1, t')}{4\pi r'^2},$$

where r_{12} is the distance from (2) to (1). Adding the contributions from all the pieces of the source means, of course, taking an integral over all regions where $r' \neq 0$; so we have

$$\phi(1, t) = \int \frac{\phi_0(x_1, t')}{4\pi r_{12}^2} dV_2. \quad (21.14)$$

That is, the field at (1) at the time t is the sum of all the spherical waves which leave the source elements at (2) at the times $t' = |x_1 - x_2|/c$. This is the solution of our wave equation for any set of sources.

We are now how to obtain a general solution of Maxwell's equations. If by ϕ we mean the scalar potential ϕ , the wave function ψ becomes ϕ_0 , etc. Or we can let ψ represent any one of the three components of the vector potential A , depending by the corresponding component of $\partial \phi_0 / \partial t$. Thus, if we knew the charge density $\rho(x_1, y_1, z_1)$ and the current density $j(x_1, y_1, z_1, t)$ everywhere, we can immediately write down the solutions of Eqs. (21.9) and (21.5). They are

$$\phi(1, t) = \int \frac{\rho(x_1, t' - r_{12}/c)}{4\pi r_{12}^2} dV_2, \quad (21.15)$$

and

$$A(1, t) = \int \frac{j(x_1, t' - r_{12}/c)}{4\pi c r_{12}^2} dV_2. \quad (21.16)$$

The fields E and B can then be found by differentiating the potentials using Eqs. (21.2) and (21.3). [Incidentally, it is possible to verify that the ϕ and A obtained from Eqs. (21.15) and (21.16) do satisfy the equality (21.6).]

We have solved Maxwell's equations. Given the currents and charges in any circumstance, we can find the potential ϕ directly by a few integrals and then differentiate and get the fields. So we have finished off the Maxwell theory. At the moment we close the ring back to our theory of light, because to connect with our earlier work on light, we need only calculate the electric field from a

moving charge. All that remains is to take a moving charge, calculate the potential due to it, then integrate, and then differentiate to find E from $-\nabla\phi = \epsilon_0 E/\rho$. We would get Eq. (21.13). It turns out to be lots of work, but that's the principle.

So here is the center of the universe of electromagnetism—the complete theory of electricity and magnetism, and of light; a complete description of the fields produced by any moving charges; and more. It is all here. Here is the structure built by Maxwell, complete in all its power and beauty. It is probably one of the greatest accomplishments of physics. To remind you of its importance, we will put it all together in a nice frame.

Maxwell's equations:

$$\nabla \cdot E = \frac{\rho}{\epsilon_0} \quad \nabla \cdot B = 0$$

$$\nabla \times E = -\frac{\partial B}{\partial t} \quad \nabla \times B = \frac{J}{\epsilon_0} + \frac{\partial E}{\partial t}$$

Their solutions:

$$E = -\nabla\phi - \frac{\partial A}{\partial t}$$

$$B = \nabla \times A$$

$$\phi(r, t) = \int \frac{q(1, t - r_{12}/c)}{4\pi\epsilon_0 r_{12}} dV_1$$

$$A(r, t) = \int \frac{q(2, t - r_{12}/c)}{4\pi\epsilon_0 r_{12}} dV_2$$

21-4 The fields of an oscillating dipole

We have still not lived up to our promise to derive Eq. (21.1) for the electric field of a point charge in motion. Even with the results we already have, it is a relatively complicated thing to do. We have not found Eq. (21.1) anywhere in the influential literature except in Vol. I of these lectures.* So you can see that it is not easy to derive. (The fields of a moving charge have been written in many other forms that are equivalent, of course.) We will have to limit ourselves here just to showing that, in a few examples, Eqs. (21.15) and (21.16) give the same results as Eq. (21.1). First, we will show that Eq. (21.1) gives the correct fields with only the restriction that the motion of the charged particle is nonrelativistic. (This, the special case will take care of 99 percent, or more, of what we said about light.)

We consider a situation in which we have a blob of charge that is moving about in some way, in a small region, and we will find the fields far away. To put it another way, we are finding the field at any distance from a point charge that is shaking up and down in very small motion. Since light is usually emitted from neutral objects such as atoms, we will consider that our wiggling charge has greater near equal and opposite charges at rest. If the separation between the centers of the charges is a , the charges will have a dipole moment $\mathbf{p} = qa$, which we take to be a function of time. Now we should expect that if we look at the fields close to the charges, we won't have to worry about the delay: the electric field will be exactly the same as the one we have calculated earlier for an electro-dipole

* This formula was worked out by R. P. Feynman, in about 1990, and given in his lectures as a good way of thinking about synchrotron radiation.

—using, of course, the instantaneous dipole moment $\rho(t)$. But if we go very far out, we ought to find a term in the field that goes as $1/c$ and depends on the acceleration of the charge perpendicular to the line of sight. Let's see if we get such a result.

We begin by calculating the vector potential A , using Eq. (21.16). Suppose that our moving charge is in a small blob whose charge density is given by $\rho(x, y, z)$, and the wave function is harmonic at any time and with the velocity v . Then the current density $j(x, y, z, t)$ will be equal to $v \rho(x, y, z)$. It will be convenient to take our coordinate system so that the z -axis is in the direction of v ; then the geometry of our problem is as shown in Fig. 21-2. We want the integral

$$\int \frac{R^2(t) - r^2/3}{r^3} dV_2. \quad (21.17)$$

Now if the size of the charge blob is really very small compared with r_{12} , we can set the r_{12} term in the denominator equal to r , the distance to the center of the blob, and take it outside the integral. Next, we are also going to set $r_{12} = r$ in the numerator, although that is not really quite right. It is not right because we should take r at, say, the top of the blob and slightly different time than we used for just the bottom of the blob. When we set $r_{12} = r$ in $j(t) = v \rho(x, y, z)$, we are taking the current density for the whole blob at the same time ($t = \tau/v$). That is an approximation that will be good only if the velocity v of the charge is much less than c . So we are making a nonrelativistic calculation. Replacing r by $r\tau$, the integral (21.17) becomes

$$\frac{1}{r} \int \rho(t) (1 - r/\tau)^2 dV_2.$$

Since all the charge has the same velocity, this integral is just v/c times the total charge ρ . But ρ is just $d\rho/dt$, the rate of change of charge content, which is, of course, to be evaluated at the instant $t = \tau/v$. We will write it as $\rho(t - \tau/v)$. So we get for the vector potential

$$A(t, r) = \frac{1}{4\pi c r^2} \frac{\rho(t - \tau/v)}{v}. \quad (21.18)$$

Our result says that the current in a varying dipole produces a vector potential in the form of spherical waves whose source strength is $\rho/16\pi c^2 v^2$.

We can now get the magnetic field from $B = \nabla \times A$. Since ρ is really in the z -direction, A has only a z -component; there are only two directions perpendicular to the curl. So $B_x = i \partial A_y / \partial z$ and $B_z = -i \partial A_y / \partial x$. Let's first look at B_x ,

$$B_x = \frac{i A_y}{\partial z} = \frac{1}{4\pi c r^2} \frac{\partial \rho(t - \tau/v)}{\partial y}. \quad (21.19)$$

Every time on the calculation, we must remember that $x = \sqrt{r^2 + y^2 + z^2}$, so

$$B_x = \frac{1}{4\pi c r^2} \rho(t - \tau/v) \frac{\partial}{\partial y} \left(\frac{1}{\sqrt{r^2 + y^2}} \right) = \frac{1}{16\pi c^2 r^3} \frac{\partial \rho(t - \tau/v)}{\partial y}. \quad (21.20)$$

Remembering that $y/\partial y = -v/r$, the first term gives

$$\frac{1}{4\pi c r^2} \frac{\rho(t - \tau/v)}{r^3}, \quad (21.21)$$

which is proportional to $1/r^3$ like the fields of a static dipole (because ρ/t is constant for a given direction).

The second term in Eq. (21.20) gives us the new effect. Carrying out the differentiation, we get

$$= \frac{1}{4\pi c r^2} \frac{\partial}{\partial z} \rho(t - \tau/v), \quad (21.22)$$

where τ means, of course, the second derivative of ρ with respect to t . The term,

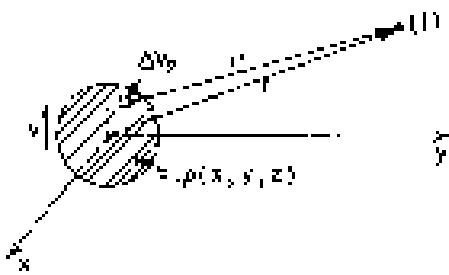


Fig. 21-2. The potentials at (1) are given by integrals over the charge density ρ .

which comes from differentiating the monopole, is responsible for radiation. First, it describes a field which decreases with distance only as $1/r$. Second, it depends on the acceleration \ddot{r} . That's why it's important to see how we're going to get a result like Eq. (21.1), which requires the radiation velocity.

Let's examine in a little more detail how this radiation term looks about—it is such an interesting and important result. We start with the expression (21.18), which has a $1/r$ dependence and is therefore like a Coulomb potential, except for the delay term in the numerator. Why is it true that, when we differentiate with respect to space around nodes to get the fields, we don't just get a $1/r^2$ field—well, of course, the source has loop time delays?

We can see why it is following way: Suppose that we let our dipole oscillate up and down in a sinusoidal motion. Then we would have

$$\vec{p} = \vec{p}_0 + p_0 \sin \omega t$$

and

$$J_r = \frac{1}{4\pi c r^2} \frac{\omega p_0 \cos \omega t - \dot{p}_0}{r}.$$

If we plot a graph of J_r as a function of r at a given instant, we get the curve shown in Fig. 21-5. The peak amplitude decreases as $1/r$, but there is, in addition, an oscillation in space, bounded by the $1/r$ envelope. When we take the spatial derivatives, they will be proportional to the slope of the curve. In fact the figure we see that there are slopes much steeper than the slope of the $1/r$ envelope itself. This, in fact, is clear, that for a given frequency the peak slopes are proportional to the amplitude of the wave, which varies as $1/r$. So that explains the drop-off rate of the radiation term.

It all comes about because the variations with time at the source are transformed into variations in space as the waves are propagated outward, and the magnetic fields depend on the spatial derivatives of the potential.

Let's go back and finish our calculation of the magnetic field. We have for B_r the two terms (21.21) and (21.22), so

$$B_r = \frac{1}{4\pi c \mu_0 r^2} \left[\frac{p_0 \theta(r - r_0)}{r^2} - \frac{p_0 \theta'(r - r_0)}{r^2} \right]$$

With the same kind of mathematics, we get

$$B_r = \frac{1}{4\pi c \mu_0 r^2} \left[\frac{\partial^2 \theta(r - r_0)}{r^2} - \frac{\partial \theta'(r - r_0)}{r^2} \right]$$

Or we can add it all together in a nice vector formula:

$$\vec{B} = \frac{1}{4\pi c \mu_0 r^2} \left[\vec{p} + (\vec{v}_0 c) \vec{s}_0 \cdot \hat{r} \hat{r} \times \vec{r} \right] \quad (21.23)$$

Now let's look at this formula. First of all, if we go very far out in r , only the \vec{p} term remains. The direction of \vec{B} is given by $\vec{p} \times \vec{r}$, which is at right angles to the radius \vec{r} and also at right angles to the acceleration, as in Fig. 21-4. Everything is at once at right; Let's take the result, we get from Eq. (21.17)

Now let's look at what we are not used to—at what happens closer in. In Section 14-9 we worked out the law of Biot and Savart for the magnetic field of an elemental current. We found that a current element $d\vec{l}$ contributes to the magnetic field the amount

$$d\vec{B} = \frac{1}{4\pi c \mu_0 r^2} \frac{\vec{l} \times \vec{r}}{r^3} dI. \quad (21.24)$$

You see that this formula looks very much like the first term of Eq. (21.23), if we remember that \vec{p} is the current. But there is one difference. In Eq. (21.23), the current is to be evaluated at the origin ($r = r_0$), which doesn't appear in Eq. (21.24). Actually, however, Eq. (21.24) is still very good for small r , because the second

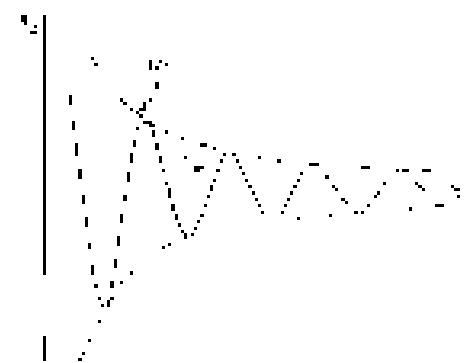


Fig. 21-5. The magnitude of A as a function of r at the instant t for the spherical wave from an oscillating dipole.

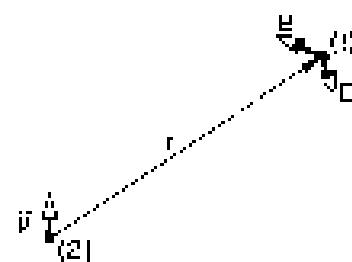


Fig. 21-4. The radiation fields B and E of an oscillating dipole.

term of Eq. (21.23) tends to cancel out the effect of the retardation in the first term. The two together give a result very near to Eq. (21.24) when r is small.

We can see that this way, when r is small, $(r - r_0c)$ is not very different from r_0 , so we can expand the bracket in Eq. (21.23) in a Taylor series. For the first term,

$$\rho(r - r_0c) = \rho(r) - \frac{1}{c} \dot{\rho}(r) + \text{etc.}$$

and to the same order in c/r ,

$$\rho(r - r_0c) = \bar{\rho}(r)$$

When we take the sum, the two terms in ρ cancel, and we are left with the zero-retarded current ρ ; that is, $\rho(r) = \rho_0(r)$ plus terms of order $(r/c)^2$ or higher (e.g., $\rho(r/c)^2 \ddot{\rho}$) which will be very small for r small enough that ρ does not alter markedly in the time c/r .

So Eq. (21.23) gives fields very much like the instantaneous theory—much closer than the instantaneous theory with a delay; the first-order effects of the delay are taken care of by the second term. The static formulas are very accurate, much more accurate than you might think. Of course, the compensation only works for points close in. For points far out the correction becomes very bad, because the terms of ρ_0 produce a very large effect, and we get the important $1/r^2$ term of the radiation.

We still have the problem of computing the electric field and current density that is the same as Eq. (21.1). For large distances we can see that the answer is going to come out all right. We know that far from the sources, where we have a propagating wave, it is perpendicular to E and also to B , as in Fig. 2.1-1, and that $eB = E$. So E is proportional to the acceleration p , as expected from Eq. (21.1).

To get the electric field completely for all distances, we need to solve for the electrostatic potential. When we compute the current integral for A to get Eq. (21.18), we make an approximation by disregarding the slight variation of r in the delay term. This will not work for the electrostatic potential, because we would then get A times the integral of the charge density, which is a constant. This approximation is too rough. We need to go to one higher order. Instead of being involved in that higher-order computation directly, we can do something else—we can determine the scalar potential from Eq. (21.6), using the vector potential we have already found. The divergence of A , in our case, is just $\partial A/\partial r$ —since A_r and A_θ are identically zero. (We're working in the same way that we did above to find B .)

$$\begin{aligned}\nabla \cdot A &= \frac{1}{4\pi\epsilon_0 c^2} \left[\rho(r - r_0c) \frac{\partial}{\partial r} \left(\frac{1}{r} \right) + \frac{1}{c} \frac{\partial}{\partial r} \rho(r - r_0c) \right] \\ &= \frac{1}{4\pi\epsilon_0 c^2} \left[-\frac{\rho_0(r)}{r^2} - \frac{\rho'(r)}{r} - \frac{\rho_0(r - r_0c)}{c r^2} \right].\end{aligned}$$

Or, in vector notation,

$$\nabla \cdot A = -\frac{1}{4\pi\epsilon_0 c^2} \frac{A_r - \rho(r)\hat{r}}{r^2}.$$

Using Eq. (21.6), we have an equation for ϕ :

$$\frac{\partial \phi}{\partial r} = \frac{1}{4\pi\epsilon_0 c^2} \frac{[p + (\rho/c)\hat{r}] - \rho(r)\hat{r}}{r^2}.$$

Integrating with respect to r just removes one \hat{r} from each of the $\rho\hat{r}$'s, so

$$\phi(r, t) = \frac{1}{4\pi\epsilon_0 c^2} \frac{[p + (\rho/c)\hat{r}] - \rho(r)\hat{r}}{r^2}. \quad (21.83)$$

(The constant of integration would correspond to some superposed static field which could, of course, exist. But the oscillating dipole we have taken, there is no static field.)

We are now able to find the electric field \mathbf{E} here:

$$\mathbf{E} = -\nabla \phi = \frac{d\phi}{dt}.$$

Since the steps are tedious but straightforward [provided you remember that $\rho(t) = r/c$ and its time derivatives depend on x , y , and z through the retardation r/c], we will just give the result:

$$\mathbf{E}(r, t) = \frac{1}{4\pi\epsilon_0 c^2 r^3} \left[-\rho^* + \frac{\rho'(t-r/c)}{r^2} \mathbf{r} + \frac{1}{c^2} (\rho(t-r/c) \times \mathbf{r}) \times \mathbf{r} \right] \quad (21.26)$$

with

$$\rho^* = \rho(t-r/c) + \frac{r}{c} \dot{\rho}(t-r/c). \quad (21.27)$$

Although it looks rather complicated, the result is easily interpreted. The vector ρ^* is the dipole moment retarded and then "corrected" for the retardation, so the two terms with ρ^* give just the static dipole field when r is small. (See Chapter 6, Eq. 16.14.) When r is large, the term in ρ dominates, and the electric field is proportional to the acceleration of the charges, at right angles to \mathbf{v} , and, in fact, directed along the projection of \mathbf{v} in a plane perpendicular to \mathbf{r} .

This result agrees with what we would have gotten using Eq. (21.1). Of course, Eq. (21.1) is more general; it works with any motion, while Eq. (21.26) is valid only for small motions for which we can take the retardation r/c as constant over the source. At any rate, we have now provided the underpinnings for our entire previous discussion of light (excepting some matters discussed in Chapter 26 of Vol. 1), for it all hangs on the last term of Eq. (21.26). We will discuss next how the fields can be obtained for more rapidly moving charges (concerning the relativistic effects of Chapter 16 of Vol. 1).

21.5 The potentials of a moving charge: the general solution of Trémaud and Whehere

In the last section we made a simplification in calculating our integral for ϕ by considering only y -axis motion. But in doing so we missed two important points and also one where it is easy to go wrong. We will therefore take up now a calculation of the potentials for a point charge moving in any way whatever— even with a relativistic velocity. Once we have this result, we will know the complete electromagnetism of electric charges. Even Eq. (21.1) can then be derived by taking derivatives. The story will be complete. So here we go.

Let's try to calculate the scalar potential $\phi(1, t)$ for the point (x_1, y_1, z_1) produced by a point charge such as an electron, moving in any manner whatsoever. By a "point" charge we mean a very small ball of charge, shrunk down as small as you like, with a charge density $\rho(x_1, y_1, z_1)$. We can find ϕ from Eq. (21.15):

$$\phi(1, t) = \frac{1}{4\pi\epsilon_0 c} \int \frac{\rho(2, t')}{r_{12}} dV_2. \quad (21.28)$$

The answer would seem to be—and almost everyone would, at first, think—but the integral of a ever-steady "point" charge is 0 , the total charge is so 0 !

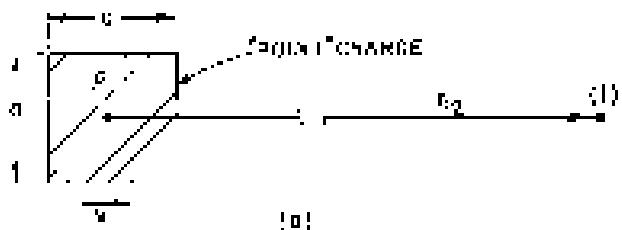
$$\phi(1, t) = \frac{1}{4\pi\epsilon_0 c} \frac{q}{r_{12}}. \quad (\text{wrong!})$$

By r_{12} we mean the radius vector from the charge at point (2) to point (1) at the retarded time $t' = r_{12}/c$. It is wrong.

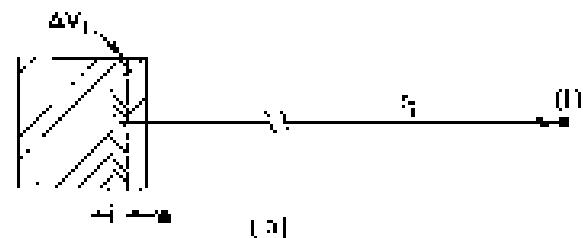
The correct answer is

$$\phi(1, t) = \frac{1}{4\pi\epsilon_0 c} \frac{q}{r_{12}} \cdot \frac{1}{1 + v_{12}/c}, \quad (21.29)$$

where v_{12} is the component of the velocity of the charge at retarded time r_{12} , namely, toward point (1). We will now show you why. To make the argument easier to



(a)



(b)

Fig. 21-5. (a) "Point" charge considered as a small cubical distribution of charge moving with the speed v toward point (1). (b) The volume element ΔV_1 used for calculating the potential.

follow, we will write the calculation first for a "point" charge which is in the form of a little cube of charge moving toward the point (1) with the speed v , as shown in Fig. 21-5(a). Let the length of a side of the cube be a , which we take to be much, much less than x_{10} , the distance from the center of the charge to the point (1).

Now to evaluate the integral of Eq. (21.28), we will return to basic principles; we will write it as the sum

$$\sum \frac{\rho_0}{r} \frac{e^2 k_e}{\epsilon_0}, \quad (2.30)$$

where a is the distance from point (1) to the s th volume element ΔV_s , and ρ_0 is the charge density $\rho_0 = \epsilon_0 E_0 / (v - c/v)$. Since $x_1 < a$, always, it will be convenient to take out ΔV_s in a form of thin, rectangular slices perpendicular to the x -axis, as shown in Fig. 21-5(b).

Suppose we start by using the volume elements ΔV_s with very thickness Δx_s much less than a . The individual elements will appear as shown in Fig. 21-6(a), where we have put in more than enough to cover the charge. But we have not shown the charge, and for a good reason. Where should we draw it? For each volume element ΔV_s , we are to take ρ at the time $t_s = (t - x_s/v)$, but since the charge is moving, it is in a different place for each volume element ΔV_s .

Let's say that we began with the volume element labeled "1" in Fig. 21-6(a), chosen so that at the time $t_1 = (t - x_1/v)$ the "back" edge of the charge occupies ΔV_1 , as shown in Fig. 21-6(b). Then when we evaluate $\rho_1/\Delta V_1$, we must use the position of the charge at the slightly later time $t_2 = (t - x_2/v)$, when the charge will be in the position shown in Fig. 21-6(c). And so on, for $\Delta V_2/\Delta V_2$, etc. Now we can evaluate the sum:

Since the thickness of each ΔV_s is Δx_s , its volume is $a^2 \Delta x_s$. Then each volume element just overlaps the charge distribution by an amount of charge Δq_s , where ρ is the density of charge within the cube, which we take to be uniform. When the distance from the charge to point (1) is large, we will make a negligible error by setting all the x 's in the denominators equal to some average value, say the retarded position t' of the center of the charge. Then the sum (21.30) is

$$\sum_s \frac{\rho a^2}{r'} \frac{\Delta x_s}{\epsilon_0},$$

where Δx_s is the Δx_s 's that overlaps the charge distribution and r' is as shown in Fig. 21-6(e). The sum is, already,

$$R \frac{\rho a^2}{r'} = \frac{a^2}{r'} \left(\frac{Mv}{\sigma} \right),$$

Note ρa^2 is the total charge per unit x and Mv is the length R shown in part (c) of the figure. So what's new?

$$e = \frac{q}{4\pi\epsilon_0 R} \left(\frac{R}{a} \right) \quad (2.31)$$

What is δt ? It is the length of the path of charge increased by the distance moved by the charge between $t_1 = (t - r_1/c)$ and $t_2 = (t - r_2/c)$ —which is the distance the charge moves in the time

$$\Delta t = t_2 - t_1 = (r_2 - r_1)/c = b/c.$$

Since the speed of the charge is v , the distance moved is $v\Delta t = v/b$. But the length b is this distance added to a :

$$b = a + \frac{v}{c}b.$$

Substituting b , we get

$$b = \frac{a}{1 - (v/c)}.$$

Of course, by v we mean the velocity at the retarded time $t' = (t - r'/c)$, which we can indicate by writing $t' = v/v_{rel}$, and Eq. (21.21) for the potential becomes

$$\phi(t, \vec{r}) = \frac{q}{4\pi\epsilon_0 c^2 t'} - \frac{1}{(v \cdot \vec{r})_{rel}}.$$

This result agrees with our assertion, Eq. (21.20). There is a correction term which comes from the fact that the charge moving as it moves “sweeps over the charge.” When the charge is moving toward the point C , its contribution to the integral is increased by the ratio b/a . It reduces the overall integral of q/r' multiplied by b/a , which is $1/(1 - v/c)^2$.

If the velocity of the charge is not directed toward the observation point C , you can see just what matters is the component of the velocity toward point C . Calling this velocity component v_{rel} , it is related to v by $v = v_{rel}/\cos\theta$. Also, the analysis we have made goes exactly the same way for $v < c$ (up to the condition of no source—it doesn’t have to be a cube). Finally, since the “size” of the charge q doesn’t enter into the final result, the same result holds when we let the charge shrink to any size—even to a point. The general result is that the scalar potential for a point charge moving with any velocity is

$$\phi(t) = \frac{q}{4\pi\epsilon_0 c^2 (1 - (\vec{v} \cdot \vec{r}))_{rel}}. \quad (21.22)$$

This equation is often written in the equivalent form

$$\phi(\vec{r}, t) = \frac{q}{4\pi\epsilon_0 c^2} \frac{1}{(\vec{v} \cdot \vec{r})_{rel}}, \quad (21.23)$$

where \vec{r} is the vector from the charge to the point (t) , where ϕ is being evaluated, and all the quantities in the bracket are to have their values at the retarded time $t' = t - r'/c$.

The same thing happens when we compute \vec{A} for a point charge, from Eq. (21.16). The retarded density is p_0 and the integral $\int d\tau$ is the same as we found for ϕ . The scalar potential is

$$A(t, \vec{r}) = \frac{q\vec{v}_0}{4\pi\epsilon_0 c^2 t'} - \frac{\vec{v}_0}{(v \cdot \vec{r})_{rel}}. \quad (21.24)$$

The potentials for a point charge were first deduced in this form by Lienard and Wiechert and are called the Lienard-Wiechert potentials.

To close the ring back to Eq. (21.1) it’s only necessary to compute E and B from these potentials (using $\vec{E} = \nabla \times \vec{A}$ and $\vec{B} = -\nabla \phi - (1/c)\partial \vec{A}/\partial t$). It is now only an exercise. The arithmetic, however, is fairly involved, so we will not write out the details. Perhaps you will trust our word for it that Eq. (21.1) is equivalent to the Lienard-Wiechert potentials we have derived.¹⁷

¹⁷ If you have a lot of paper and time you can try to work it through yourself. We would then make two suggestions. First, don’t forget that the derivatives of \vec{r} are complicated, since it is a function of t' . Second, don’t try to derive Eq. (21.1) but carry out the derivatives on it and then compare what you get with the \vec{E} obtained from the potentials (2.230) and (2.232).

21-6 The potentials for a charge moving with constant velocity: the Lorentz formula

We want next to use the Liénard-Wiechert potentials for a special case: to find the fields of a charge moving with uniform velocity in a straight line. We will do it again later, using the principle of relativity. We already know what the potentials are when we are standing in the rest frame of a charge. When the charge is moving, we can figure everything out by a relativistic transformation from one system to the other. But relativity had its origin in the theory of electricity and magnetism. The formulas of the Lorentz transformation (Chapter 15, Vol. I) were discoveries made by Lorentz when he was studying the equations of electricity and magnetism. So that you can appreciate where things have come from, we would like to show that the Maxwell equations do lead to the Lorentz transformation. We begin by calculating the potentials of a charge moving with uniform velocity, directly from the electrodynamics of Maxwell's equations. We have shown that Maxwell's equations lead to the potentials for a moving charge that we get in the last section. So when we get these potentials, we are up to Maxwell's theory.

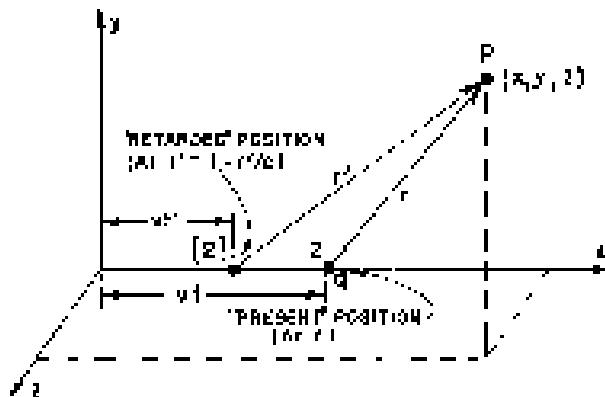


Fig. 21-7. Finding the potentials ϕ & ψ of a charge moving with uniform velocity v along the x -axis.

Suppose we have a charge moving along the x -axis with the speed v . We want the potentials at the point (x, y, z) , as shown in Fig. 21-7. If $t = 0$ is the instant when the charge is at the origin, at $(0, 0, 0)$, then the charge is at $(x - vt, y, z - 0)$. What we want to know, however, is its position at the retarded time

$$t'' = t - \frac{r'}{c}, \quad (21.35)$$

where r' is the distance from the point P to the x -axis, r , of the retarded time. At the retarded time t'' , the charge was at $(x - vt', y, 0)$.

$$r' = \sqrt{(x - vt')^2 + y^2} = \sqrt{x^2 + y^2}. \quad (21.36)$$

To find v' or r' we have to combine this equation with Eq. (21.35). First, we eliminate r' by solving Eq. (21.35) for t' and substituting in Eq. (21.36). Then, squaring both sides, we get

$$c^2(r - vt)^2 = (x - vt')^2 + y^2 + z^2,$$

which is a quadratic equation in t' . Expanding the squared binomials and collecting like terms in t' , we get

$$(c^2 - v^2)r^2 - 2vtv + c^2vt' + x^2 + v^2 - z^2 - (vt')^2 = 0$$

Solving for t' ,

$$\left(1 - \frac{v^2}{c^2}\right)t' = \frac{v^2}{c^2} - \frac{1}{c} \sqrt{(x^2 + y^2 + z^2) - \left(\frac{c^2 - v^2}{c^2}\right)(r^2 - x^2)}. \quad (21.37)$$

To get ψ we have to substitute this expression for r' into

$$r' = c(t - t').$$

Now we are ready to find ϕ from Eq. (21.33), which, since ρ is constant, becomes

$$\phi(x, y, z, t) = \frac{q}{4\pi\epsilon_0} \frac{1}{r^2 - (y^2 + z^2/c^2)}, \quad (21.38)$$

The component of v in the direction of r' is $c \propto (x - ct)/r'$, so $v \cdot r'$ is just $c \propto (x - ct)$, and the whole denominator is

$$r'(t - t') = \sqrt{(x - ct)^2 + v^2} = c \left[t - \frac{v^2}{c^2} - \left(1 - \frac{v^2}{c^2} \right) t' \right].$$

Substituting for $(1 - v^2/c^2)t'$ from Eq. (21.37), we get for ϕ

$$\phi(x, y, z, t) = \frac{q}{4\pi\epsilon_0} \frac{1}{\sqrt{(x - ct)^2 + \left(1 - \frac{v^2}{c^2} \right)(y^2 + z^2)}}$$

This equation is more understandable if we rewrite it as

$$\phi(x, y, z, t) = \frac{q}{4\pi\epsilon_0} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}} \left[\sqrt{\frac{(x - ct)^2}{\left(\sqrt{1 - v^2/c^2} \right)^2 + y^2 + z^2}} \right]^{1/2}} \quad (21.39)$$

The vector potential A is the same expression with an additional factor of cv/c^2 :

$$A = \frac{v}{c^2} \phi.$$

In Eq. (21.39) you can easily see the beginning of the Lorentz transformation. If the charge were at the origin in its own rest frame, its potential would be

$$\phi(x, y, z) = \frac{q}{4\pi\epsilon_0} [x^2 + y^2 + z^2]^{-1/2}$$

We are seeing it in a moving coordinate system, and it appears that the coordinates should be transformed by

$$\begin{aligned} x &\rightarrow \frac{x - vt}{\sqrt{1 - v^2/c^2}}, \\ y &\rightarrow y, \\ z &\rightarrow z. \end{aligned}$$

That is, the Lorentz transformation, and what we have done is essentially the way Lorentz discovered it.

But what about that extra factor $1/\sqrt{1 - v^2/c^2}$ that appears at the front of Eq. (21.39)? Also, how does the vector potential A appear when it is everywhere zero in the rest frame of the particle? We will soon show that A and ϕ together constitute a four-vector, like the momentum p and the total energy E of a particle. The extra $1/\sqrt{1 - v^2/c^2}$ in Eq. (21.39) is the same factor that always comes in when one transforms the components of a four-vector—just as the charge density ρ transforms to $\rho/\sqrt{1 - v^2/c^2}$. In fact, it is almost apparent from Eqs. (21.1) and (21.3) that A and ϕ are components of a four-vector, because we have already shown in Chapter 13 that j and ρ are the components of a four-vector.

Later we will take up in more detail the relativity of electrodynamics; here we only wished to show how naturally the Maxwell equations lead to the Lorentz transformation. You will not, then, be surprised to find that the laws of electricity and magnetism are already correct for Einstein's relativity. We will not have to "fix them up," as we had to do for Newton's laws of mechanics.

22. Circuits

22-1 Impedances

Most of our work in this course has been aimed at rendering the complete equations of Maxwell. In the last two chapters we have been discussing the consequences of these equations. We have found that the equations contain all the static phenomena we had worked out earlier, as well as the phenomena of electromagnetic waves and light that we had gone over in some detail in Volume I. The Maxwell equations give both phenomena, depending upon whether one computes the fields close to the currents and charges, or very far from them. There is not much interesting to say about the intermediate region; no special phenomena appear there.

There still remain, however, several subjects in electromagnetism that we want to take up. We can start by discussing the question of causality and the Maxwell equations: what happens when one looks at the Maxwell equations with respect to moving coordinate systems. There is also the question of the conservation of energy in electromagnetic systems. Then there is the basic subject of the electromagnetic properties of materials: so far, except for the study of the properties of dielectrics, we have considered only the electromagnetic fields in free space. And although we covered the subject of light in some detail in Volume I, there are still a few things we would like to do again from the point of view of the field equations.

In particular, we want to take up again the subject of the index of refraction, particularly for dense materials. Finally, there are the phenomena associated with waves confined in a limited region of space. We touched on this kind of problem briefly when we were studying sound waves. Maxwell's equations lead also to solutions which represent confined waves of the electric and magnetic fields. We will take up this subject, which has important technical applications, in some of the following chapters. In order to lead up to that subject, we will begin by considering the properties of electrical circuits at low frequencies. We will then be able to make a comparison between those situations in which the almost static approximations of Maxwell's equations are applicable and those situations in which high frequency effects are dominant.

So we descend from the great and esoteric heights of the last few chapters and turn to the relatively low-level subject of electrical circuits. We will see, however, that even such a mundane subject, when looked at in sufficient detail, can contain great complexities.

We have already discussed some of the properties of electrical circuits in Chapters 23 and 25 of Vol. I. Now we will cover some of the same material again, but in greater detail. Again we are going to consider only *ac* linear systems and with voltages and currents which all vary sinusoidally; we can then represent all voltages and currents by complex numbers, using the exponential notation described in Chapter 22 of Vol. I. Thus a time-varying voltage $V(t)$ will be written

$$V(t) = V e^{i\omega t}, \quad (22.1)$$

where V represents a complex number that is independent of t . It is, of course, understood that the actual time-varying voltage $V(t)$ is given by the real part of the complex function on the right-hand side of the equation.

22-1 Impedances

22-2 Generators

22-3 Networks of ideal elements: Kirchhoff's rules

22-4 Equivalent circuits

22-5 Energy

22-6 A ladder network

22-7 Filters

22-8 Other circuit elements

Review: Chapter 22, Vol. I, Appendix
Chapter 21, Vol. I, Resonance
Chapter 18, Vol. I, Linear
Systems and Review

Similarly, all of our other time-varying quantities will be taken to vary sinusoidally at the same frequency ω . So we write

$$\begin{aligned} I &= \hat{I} e^{i\omega t} \quad (\text{current}), \\ \Phi &= \hat{\Phi} e^{i\omega t} \quad (\text{flux}), \\ E &= \hat{E} e^{i\omega t} \quad (\text{electric field}), \end{aligned} \quad (22.2)$$

and so on.

Most of the time we will write our equations in terms of V, I, Φ, \dots instead of in terms of $\hat{V}, \hat{I}, \hat{\Phi}, \dots$, remembering though that the time variations are as given in (22.2).

In our earlier discussion of circuits we assumed that such things as inductances, capacitances, and resistances were familiar to you. We will now look in a little more detail at what is meant by these idealized circuit elements. We begin with the inductance.

An inductance is made by winding many turns of wire in the form of a coil and bringing the two ends out to terminals a and b some distance from the coil, as shown in Fig. 22.1. We want to assume that the magnetic field produced by current I in the coil does not spread out strongly all over space and instead is in other parts of the circuit. This is usually arranged by winding the coil in a doughnut-shaped form, or by confining the magnetic field by winding the coil on a suitable iron core, or by placing the coil in some suitable metal box, as indicated schematically in Fig. 22.1. In any case, we assume that there is a negligible magnetic field in the external region near the terminals a and b . We are also going to assume that we can neglect any electrical resistance in the wire of the coil. Finally, we will assume that we can neglect the amount of electrical charge that appears on the surface of a wire in building up the electric fields.

With all these approximations we have what we call the "ideal" inductance. (We will see what happens when things are not as simple.) For our ideal inductance we say that the voltage across the terminals is equal to $L(dI/dt)$. Let's see why that is so. When there is a current flowing through the inductor, a magnetic field proportional to the current is built up inside the coil. If the current changes with time, the magnetic field also changes. In general, the sum of it is equal to

$\partial B / \partial t$ at a , just like every, the line integral of E all the way around any closed path is equal to $-i$ times ∂ of the total magnetic flux Φ through the loop. Now suppose we consider the following path: Begin at terminal a and go along the coil (going always inside the wire) to terminal b , the current then goes back to terminal a through the air in the space outside the inductor. The line integral of E around this closed path can be written as the sum of two parts:

$$\int E \cdot ds = \int_{\text{coil}}^a E \cdot ds + \int_b^a E \cdot ds. \quad (22.3)$$

As we have seen before, there can be no electric fields inside a perfect conductor. (The smallest fields would produce infinite currents.) Therefore the integral from a to b via the coil is zero. The whole contribution to the line integral of E comes from the path outside the inductor from terminal b to terminal a . Since we have assumed that there are no magnetic fields in the space outside of the "box," this part of the integral is independent of the path chosen and we can define the potential's of the two terminals. The difference of these two potentials is what we call the voltage difference, or simply the voltage V , so we have

$$V = \int_a^b E \cdot ds = - \int_b^a E \cdot ds$$

The complete line integral is what we have before called the electromotive force ϵ and is, of course, equal to the rate of change of the magnetic flux in the coil. We have seen earlier that this ϵ is equal to the negative rate of change of Φ :

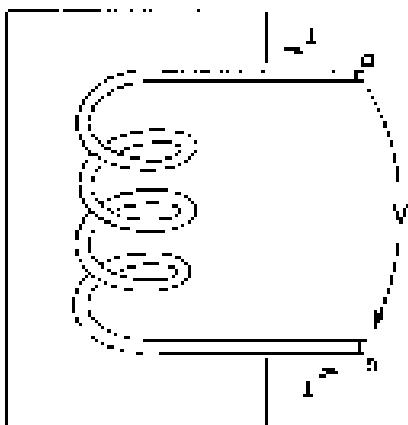


Fig. 22.1. An inductance.

the current, so we have

$$V = -i = L \frac{di}{dt},$$

where L is the inductance of the coil. Since $di/dt = idL$, we have

$$V = idL. \quad (22.4)$$

The way we have described the ideal inductor illustrates the general approach to other real circuit elements usually called "lumped" elements. The properties of the element are described completely in terms of currents and voltages that appear at the terminals. By making suitable approximations, it is possible to ignore the great complexities of the fields that appear inside the object. A separation is made between what happens inside and what happens outside.

For all the circuit elements we will find a relation like the one in Eq. (22.1), in which the voltage is proportional to the current, with a proportionality constant that is, in general, a complex number. This complex coefficient of proportionality is called the *impedance* and is usually written as z (not to be confused with the z -coordinate). It is, in general, a function of the frequency ω . So for any lumped element we write

$$\frac{V}{I} = \frac{\dot{V}}{\dot{I}} = z. \quad (22.5)$$

For an inductor, we have

$$\text{inductance} = z_L = idL. \quad (22.6)$$

Now let's look at a capacitor from the same point of view.* A capacitor consists of a pair of conducting plates from which two wires are brought out to suitable terminals. The plates may be of any shape whatsoever, and are often separated by some dielectric material. We'll draw such a situation schematically in Fig. 22-2. Again we make several simplifying assumptions. We assume that the plates and the wires are perfect conductors. We also assume that the insulation between the plates is perfect, so that no charges can flow across the insulation from one plate to the other. Next, we assume that the two conductors are close to each other but far from all others, so that all field lines which leave one plate end up on the other. Then there are always equal and opposite charges on the two plates and the charges on the plates are much larger than the charges on the surfaces of the lead-in wires. Finally, we assume that there are no magnetic fields close to the capacitor.

Suppose now we consider the line integral of E around a closed loop which starts at terminal a , goes along inside the wires to the top plate of the capacitor, jumps across the space between the plates, passes from the lower plate to terminal b through the wire, and returns to terminal a in the space outside the capacitor. Since there is no magnetic field, the line integral of E around this closed path is zero. The integral can be broken down into three parts:

$$\oint E \cdot d\ell = \int_{\text{wires}} E \cdot d\ell + \int_{\text{space}} E \cdot d\ell + \int_{\text{plates}} E \cdot d\ell. \quad (22.7)$$

The integral along the wires is zero, because there are no electric fields inside perfect conductors. The integral from b to a outside the capacitor is equal to the negative of the potential difference between the terminals. Since we imagined that the two plates are in some way isolated from the rest of the world, the total charge on

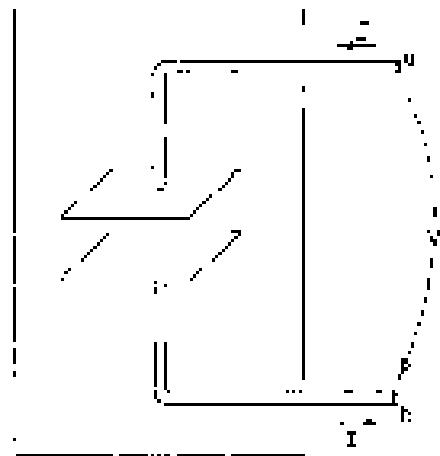


Fig. 22-2. A capacitor (or condenser).

* There are people who say we should call the objects by the names "inductor" and "capacitor" and call the properties "inductance" and "capacitance" (by analogy with "resistor" and "resistance"). We would rather let the words you will hear in the laboratory. Most people will say "inductance" for both the physical entity and its definition. If the word "capacitor" seems to have caught on, though, you will still hear "capacitance" fairly often, and vice versa still pretty frequently if "inductance" is "capacitance."

the two plates must be zero; if there is a charge Q on the upper plate, there is an equal, opposite charge $-Q$ on the lower plate. We have seen earlier that if two conductors have equal and opposite charges plus and minus Q , the potential difference between the plates is equal to Q/C , where C is called the capacity of the two conductors. From Eq. (22.7) the potential difference between the terminals a and b is equal to the potential difference between the plates. We have, therefore, that:

$$V = \frac{Q}{C}.$$

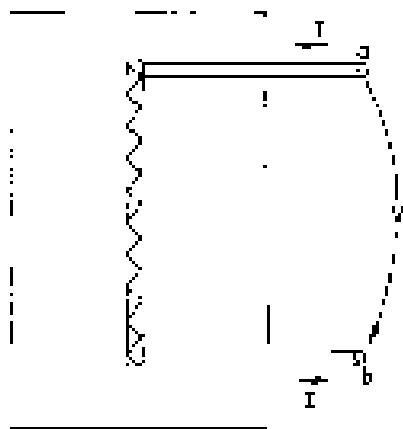


Fig. 22-3. A resistor.

The electric current I entering the capacitor through terminal a (and leaving through terminal b) is equal to dQ/dt , the rate of change of the electric charge on the plates. Writing dQ/dt as i_C , we can put the voltage equation V in analogy for a capacitor in the following way:

$$\text{for } C = \frac{Q}{V},$$

$$V = \frac{Q}{i_C C}. \quad (22.5)$$

The impedance z of a capacitor is then

$$z(\text{capacitor}) = z_C = \frac{1}{i_C C}. \quad (22.6)$$

The third element we want to consider is a resistor. However, since we have not yet discussed the electrical properties of real materials, we are not yet ready to talk about what happens inside a real conductor. We will just have to accept as fact that electric fields can exist inside real materials, that these electric fields give rise to a flow of electric charge—that is, to a current—and that this current is proportional to the integral of the electric field from one end of the conductor to the other. We then imagine a ideal resistor symbolized as in the diagram of Fig. 22-3. Two wires which we assume to be perfect conductors go from the terminals a and b to the two ends of a bar of resistive material. Following our usual line of argument, the potential difference between the terminals a and b is equal to the line integral of the external electric field, which is also equal to the line integral of the electric field through the bar of resistive material. It then follows that the current I through the resistor is proportional to the terminal voltage V :

$$I = \frac{V}{R},$$

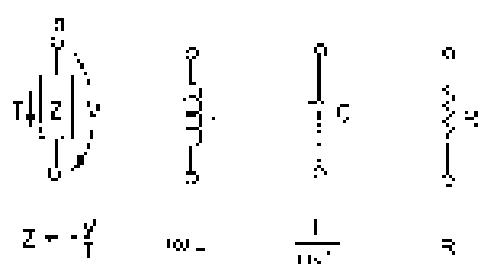


Fig. 22-4. The four lumped circuit elements (passive).

where R is called the resistance. We will see later that the relation between the current and the voltage for real conducting materials is only approximately linear. We will also see that the approximate proportionality is expected to be independent of the frequency of variation of the current and voltage only if the frequency is not too high. For alternating currents, then, the voltage across a resistor is proportional to the current, which means that the impedance is a real number:

$$z(\text{resistor}) = z_R = R. \quad (22.7)$$

Our results for the three lumped circuit elements—the inductor, the capacitor and the resistor—are summarized in Fig. 22-4. In the inductor as well as in the presenting case, we have indicated the voltage by an arrow that is directed from terminal a to terminal b . If the voltage is "positive"—that is, if the terminal a is the higher potential than the terminal b —the arrow indicates the direction of a positive "voltage drop."

Although we are talking about alternating currents, we can of course include the special case of circuits with steady currents by taking ω to be the frequency $\omega = 0$. At zero frequency—that is, for DC—the impedance of an inductor goes to zero, there is a short circuit, but of course the impedance of a capacitor

goes to infinity it becomes an open circuit. Since the impedance of a resistor is independent of frequency, it is the only element left when we analyze a circuit for us.

In the circuit elements we have described so far, the current and voltage are proportional to each other. If one is zero, so also is the other. We usually think in terms like these: An applied voltage is "responsible" for the current, or a current "gives rise to" a voltage across the terminals; so in a sense the elements "respond" to the "applied" external conditions. For this reason these elements are called *passive* elements. They can thus be contrasted with the active elements, such as the generator, we will consider in the next section, which are the *source*s of the oscillating currents or voltages in a circuit.

22-2 Generators

Now we want to talk about one active circuit element—one that is a source of the currents and voltage in addition to, recently, a passive one.

Suppose that we have a coil fixed in an inductor except that it has very low losses so that we may neglect the resistive loss of its own current. This coil, however, lies in a changing magnetic field which might be produced by a rotating magnet, as sketched in Fig. 22-5. (We have seen earlier that such a rotating magnetic field could also be produced by a suitable set of coils with alternating currents.) Again we must make several simplifying assumptions. The assumptions we will make are all the ones that we discussed for the case of the inductance. In particular, we assume that the varying magnetic field is restricted to a definite region in the vicinity of the coil and does not appear outside the generator in the space between the terminals.

Following closely the analysis we made for the inductance, we consider the line integral of \mathbf{B} around a complete loop that starts at terminal *a*, goes through the coil to terminal *b* and returns to its starting point in the space between the two terminals. Again we conclude that the potential difference between the terminals is equal to the total line integral of \mathbf{B} around the loop:

$$V_{ab} = \oint \mathbf{B} \cdot d\mathbf{l}$$

This line integral is equal to the emf in the system, so the potential difference V across the terminals of the generator is also equal to the rate of change of the magnetic flux linking the coil:

$$V = \dot{\phi}_m = \frac{d}{dt} (\text{flux}). \quad (22-1)$$

For an ideal generator we assume that the magnetic flux linking the coil is determined by external conditions such as the angular velocity of a rotating magnetic field and is not influenced in any way by the currents through the generator. Thus a "generator," at least the ideal generator we are considering, is not an impedance. The potential difference across its terminals is determined by the chirality assigned to the coil terminals *a* and *b*. Such an ideal generator is represented by the symbol shown in Fig. 22-6. The little arrow represents the direction of the emf when it is positive. A positive emf in the generator of Fig. 22-6 will produce a voltage $V = \dot{\phi}_m$ with the terminal *a* at a higher potential than the terminal *b*.

There is another way to make a generator which is quite different in the inside but which is indistinguishable from the one we have just described because what happens beyond its terminals. Suppose we have a coil of wire which is rotated in a fixed magnetic field, as indicated in Fig. 22-7. We can now let magnet to indicate the presence of a magnetic field; it could, of course, be replaced by any other source of a steady magnetic field, such as an alternating coil carrying a steady current. As shown in the figure, suppose we let the rotating coil *not* move in the outside world by means of strong contacts or "slip rings." At first we are interested in the potential difference that appears across the two terminals

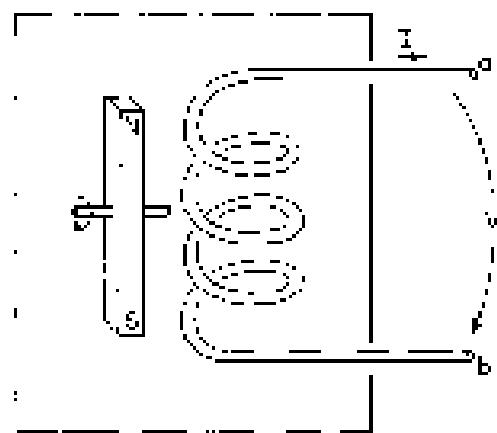


Fig. 22-5. A generator consisting of a fixed coil and a rotating magnetic field.

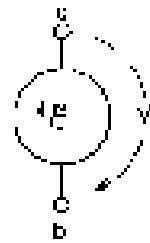


Fig. 22-6. Symbol for an ideal generator.

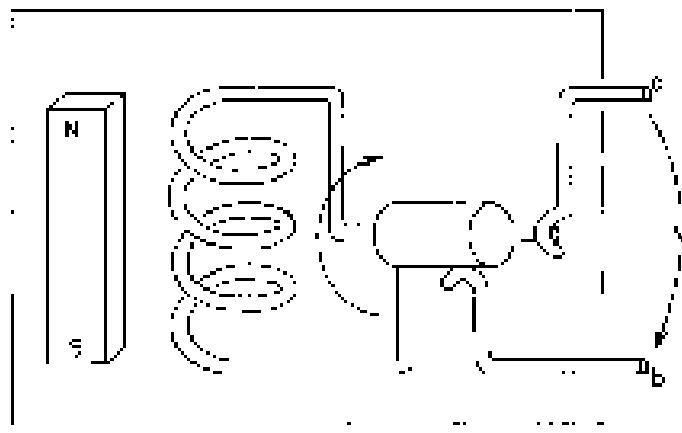


Fig. 22-7. A generator consisting of a coil rotating in a fixed magnetic field.

a and b, which is of course the integral of the electric field from terminal a to terminal b along a path outside the generator.

Now in the system of Fig. 22-7 there are no changing magnetic fields, so we might at first wonder how any voltage could appear at the generator terminals. In fact, there are no electric fields anywhere inside the generator. We are, as usual, assuming for our ideal element is that the wires inside are made of a perfectly conducting material, and as we have said many times, the electric field inside a perfect conductor is equal to zero. But that is not true. It is not true when a conductor is moving in a magnetic field. The true statement is that the total force on any charge inside a perfect conductor must be zero. Otherwise there would be an infinite flow of the free charges. So what is always true is that the sum of the electric field E and the cross product of the velocity of the conductor and the magnetic field B —which is the total force on a unit charge—must have the value zero inside the conductor:

$$\nabla \cdot E + v \times B = 0 \quad (\text{in a perfect conductor}). \quad (22.12)$$

where v represents the velocity of the conductor. (By another statement that I say is no electric field it is a perfect conductor, it will mean if the velocity v of the conductor is zero, or if even the current statement is given by Eq. (22.12).

Returning to our generator of Fig. 22-7, we know now that the line integral of the electric field E from terminal a to terminal b through the connecting path of the generator must be equal to the total force of $v \times B$ on the same path:

$$\int_{\text{path}}^b E \cdot dr = - \int_a^b (v \times B) \cdot dr. \quad (22.13)$$

It is still true, however, that the line integral of E around a complete loop, including the return from b to a outside the generator, must be zero, because there are no changing magnetic fields. So the first integral in Eq. (22.13) is also equal to V , the voltage between the two terminals. It turns out that the right-hand integral in Eq. (22.13) is just the rate of change of the flux linkage through the coil and is therefore—by the flux rule—equal to the emf in the coil. So we have again that the potential difference across the terminals is equal to the electromotive force in the circuit, in agreement with Eq. (22.11). So whether we have a generator in which a magnetic field changes near a fixed coil, or one in which a coil moves in a fixed magnetic field, the external properties of the generators are the same. There is a voltage difference V across the terminals, which is independent of the current in the circuit but depends only on the arbitrarily assigned conditions inside the generator.

So long as we are trying to understand the operation of generators from the point of view of Maxwells' equations, we might also consider the ordinary chemical cell, like a flashlight battery. It is also a generator, i.e., a voltage source, although it will of course only appear in the circuit. This simplest kind of cell is represented in Fig. 22-8. We imagine two small objects immersed in space

chemical solution. We suppose that the solution contains positive and negative ions. We suppose also that one kind of ion, say the negative, is much heavier than the rest of opposite polarity, so that its motion through the solution by the process of diffusion is much slower. We suppose next that by some means or other it is arranged that the concentration of the solution is made to vary from one part of the fluid to the other, so that the number of ions of both polarities near, say, the lower plate is much larger than the concentration of ions near the upper plate. Because of their rapid mobility the positive ions will drift more readily into the region of lower concentration, so that there will be a slight excess of positive charge arriving at the upper plate. The upper plate will become positively charged and the lower plate will have a net negative charge.

As more and more charges diffuse to the upper plate, the potential of this plate will rise until the resulting electric field between the plates produces forces on the ions which just compensate for their excess mobility, so the two plates of the cell finally reach a potential difference which is characteristic of the internal construction.

Arguing just as we did for the ideal capacitor, we see that the potential difference between the terminals *a* and *b* is just equal to the line integral of the electric field between the two plates when there is no longer any net diffusion of the ions. There is, of course, an essential difference between a capacitor and such a chemical cell. If we short-circuit the terminals of a condenser for a moment, the capacitor is discharged and there is no longer any potential difference across the terminals. In the case of the chemical cell a current can be drawn from the terminals continuously without any change in the emf—until, of course, the chemicals inside the cell have been used up. In a real cell it is found that the potential difference across the terminals decreases as the current drawn from the cell increases. In keeping with the abstractions we have been making, however, we may imagine an ideal cell in which the voltage across the terminals is independent of the current. A real cell can then be looked at as an ideal cell in series with a resistor.

22-5 Networks of ideal elements; Kirchhoff's rules

As we have seen in the last section, the description of an ideal circuit element in terms of what happens outside the element is quite simple. The current and the voltage are linearly related. But what is actually happening inside the element is quite complicated, and it is quite difficult to give a precise description in terms of Maxwell's equations. Imagine trying to give a precise description of the electric and magnetic fields of the inside of a radio which contains hundreds of resistors, capacitors, and inductors. It would be an impossible task to analyze such a thing by using Maxwell's equations. But, by making the many approximations we have discussed in Section 22-2 and summarizing the essential features of the real circuit elements in terms of idealizations, it becomes possible to analyze an electrical circuit in a relatively straightforward way. We will now show how that is done.

Suppose we have a circuit consisting of a generator and several impedances connected together, as shown in Fig. 22-9. According to our approximations there is no magnetic field in the region outside the individual circuit elements. Therefore the line integral of E around any curve which does not pass through any of the elements is zero. Consider then the curve Γ shown by the broken line which goes all the way around the circuit in Fig. 22-9. The line integral of E around this curve is made up of several pieces. Each piece is the line integral from one terminal of a circuit element to the other. This line integral we have called the voltage drop across the circuit element. The complete line integral is then just the sum of the voltage drops across all of the elements in the circuit:

$$\oint E \cdot d\tau = \sum V_L$$

Since the line integral is zero, we have that the sum of the potential differences

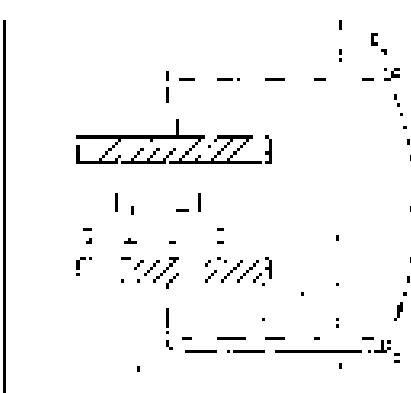


Fig. 22-8. A chemical cell.

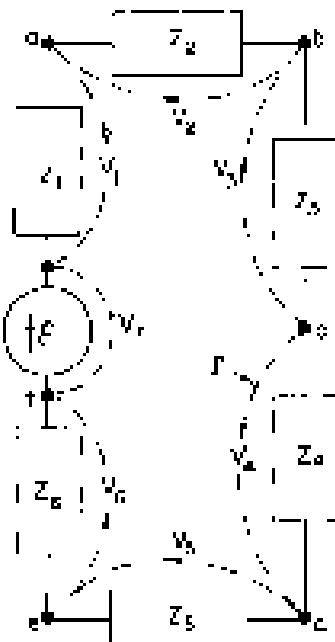


Fig. 22-9. The sum of the voltage drops around any closed path is zero.

around a complete loop of a circuit is equal to zero:

$$\sum_{\text{around}} V_i = 0. \quad (22.14)$$

This result follows from one of Maxwell's equations—that in a region where there are no magnetic fields the line integral of \mathbf{B} around any complete loop is zero.

Suppose we consider now a circuit like that shown in Fig. 22-10. The horizontal line joining the terminals a , b , c , and d is intended to show that these terminals are all connected, or that they are joined by wires of negligible resistance. In any case, this drawing means that terminals a , b , c , and d are all at the same potential V_0 and, similarly, that the terminals e , f , g , and h are also at the common potential V_1 . Then the voltage drop V across each of the four elements is the same.

Now one of our idealizations has been that negligible electrical charges accumulate on the terminals of the impedances. We now assume further that any electrical charges on the wires joining terminals can also be neglected. Then the conservation of charge requires that any charge which leaves one circuit element immediately enters some other circuit element. Or, what is the same thing, we require that the algebraic sum of the currents which enter any given junction must be zero. By a junction, of course, we mean any set of terminals a , b , c , d , e , f , g , and h which are connected. Such a set of connected terminals is usually called a "node." The conservation of charge then requires that for the network of Fig. 22-10,

$$I_1 + I_2 + I_3 + I_4 = 0. \quad (22.15)$$

The sum of the current entering the node which consists of the four terminals e , f , g , and h must also be zero:

$$I_5 + I_6 - I_7 - I_8 = 0. \quad (22.16)$$

This is, of course, the same as Eq. (22.15). The two equations are not independent. The general rule is that the sum of the currents into any node must be zero.

$$\sum_{\text{into node}} I_i = 0. \quad (22.17)$$

Our earlier conclusion that the sum of the voltage drops around a closed loop is zero must apply to any loop in a complicated circuit. Also, our result that the sum of the currents into a node is zero must be true for any node. These two equations are known as Kirchhoff's rules. With these two rules it is possible to solve for the currents and voltages in any network whatever.

Suppose we consider the more complicated circuit of Fig. 22-11. How shall we find the currents and voltages in this circuit? We can find them in the following straightforward way. We consider separately each of the four subsidiary closed loops which appear in the circuit. (For instance, one loop goes from terminal a to terminal b to terminal c to terminal d and back to terminal a .) For each of the loops we write the equation (as the first of Kirchhoff's rules) that the sum of the voltages around each loop is equal to zero. We must remember to count the voltage drop as positive if we are going in the direction of the current and negative if we are going across an element in the direction opposite to the current; and we must remember that the voltage drop across a generator is the negative of the emf in that direction. Thus if we consider the small loop that starts and ends at terminal a we have the equation

$$z_1 I_1 + z_2 I_2 + z_3 I_3 + A_1 = 0.$$

Applying the same rule to the remaining loops, we would get three more equations of the same kind.

Next, we can't write the current equation for each of the nodes in the circuit. For example, consider the currents into the node at terminal b gives the equation

$$I_1 - I_2 - I_3 = 0.$$

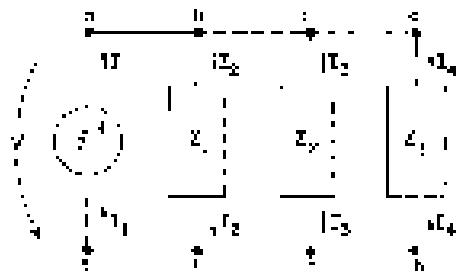


Fig. 22-10. The sum of the currents into any node is zero.

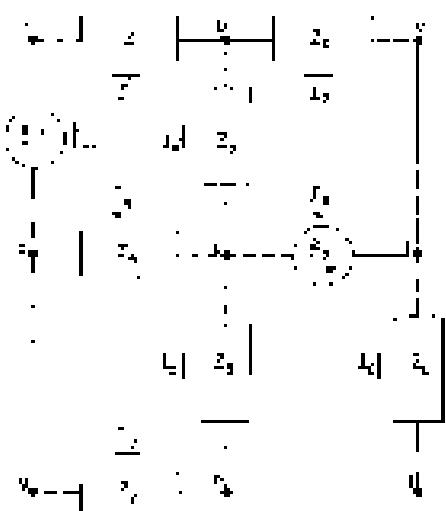


Fig. 22-11. Analyzing a circuit with Kirchhoff's rules.

Similarly, for the node between c we would have the current equation

$$I_1 + I_4 + I_5 - I_6 = 0.$$

For the circuit shown there are five such current equations. It turns out, however, that any one of these equations can be derived from the other four; there are, therefore, only four independent current equations. We thus have a total of eight independent linear equations: the four voltage equations and the four current equations. With these eight equations we can solve for the eight unknown currents. Once the currents are known the circuit is solved. The voltage drop across any element is given by the sum of those that element times its impedance (or, in the case of the voltage sources, it is already known).

We have seen that when we write the current equations, we get one equation which is not independent of the others. Generally, it is also possible to write down too many voltage equations. For example, in the circuit of Fig. 22-11, although we have considered only the four small loops, there are a large number of other loops for which we can also write the voltage equation. There is, for example, the loop along the path $a-b-f-d-a$. There is another loop which follows the path $a-b-f-e$. You can see that there are many loops. In analyzing complicated circuits it is very easy to get too many equations. There are rules which tell us how to proceed so that only the minimum number of equations is written down. But, already with a little thought it is possible to see how to get the right number of equations in the simplest form. Besides, writing an extra equation or two doesn't carry a burden. They will not lead to any wrong answers, only perhaps a little unnecessary algebra.

In Chapter 21 of Vol. I we showed that if the two impedances z_1 and z_2 are in series, they are equivalent to a single impedance z_s , given by

$$z_s = z_1 + z_2. \quad (22.8)$$

We also showed that if the two impedances are connected in parallel, they are equivalent to the single impedance z_p , given by

$$z_p = \frac{1}{(1/z_1) + (1/z_2)} = \frac{z_1 z_2}{z_1 + z_2}. \quad (22.9)$$

If you look back you will see that, in deriving these results we were in effect making use of Kirchhoff's rule. It is often possible to analyze a complicated circuit by repeated application of the formulas for series and parallel impedances. For instance, the circuit of Fig. 22-12 can be analyzed that way. First, the impedances z_1 and z_2 can be replaced by their parallel equivalent, and we also can z_3 and z_4 . Then the impedance z_5 can be combined with the parallel equivalent of z_6 and z_7 by the series rule. Proceeding in this way, the whole circuit can be reduced to a generator in series with a single impedance Z . The current through the generator is then $\mu = \mathcal{E}/Z$. Then by working backward one can solve for the currents in each of the impedances.

There are, however, quite simple circuits which cannot be analyzed by this method, as for example the circuit of Fig. 22-13. To analyze this circuit we must

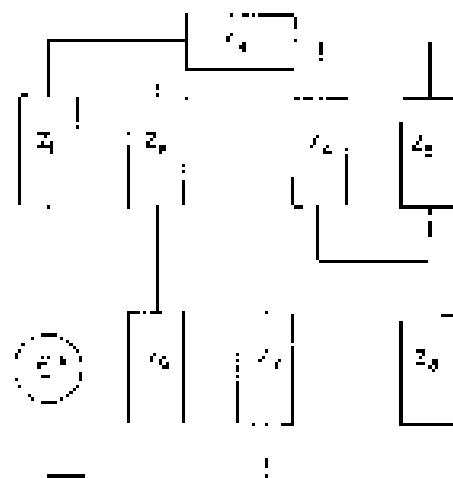


Fig. 22-12. A circuit which can be analyzed in terms of series and parallel combinations.

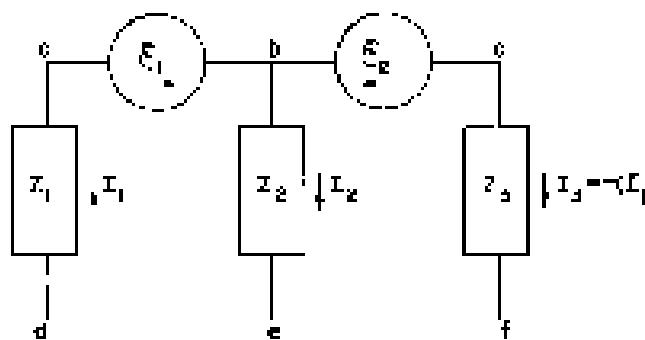


Fig. 22-13. A circuit that cannot be analyzed in terms of series and parallel combinations.

write down the current and voltage equations from Kirchhoff's rules. Let's do it. There is just one current equation:

$$I_1 + I_2 + I_3 = 0,$$

so we know immediately that

$$I_3 = -(I_1 + I_2).$$

We can save ourselves some trouble if we simply take note of this result in writing the voltage equations. But let's avoid that; there are two independent voltage equations; they are

$$-e_1 + I_2 z_2 - I_1 z_1 = 0$$

and

$$e_2 - I_1 z_1 + I_2 z_2 - I_3 z_3 = 0.$$

There are two equations and two unknown currents. Solving these equations for I_1 and I_2 , we get

$$I_1 = \frac{-z_2 e_2 - (z_2 + z_3) e_1}{z_1 z_2 - z_2 z_3 + z_1 z_3} \quad (22.20)$$

and

$$I_2 = \frac{z_1 e_2 - z_3 e_1}{z_1 z_2 - z_2 z_3 + z_1 z_3} \quad (22.21)$$

The third current is obtained from the sum of these two.

Another example of a circuit that cannot be analyzed by using the rules for series and parallel impedance is shown in Fig. 22-14. Such a circuit is called a "bridge." It appears in many instruments used for measuring impedances. With such a circuit one is usually interested in the question: How must the various impedances be related if the current through the impedance z_4 is to be zero? We leave it for you to find the conditions for which this is so.

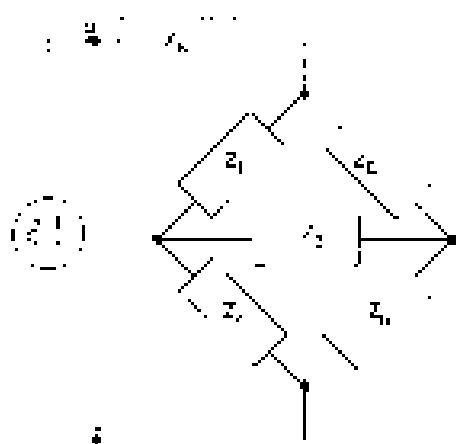


Fig. 22-14. A bridge circuit.

22-4 Equivalent elements

Suppose we connect a generator \mathcal{E} to a circuit containing some complicated interconnection of impedances, as indicated schematically in Fig. 22-15(a). All of the equations we get from Kirchhoff's rules are linear, so when we solve them for the current I through the generator, we will get that I is proportional to \mathcal{E} . We can write

$$I = \frac{\mathcal{E}}{Z_{eq}}$$

where Z_{eq} is an equivalent impedance, an algebraic function of all the elements in the circuit. If we connect another generator \mathcal{E}' after \mathcal{E} (as shown), there is no additional impedance of \mathcal{E}' . But this equation is just what we would write in the circuit of Fig. 22-15(b). So long as we are interested only in what happens to the left of the two terminals a and b , the networks of Fig. 22-15 are equivalent. We can, therefore, make the general statement that any two-terminal network of two elements can be replaced by a single impedance Z_{eq} , without changing the currents or voltages in the rest of the circuit. This statement is, of course, just a compact way of stating out of Kirchhoff's rules, and it is surely better than memory of Maxwell's equations.

The idea of a general Z is so simple that it includes generators as well as impedances. Suppose we look at such a circuit "in the point of view" of one of the terminals, which we will call a , as in Fig. 22-16(a). If we were to solve the equation for the whole circuit, we would find that the voltage V_a between the two terminals a and b is a linear function of I , which we can write

$$V_a = A - BI_{eq} \quad (22.22)$$

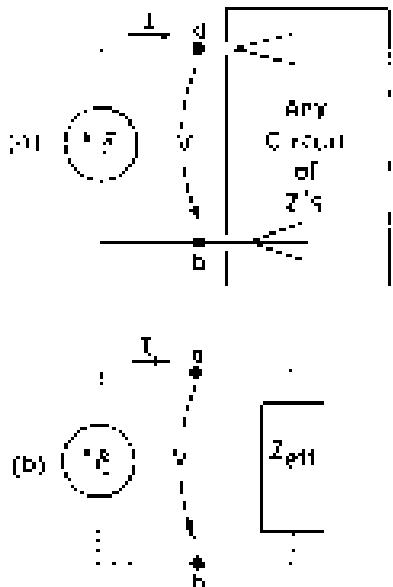


Fig. 22-15. Any two-terminal network of passive elements is equivalent to an effective impedance.

where A and B depend on the generators and impedances in the circuit to the left of a .

of the terminals. For instance, for the circuit of Fig. 22-13, we find $V_1 = I_1 z_1$. This can be written (by rearranging Eq. (22-20)) as

$$V_1 = \left[\left(\frac{z_1}{z_1 + z_2} \right) V_{\text{in}} - \frac{1}{2} \frac{z_1 z_2}{z_1 + z_2} I_1 \right]. \quad (22-23)$$

The complete solution is then obtained by combining this equation with the one for the impedance z_2 , namely, $V_2 = I_2 z_2$. As in the general case, by combining Eq. (22-22) with

$$V_2 = I_2 z_2,$$

If now we consider that z_1 is attached to a simple series circuit of a generator and a current, as in Fig. 22-15(b), the equation corresponding to Eq. (22-22) is

$$V_{\text{in}} = E_{\text{gen}} + I_{\text{gen}} z_{\text{gen}}$$

which is identical to Eq. (22-22) provided we set $E_{\text{gen}} = A$ and $I_{\text{gen}} = B$. So if we are interested only in what happens to the right of the terminals a and b , the arbitrary circuit of Fig. 22-16 can always be replaced by an equivalent circuit consisting of a generator in series with an inductance.

22-4 Energy

We have seen that to build up the current I in an inductance, the energy $G = \frac{1}{2} I^2 L$ must be provided by the external circuit. When the current falls back to zero, this energy is delivered back to the external circuit. There is no energy loss mechanism in an ideal inductance. When there is an alternating current through an inductance, energy flows back and forth between it and the rest of the circuit, but the average rate at which energy is delivered to the circuit is zero. We say that an inductance is a nondissipative element; no electrical energy is dissipated—that is, “lost”—in it.

Similarly, the energy of a condenser, $G = \frac{1}{2} C V^2$, is returned to the external circuit when a condenser is discharged. When a condenser is in an ac circuit energy flows in and out of it, but the net energy flow in each cycle is zero. An ideal condenser is also a nondissipative element.

We know that an emf is a source of energy. When a current I flows in the direction of the emf, energy is delivered to the external circuit at the rate $dG/dt = SI$. If current is driven against the emf by other generators in the circuit, the emf will absorb energy at the rate $-SI$; since I is negative, dG/dt will also be negative.

If a generator is connected to a resistor R , the current through the resistor is $I = E/R$. The energy being supplied by the generator at the rate SI is being absorbed by the resistor. This energy goes into heat in the resistor and is lost from the electrical energy of the circuit. We say that electrical energy is dissipated in a resistor. The rate at which energy is dissipated in a resistor is $dG/dt = RI^2$.

In an ac circuit the average rate of energy loss to a resistor is the average of dG/dt over one cycle. Since $I = I_0 \sin \omega t$, by which we really mean that I varies as $\sin \omega t$, the average of I^2 over one cycle is $|I|^2/2$, since the peak current is $|I|$ and the average of $\cos^2 \omega t$ is $1/2$.

What about the energy loss when a generator is connected to an arbitrary impedance z ? (By “loss” we mean, of course, conversion of electrical energy into thermal energy.) Any impedance z can be written as the sum of its real and imaginary parts. That is,

$$z = R + jX, \quad (22-24)$$

where R and X are real numbers. From the point of view of equivalent circuits we can say that any impedance z is equivalent to a resistance in series with a pure inductance or pure capacitance—called a generator—as shown in Fig. 22-17.

We have seen earlier that any circuit that contains only J 's and C 's has no dissipative loss if J is a pure inductive member. Since there is no energy loss in any of the J 's and C 's out in a source, a pure resistance containing only R 's and C 's will have no energy loss. We can see that this must be true in general for a resistance.

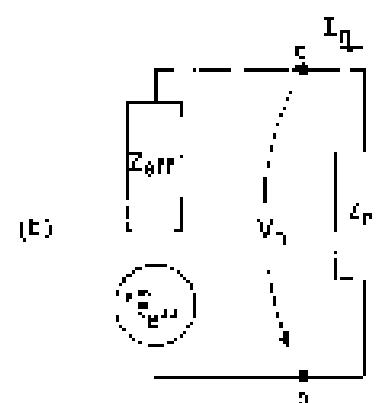
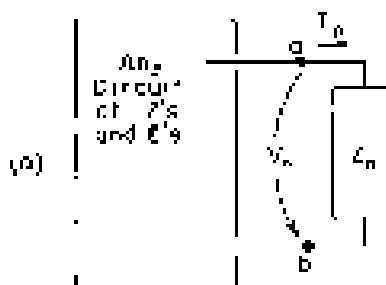


Fig. 22-16. Any a - b -terminal network can be replaced by a generator-series with an inductor.

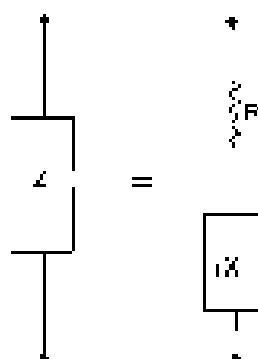


Fig. 22-17. Any impedance z is equivalent to a series combination of a pure resistance and a pure reactance.

If a generator with the emf E is connected to the impedance z of Fig. 22-17, the emf must be related to the current I from the generator by

$$E = IR = (I)z. \quad (22.25)$$

To find the average rate at which energy is delivered, we take the average of the product IR . Now we must be careful. When dealing with such products, we must deal with the real quantities $E(t)$ and $I(t)$. (The real parts of the complex quantities will represent the actual physical quantities only when we have linear equations; here we are concerned with products, which are certainly not linear.)

Suppose we choose our origin of t so that the amplitude I is a real number, let's say I_0 , then the total time variation I is given by

$$I = I_0 \cos \omega t.$$

The emf of Eq. (22.25) is the real part of

$$E_0 \cos(\omega t + \phi) (R + jX)$$

or

$$E = I_0 R \cos \omega t - I_0 X \sin \omega t. \quad (22.26)$$

The two terms in Eq. (22.26) represent the voltage drops across R and X in Fig. 22-17. We see first, the voltage drop across the resistance is in phase with the emf E , while the voltage drop across the purely reactive part is out of phase with the emf.

The average rate of energy loss, $\langle P \rangle_{av}$, from the generator is the integral of the product of over one cycle divided by the period T , in other words,

$$\langle P \rangle_{av} = \frac{1}{T} \int_0^T E I dt = \frac{1}{T} \int_0^T I_0 R \cos^2 \omega t dt = \frac{1}{T} \int_0^T P X \cos \omega t \sin \omega t dt.$$

The first integral is $\langle I_0^2 R \rangle$, and the second integral is zero. So the average energy loss in an impedance $z = R + jX$ depends only on the real part of z , and is $E_0^2 R / 2$, which is in agreement with our earlier result for the energy loss in a resistor. There is no energy loss in the reactive part.

22-6 A ladder network

We would like now to consider an interesting circuit which can be analyzed in terms of series and parallel combinations. Suppose we start with the circuit of Fig. 22-18(a). We can see right away that the impedance from terminal a to terminal b is simply $z_a - z_b$. Now let's take a little ladder circuit, the one shown in Fig. 22-18(b). We could analyze this circuit using Kirchhoff's rules, but it is also easy to handle with series and parallel combinations. We can replace the two impedances on the right hand end by a single impedance $z_3 = z_1 + z_2$, as in part (c) of the figure. Then the two impedances z_1 and z_2 can be reduced by their equivalent parallel impedance z_p , as shown in part (d) of the figure. Finally, z_1 and z_2 are equivalent to a single impedance z_{av} , as shown in part (e).

Now we may ask an amusing question: What would happen if in the network of Fig. 22-18(b) we kept on adding more sections forever—so we indicate by the dashed lines in Fig. 22-19(a)? Can we solve such an infinite network? Well, that's

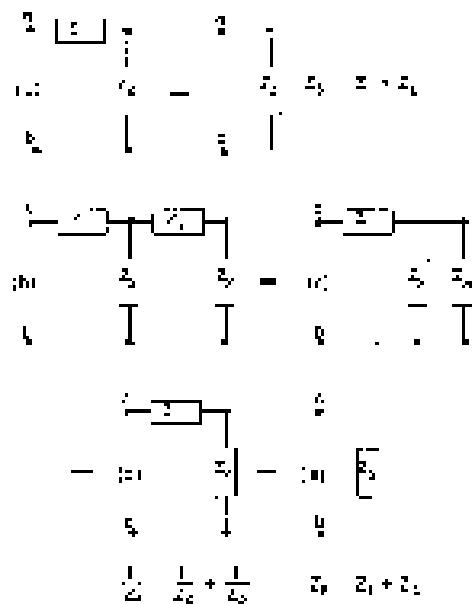


Fig. 22-18. The effective impedance of a ladder.

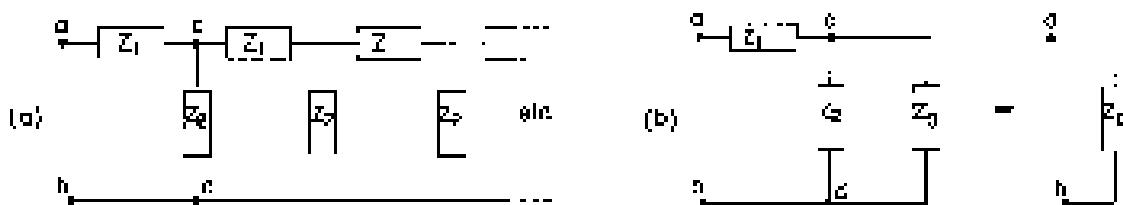


Fig. 22-19. The effective impedance of an infinite ladder.

not so hard. First, we notice that such an infinite network is unchanged if we add one more section at the "front" end. Secondly, if we add one more section to an infinite network it is still the same infinite network. Suppose we call the impedance between the two terminals a and b of the infinite network z_0 ; then the impedance of all the stuff to the right of the two terminals a and b is also z_0 . Therefore, so far as Eq. (22-19) and its consequences are concerned, we can represent the network as shown in Fig. 22-19(b). Combining the parallel combinations z_0-z_1 and adding the result in series with z_1 , we can immediately write down the impedance of this combination:

$$z_1 + z_0 = \frac{1}{(1/z_1) + (1/z_0)} \quad (22-20)$$

But this impedance is also equal to z_0 , so we have the equation

$$z_0 = z_1 + \frac{z_1 z_0}{z_0 - z_1}$$

We can solve for z_0 to get

$$z_0 = \frac{1}{\frac{1}{z_1} + \sqrt{(\frac{1}{z_1})^2 + z_1/z_0}} \quad (22-21)$$

So we have found the solution for the impedance of an infinite ladder of resistor units and parallel impedances. The impedance z_0 is called the *characteristic impedance* of such an infinite network.

Let's now consider a specific example in which the series element is an inductance L and the shunt element is a capacitance C , as shown in Fig. 22-20(a). In this case we find the impedance of the infinite network by setting $z_1 = j\omega L$ and $z_0 = j\omega C$. Notice that the first term, z_1/z_0 , in Eq. (22-21) is just equal to the impedance of the first element. It would therefore seem more natural, in at least somewhat simpler, if we were to draw our infinite network as shown in Fig. 22-20(b). Looking at the infinite network from the terminal a we would fix the characteristic impedance

$$z_0 = \sqrt{j\omega LC} = (\omega^2 LC)^{1/4}. \quad (22-22)$$

Now there are two interesting cases, depending on the frequency ω . If $\omega^2 L^2$ is less than $1/4C$, the second term in the radicand will be smaller than the first, and the impedance z_0 will be a real number. On the other hand, if ω^2 is greater than $4/LC$ the impedance z_0 will be a pure imaginary number which we can write as

$$z_0 = \pm \sqrt{j\omega LC/4} = j(\omega L^2)^{1/4}.$$

We have said earlier that $j\omega L^2$ which represents only inductive inductances, such as inductances and capacitors, will have an impedance which is purely imaginary. How can it be then that for the circuit we are now studying—which has only L 's and C 's—the impedance is a pure resistance for frequencies below $\sqrt{4/LC}$? For higher frequencies the impedance is surely being noisy, in agreement with our earlier statement. But lower frequencies the impedance is a pure resistance and will therefore absorb energy. But how can the circuit continue to absorb energy, if it is made only of inductances and capacitors? Answer: Because this is an infinite number of inductances and capacitors, so that when a source is connected to the circuit it supplies energy to the last inductance and capacitor, then to the second, to the third, and so on. In a circuit of this kind, energy is continually absorbed from the generator at a constant rate and flows continually out of the network, supplying energy which is stored in the inductive inductances close to the line.

This is a very strange situation, and about what is happening in the circuit. We would expect that, for example, a source to the front end, the effects of this source will be propagated through the network toward the infinite end. The propagation of the waves toward the line is much like the radiation from an antenna which loses energy from its driving source; that is, we expect such a propagation to occur, when the impedance is real, which occurs if ω is less than $\sqrt{4/LC}$. But since the impedance is purely imaginary, which happens for ω greater than $\sqrt{4/LC}$, we would not expect to see any such propagation.

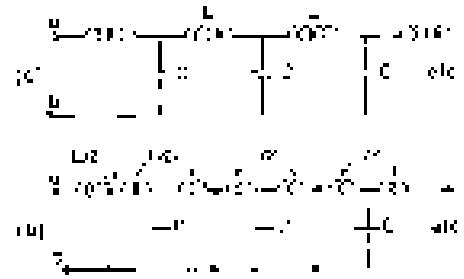


Fig. 22-20. An L-C ladder, shown in two equivalent ways.

22-7 Ladders

We saw in (Lecture 20) that the infinite ladder network of Fig. 22-20 above can carry energy continuously if it is driven at a frequency below a certain critical frequency $\sqrt{A/LC}$, which we will call the "cutoff frequency" ω_0 . We suggested that this effect could be understood in terms of a continuous transport of energy down the line. On the other hand, at high frequencies, for $\omega > \omega_0$, there is no continuous absorption of energy; we should then expect that perhaps the currents don't "propagate" very far down the line. Let's see whether these ideas are right.

Suppose we have the input end of the ladder connected to some AC generator and we ask what the voltage looks like at, say, the 75th section of the ladder. Since the network is infinite, whatever happens to the voltage from one section to the next is always the same; so let's just look at what happens when we go from some section, say the n th to the next. We will define the currents I_n and voltages V_n as shown in Fig. 22-21(a).

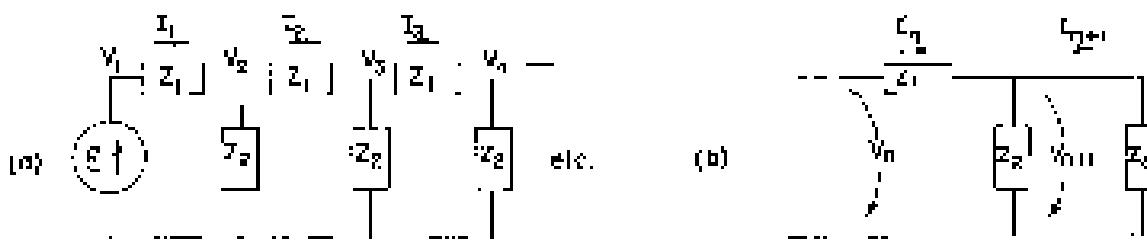


Fig. 22-21. Finding the propagation factor of a ladder.

We can get the voltage V_{n+1} from V_n by remembering that we can always replace the rest of the ladder after the n th section by its characteristic impedance γ_n ; then we need only analyze the circuit of Fig. 22-21(b). First, we notice that any V_{n+1} since I_n is across Z_n , must again I_{n+1} . Also, the difference between V_n and V_{n+1} is just $I_n Z_n$:

$$V_n - V_{n+1} = I_n Z_n \approx V_n \frac{\gamma_n}{\gamma_n}$$

So we get the ratio

$$\frac{V_{n+1}}{V_n} = -\frac{Z_n}{Z_{n+1}} \approx \frac{\gamma_n}{\gamma_{n+1}}$$

We can call this ratio the *propagation factor* for one section of the ladder; we'll call it α . It is, of course, the same for all sections:

$$\alpha = \frac{\gamma_n}{\gamma_{n+1}} \quad (22.29)$$

The voltage after the n th section is then

$$V_n = \alpha^n V_0 \quad (22.30)$$

You can now link the voltage after 254 sections, V_{254} , just to the initial power times α .

Suppose we set up a circuit for the $L-C$ ladder of Fig. 22-20(p) (using π_0 from Eq. (22.23)), and $\omega = \omega_0$; we get

$$\alpha = \frac{\sqrt{A/C} + (\omega C/4)}{\sqrt{A/C} - (\omega C/4)} = (\omega d/2) \quad (22.31)$$

If the driving frequency is below the cutoff frequency $\omega_0 = \sqrt{A/LC}$, the radical is a real number, and the magnitudes of the complex numbers in the numerator and denominator are equal. Therefore, the magnitude $|\alpha|$ is one; we can write

$$\alpha = e^{j\phi}$$

which means that the magnitude of the voltage is the same at every section, only

dephase phase. The phase change ϕ , is, in fact, a negative number and represents the "delay" of the voltage as it passes through the network.

For frequencies above the cutoff frequency ω_0 , it is useful to factor out ω/ω_0 from the numerator and denominator of Eq. (22-31) and rewrite it as

$$\alpha = \frac{\sqrt{L/C}(\omega/4)}{\sqrt{(4\omega^2/4) - (L/C)} - (L/C)} = \frac{1}{\omega\omega_0}. \quad (22-32)$$

The propagation factor α is now a real number, and a number less than one. That means that the voltage at any section is always less than the voltage at the preceding section by the factor α . For any frequency above ω_0 , the voltage decreases rapidly as we go along the network. A plot of the absolute value of α as a function of frequency looks like the graph in Fig. 22-22.

We see that the behavior of α , both above and below ω_0 , agrees with our interpretation that the network propagates, stores energy, and blocks it for $\omega > \omega_0$. We say that the network "passes" low frequencies and "rejects" or "blocks" the high frequencies. Any network designed to have these characteristics says in a prescribed way with frequency is called a "filter." We have been analyzing a "low-pass filter."

You may be wondering why all this discussion of an infinite network which obviously cannot actually occur. The point is that the same characteristics are found in a finite network if we finish it off at the end with an impedance equal to the characteristic impedance ω_0 . Now in practice it is not possible to exactly reproduce the characteristic impedance with a few simple elements like resistors, inductors, and capacitors. But it is often possible to do so with a fair approximation for a certain range of frequencies. In this way one can make a finite filter network whose properties are very nearly the same as those for the infinite case. For instance, the L-C ladder behavior which we have described in (c) is contained in the pure resistance $R = \sqrt{L/C}$.

If in our L-C ladder we interchange the positions of the L's and C's, to make the ladder shown in Fig. 22-23(a), we can have a filter that propagates high frequencies and rejects low frequencies. It's easy to see what happens with this network by using the results we already have. You will notice first whenever we change an L to a C and vice versa, we also change every ω to $1/\omega_0$. So whatever happens at ω before will now happen at $1/\omega$. In particular, we can see how α will vary with frequency by using Fig. 22-22 and changing the label on the axis to $1/\omega_0$, as we have done in Fig. 22-23(b).

The low-pass and high-pass filters we have described have various technical applications. An L-C low-pass filter is often used as a "smoothing" filter in a power supply. If we want to manufacture \propto power from an ac source, we begin with a rectifier which permits current to flow only in one direction. From the rectifier we get a series of pulses that looks like the function $V(t)$ shown in Fig. 22-24, which is noisy ω , because it wobbles up and down. Suppose we would like a nice pure \propto , such as a battery provides. We can come close to that by putting a low-pass filter between the rectifier and the load.

We know from Chapter 8 of Vol. I that the time function in Fig. 22-24 can be represented as a superposition of a constant voltage plus a sine wave, plus a higher frequency sine wave, plus a still higher frequency sine wave, etc.—by a Fourier series. If our filter is linear (i.e., as we have been assuming, the L's and C's don't vary with the currents or voltages), then what comes out of the filter is the superposition of the outputs for each component of the input. If we arrange that the cutoff frequency ω_0 of our filter is well below the lower frequency in the function $V(t)$, the $\omega = 0$ (for which $\omega = 0$) goes through fine, but the amplitude of the first harmonic will be cut down a lot. And amplitudes of the higher harmonics will be cut down even more. So we can get the output as smooth as we wish, depending only on how many filter sections we are willing to buy.

A high-pass filter is used in one way to reject certain low frequencies. For instance, in a phonograph amplifier a high-pass filter may be used to let the music

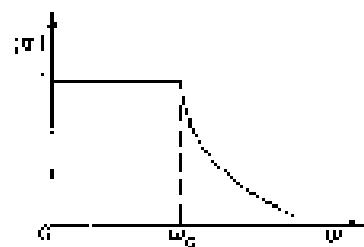
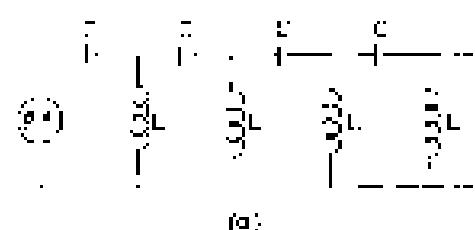
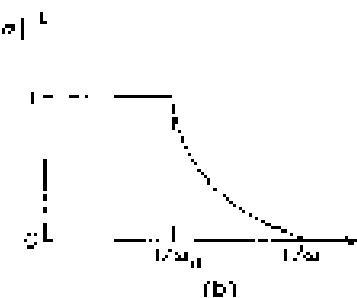


Fig. 22-22. The propagation factor as a function of frequency in an L-C ladder.



(a)



(b)

Fig. 22-23. (a) A high-pass filter; (b) its propagation factor as a function of $1/\omega$.

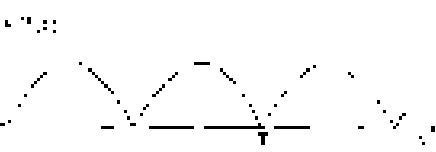


Fig. 22-24. The output voltage $v(t)$ of a full-wave rectifier.

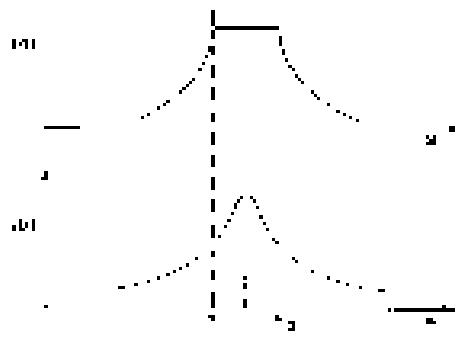


Fig. 22-25. (a) A band-pass filter.
(b) A simple resonant filter.

through, while keeping out the low-passed sounding from the music of the turntable.

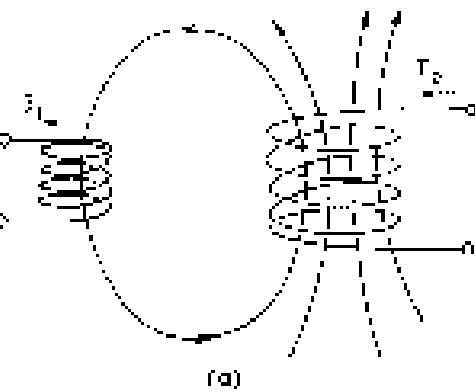
It is also possible to make "band-pass" filters that, to cut frequencies below some frequency ω_1 and above another frequency ω_2 (greater than ω_1), but pass the frequencies between ω_1 and ω_2 . This can be done simply by putting together a high-pass and a low-pass filter, but it is more usually done by making a "filter in which the impedances ω_1 and ω_2 is more complicated—here, and a combination of ω_1 's and ω_2 's. Such a band-pass filter might have a propagation constant like that shown in Fig. 22-25(a). It might be used, for example, in separating signals that occupy only an interval of frequencies, such as most of the music wave channels in a high-frequency telephone system, or the modulated carrier of a radio transmission.

We have seen in Chapter 25 of Vol. I that such filtering can also be done using the selectivity of an ordinary resonator curve, which we have shown in comparison in Fig. 22-25(b). But the resonator filter is not as good for some purposes as the band-pass filter. You will remember (Chapter 25, Vol. I) that when a carrier of frequency ω_1 is modulated with a "signal" frequency ω_2 , the total signal contains not only the carrier frequency but also the two side-band frequencies $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$. With a resonator filter, these side-bands are always attenuated somewhat, and the attenuation is more, the higher the signal frequency, as you see from the figure. So there is a poor "frequency response." The higher musical tones don't get through. But if the filtering is done with a band-pass filter, designed so that the width $\omega_2 - \omega_1$ is at least twice the highest signal frequency, the frequency response will be "flat" for the signals wanted.

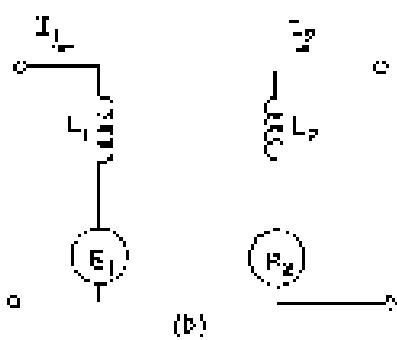
We want to make one more point about the ladder filter: the *L-C* ladder of Fig. 22-20 is often an approximate representation of a transmission line. If we have a long conductor that runs parallel to another conductor—such as a wire in a coaxial cable, or a wire suspended above the earth—there will be some capacitance between the two conductors and also some inductances due to the magnetic field between them. If we imagine the line as broken up into small lengths Δz , each length will look like one section of the *L-C* ladder with a series inductance ΔL and a shunt capacitance ΔC . We can then use our results for the ladder filter. If we take the limit as Δz goes to zero, we have a good description of the transmission line. Notice that as Δz is made smaller and smaller, both ΔL and ΔC decrease but in the same proportion, so that the ratio $\Delta L/\Delta C$ remains constant. So if we take the circuit of Eq. (22-28) as Δz and ΔC go to zero, we find that the characteristic impedance ω_0 is a pure resistance whose magnitude is $\sqrt{\Delta L/\Delta C}$. We can also write the ratio $\Delta L/\Delta C$ as L_0/C_0 , where L_0 and C_0 are the inductance and capacitance of a unit length of the line; then we have

$$\omega_0 = \sqrt{L_0/C_0} \quad (22-39)$$

You will also notice that as ΔL and ΔC go to zero, the cutoff frequency $\omega_c = \sqrt{4/LC}$ goes to infinity. There is no cutoff frequency for an ideal transmission line.



(a)



(b)

Fig. 22-26. Equivalent circuit of a valve inductor.

22-8 Other circuit elements

We have so far defined only the ideal circuit impedances—the inductance, the capacitance, and the resistance—as well as the ideal voltage generator. We would now be able to show that other elements, such as mutual inductances or varactors diodes, can be described by using only the same three elements. Suppose that we have two coils and that on purpose, or otherwise, some flux from one of the coils links the other, as shown in Fig. 22-28(a). Then the two coils will have a mutual inductance M such that when the current varies in one of the coils, there will be a voltage generated in the other. Can we take account of such a effect in our equivalent circuits? We can in the following way. We have seen that the

induced emf's in each of two inexciting coils can be written as the sum of two parts:

$$\begin{aligned}\delta_1 &= -L_1 \frac{dI}{dt} + M \frac{dI_2}{dt}, \\ \delta_2 &= -L_2 \frac{dI_2}{dt} + M \frac{dI_1}{dt}.\end{aligned}\quad (22.34)$$

The first term comes from the self-inductance of the coil, and the second term comes from its mutual inductance with the other coil. The sign of the second term can be plus or minus, depending on the way the flux from one coil links the other. Making the same approximations we used in describing an ideal inductance, we would say that the potential difference across the terminals of each coil is equal to the electromotive force in the coil. Then the two equations of (22.34) are the same as the ones we would get from the circuit of Fig. 22-26(b), provided the electromotive force in each of the two circuits shown depends on the current in the opposite circuit according to the relations

$$E_1 = -2\pi M I_2, \quad E_2 = -2\pi M I_1. \quad (22.35)$$

So what we can do is represent the effect of the self-inductance in a normal way and replace the effect of the mutual inductance by an auxiliary ideal voltage generator. We must, in addition, of course, have the equation that relates this emf to the current in some other part of the circuit; but so long as this equation is linear, we have just added more linear equations to our circuit equations, and all of our earlier conclusions about equivalent circuits and so forth are still correct.

In addition to mutual inductances there may also be mutual capacitances. So far, when we have talked about condensers we have always imagined that there were only two electrodes, but, in many situations, for example in a vacuum tube, there may be many electrodes close to each other. If we put an electric charge on any one of the electrodes, its electric field will induce charges on each of the other electrodes and affect its potential. As an example, consider the arrangement of four plates shown in Fig. 22-27(a). Suppose these four plates are connected to external circuits by means of the wires A, B, C, and D. So long as we are only worried about electrostatic effects, the equivalent circuit of such an arrangement of electrodes is as shown in part (b) of the figure. The electrostatic interaction of any electrode with each of the others is equivalent to a capacity between the two electrodes.

Finally, let's consider how we should represent such complicated devices as transistors and radio tubes in an ac circuit. We should point out at this stage that such devices are often operated in such a way that the relationship between the currents and voltages is not at all linear. In such cases, those statements we have made which depend on the linearity of equations are, of course, no longer correct. On the other hand, in many applications the operating characteristics are sufficiently linear that we may consider the transistors and tubes to be linear devices. By this we mean that the alternating currents in, say, the plate of a vacuum tube are linearly proportional to the voltages that appear on the other electrodes, say the grid voltage and the plate voltage. When we have such linear relationships, we can incorporate the device into our equivalent circuit representation.

As in the case of the mutual inductance, our representation will have to include auxiliary voltage generators which describe the influence of the voltages or currents in one part of the device on the currents or voltages in another part. For example, the plate circuit of a tube can usually be represented by a resistance in series with an ideal voltage generator whose strength is proportional to the grid voltage. We get the equivalent circuit shown in Fig. 22-28(c). Similarly, the collector circuit

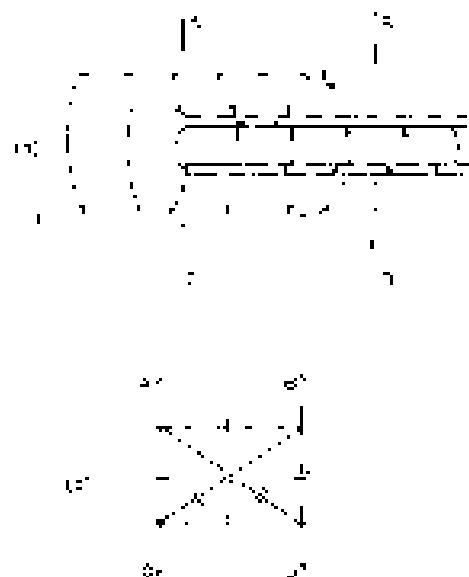


Fig. 22-27. Equivalent circuit of mutual capacitance.

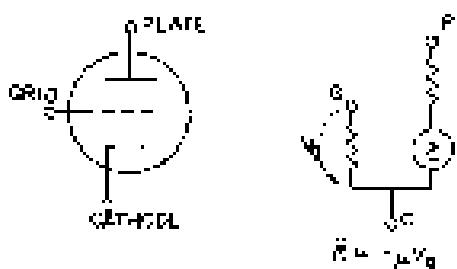


Fig. 22-28. A low-frequency equivalent circuit of a vacuum tube.

* The equivalent circuit, shown here is only for low frequencies. For high frequencies the equivalent circuit gets much more complicated and will include various so-called "parasitic" capacitances and inductances.

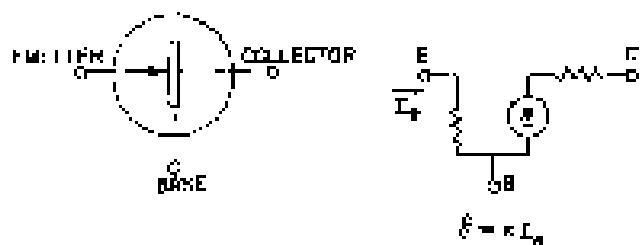


Fig. 22-29. A low-frequency equivalent circuit of a transistor.

of a capacitor is conveniently represented as a resistor in series with an ideal voltage generator whose output voltage is proportional to the current from the emitter to the base of the transistor. The equivalent circuit is then like that in Fig. 22-29. So long as the resistances which describe the elements are linear, we can use such representations for tubes or transistors. (Note, when tubes are incorporated in a complicated network, our general conclusion about the equivalent representation of any arbitrary connection of elements is still valid.)

There is one remarkable thing about transistors and radio tube circuits which is different from circuits containing only impedances: the real part of the effective impedance γ_{AB} can be truly negative. We have seen that the real part of γ represents the loss of energy. But, it is the characteristic of transistors and tubes that they *supply* energy to the circuit. (Of course they don't just "pass" energy; they also convert from the DC currents of the power supplies into currents at high AC voltage.) So it is possible to have a circuit with a negative real-term. Such a circuit has the property that if you connect it to an input with a positive real part, i.e., a positive resistance, and accept emitters as the sign of the two real parts is exactly *opp*, then there is no dissipation in the connected circuit. If there is no loss of energy, any alternating voltage once started will continue forever! This is the basic idea behind the operation of an oscillator or signal generator which can be used as a source of alternating voltage at any desired frequency.

Circuit Resonance

23-1 Real circuit elements

When looked at from any one point in a circuit, any arbitrary circuit will look up to an impedance and admittance as, at any given frequency, equivalent to a generator & load in series with an impedance. This makes sense because if we put a voltage V across the terminals and solve all the equations to find the current I , we must get a linear relation between the current and the voltage. Since all the equations are linear, the result for I must also depend on V linearly or $I = KV$. The most general linear form can be expressed as

$$I = \frac{1}{Z} (V - R) \quad (23.1)$$

In general, both Z and R may depend in some complicated way on the frequency ω . Equation (23.1), however, is the relation we would get if below the two terminals there was just a generator "in series with its impedance Z ".

There is also the opposite kind of question. If we have any electronic circuit device at which we know the impedances and we measure the voltage between two points to determine certain functions of frequency, can we find a equivalent circuit of real elements that is equivalent to the stated impedances? The answer is that for any reasonably simple & physically reasonable circuit it is impossible to approximate the situation to as high a degree of accuracy as you wish with a finite number of real elements. We can't want to consider the general problem now, but only look at what might be expected from physical arguments for a few cases.

If we think of a real resistor, we know that the current through it will produce a magnetic field. At any rate, resistors almost always have some inductance. After all, when a resistor has a potential difference across it, there must be charges on the ends of the resistor to produce the necessary electric field. As the voltage changes, the charges will change in proportion, so the resistor will also have some capacitance. We expect that a real resistor *in the form* the equivalent circuit shown in Fig. 23-1. In a well designed resistor, the so called "parasitic" elements A and C are small, so that at the frequencies for which it's intended, ω_A is much less than ω , and ω_C is much greater than ω . It may therefore be possible to neglect them. As the frequency is raised, however, they will eventually become important, and a resistor begins to look like a resonant circuit.

A real inductor is also not equal to the idealized inductor, whose impedance is ωL . A real coil of wire will have some resistance, so at low frequencies the coil is really equivalent to an inductor in series with some resistance, as shown in Fig. 23-2(a). But, you are thinking, the resistance and inductance are *together* in a real coil—the resistance is spread all along the wire so it's mixed in with the inductance. We should probably use a circuit more like the one in Fig. 23-2(b), which has several little R 's and L 's in series. But the total impedance of such a circuit is just $\sum R + \sum L\omega$. That's what is equivalent to the simpler diagram of part (a).

As we go up in frequency ω , in a real coil the representation of an inductor as a pure inductor is no longer very good. The charges that must build up on the wires to make the voltage will become non-linear. It's as if there were little resistors across the turns of the n_1 and n_2 sketched in Fig. 23-2(a). We might try to approximate the real coil by the circuit of Fig. 23-2(b), but in fact, this circuit can be converted fairly easily to the simpler form of part (a) of the figure (which is again the only reasonable circuit we could find for the high-frequency theory of a resistor). For higher frequencies, however, the more complexer circuit of

23-1 Real circuit elements

23-2 A capacitor at high frequencies

23-3 A resonant cavity

23-4 Cavity modes

23-5 Cavities and resonant circuits

Review: Chapter 21, Vol. I, Resonance
Chapter 19, Vol. I, Magnetism

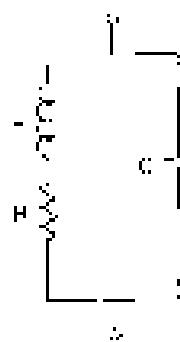


Fig. 23-1. Equivalent circuit of a real resistor.

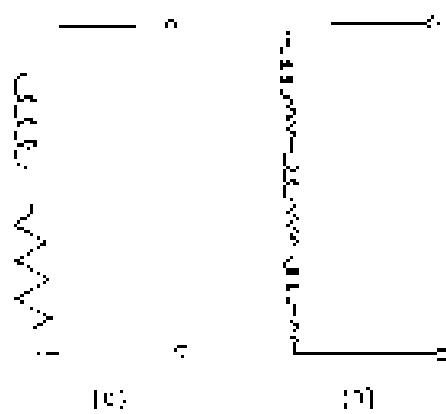


Fig. 23-2. The equivalent circuit of a real inductor at low frequencies.

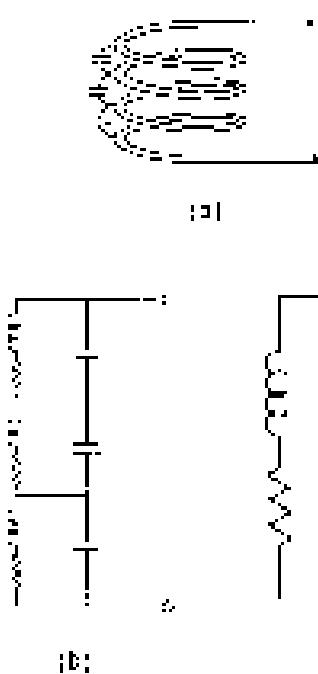


Fig. 23-3. The negative ωL circuit of a real inductance at higher frequencies.

Fig. 24. (a) As you learn, to build the more accurately you wish to approach the actual impedance of a real physical inductor, the more ideal elements you will have to use in the artificial model at it.

Let's look a little more closely at what goes on in a real coil. The impedance of an inductor depends on what it is wound around. For example, if it is wound around a thin wire, the reactance of the wire is small. As we go up in frequency, ω , one becomes much larger than R , and the coil looks mostly like a real inductor. At very low frequencies, the inductor becomes unimportant. At very high frequencies, a resistor is in "series" across the coil, and when it is in parallel with something else, it allows no current. But at high frequencies, the current prefers to flow into the capacitance between the turns, rather than through the inductor. So the current in the coil jumps from one turn to the other and doesn't have to go around and around where it has to stick the ear! So although we may have imagined that the current should go around the loop, it will take the easier path—the path of least impedance.

If the subject had been one of popular interest, this effect would have been called "the high frequency barrier," or some such name. The same kind of thing happens in all subjects. In aerodynamics, if you try to make things go faster than the speed of sound when they were designed for lower speeds, they don't work. It doesn't mean that there is a great "barrier" there; it just means that the object should be redesigned. So this coil which we designed as an "inductance" is not going to work as a good inductance, but as some other kind of thing at very high frequencies. For high frequencies, we have to find a new design.

23-2. A capacitor at high frequencies

Now we want to discuss in detail the behavior of a capacitor—a geometrically ideal capacitor—as the frequency gets larger and larger, so we can see the transition off its properties. (We prefer to use a capacitor instead of an inductance, because the geometry of a pair of plates is much less complicated than the geometry of a coil.) We consider the capacitor shown in Fig. 23-4(a), which consists of two parallel circular plates connected to an external generator by a pair of wires. If we charge the capacitor well DC, there will be a positive charge on one plate and a negative charge on the other, and there will be a uniform electric field between the plates.

Now suppose that instead of DC, we put in $V_0 \cos \omega t$ of low frequency on the plates. (We will find out later what is "low" and what is "high.") Say we repeat the capacitor as a lower-frequency generator. As the voltage alternates, the positive charge on the top plate is taken off and moves to the bottom plate. While that is happening, the electric field disappears and then, in due time, it appears in reverse.

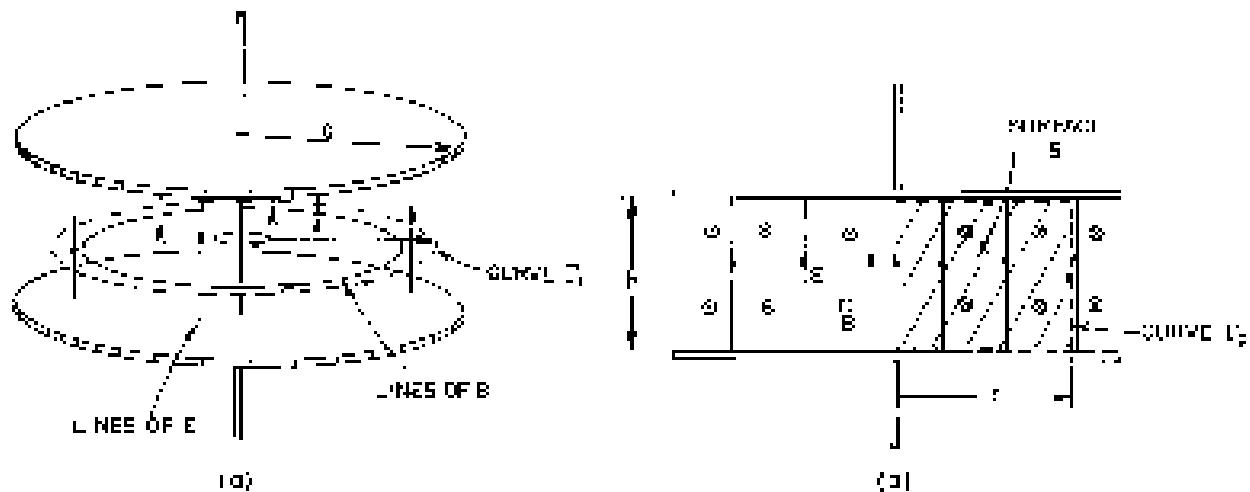


Fig. 23-4. The electric and magnetic fields between the plates of a capacitor.

As the charge shuffles back and forth slowly, the electric field follows. At each instant the electric field is uniform as shown in Fig. 22-1(b), except for some edge effects which we are going to disregard. We can write the magnitude of the electric field as

$$E = E_0 e^{i\omega t}, \quad (22.2)$$

where E_0 is a constant.

Now will that continue to be right as the frequency goes up? No, because as the electric field is going up and down, there is a flux of electric field through any loop like "1" in Fig. 22-4(a). And as you know, a changing electric field sets up a magnetic field. One of Maxwell's equations says that when there is a varying electric field, as there is here, the \mathbf{E} has got to be a line integral of the magnetic field. (Or integrals of the magnetic field are on a closed loop, and induced by \mathbf{E} , so if you go around the loop twice, the change in \mathbf{E} there are no net work.)

$$\epsilon^2 \oint_{C_1} \mathbf{B} \cdot d\mathbf{s} = \frac{\partial}{\partial t} \int_{\text{inside } C_1} \mathbf{E} \cdot d\mathbf{s}. \quad (22.3)$$

So now what magnetic field is there? That's not very hard. Suppose that we take the loop C_1 , which is a circle of radius r . We can see from symmetry that the magnetic field goes around as shown in the figure. Then the line integral of \mathbf{B} is $2\pi r B$. And, since the electric field is uniform, the flux of the electric field is simply E multiplied by πr^2 , the area of the circle:

$$r' B \cdot 2\pi r = \frac{\partial}{\partial t} (E \cdot \pi r^2). \quad (22.4)$$

The derivative of E with respect to time is, for our alternating field, simply $i\omega E_0 e^{i\omega t}$. So we find that our capacitor has the magnetic field

$$B = \frac{\mu_0}{2\pi r} E_0 e^{i\omega t}. \quad (22.5)$$

In other words, the magnetic field also oscillates and with a strength proportional to r .

What is Earth's law of that? When there is a magnetic field that is varying, there will be an induced electric field, or the capacitor will begin to act a little bit like an inductor. As the frequency goes up, the magnetic field gets stronger; it is proportional to the rate of change of E , and so to $i\omega$. The resistance of the capacitor will no longer be simply R/r .

Let's continue to raise the frequency and to analyze what happens more carefully. We have a magnetic field that goes shushing back and forth. But then the electric field cannot be uniform, as we have assumed! When there is a varying magnetic field, there must be a line integral of the electric field—because of Faraday's law. So if there is an appreciable magnetic field, as begins to happen at higher frequencies, the electric field cannot be the same at all distances from the center. The electric field must change with r so that the line integral of the electric field can equal the changing flux of the magnetic field.

Let's see if we can figure out the correct electric field. We can do that by computing a "correction" to the uniform field we originally assumed for low frequencies. Let's call the uniform field E_0 , which will still be $E_0 e^{i\omega t}$, and we'll call the correct field as

$$E = E_0 + E_1,$$

where E_1 is the correction due to the changing magnetic field. The way we will write the field at the center of the capacitor as $E_0 e^{i\omega t}$ (barely defining E_0), so that we have no correction at the center: $E_1 = 0$ at $r = 0$.

To find E_1 we can use the integrated form of Faraday's law

$$\oint_C \mathbf{E} \cdot d\mathbf{s} = - \frac{\partial}{\partial t} (\text{flux of } \mathbf{B}).$$

For simplicity, consider if we take the path for the curve Γ shown in Fig. 23-4(b), which goes up along the left edge, outwards by the distance a , then down the right edge to the vertical dashed line to the axis. The line integral of B around this curve is $\oint \mathbf{B} \cdot d\mathbf{l}$, equal to zero, as only B_x contributes, and its integral is just $B_x a$, where a is the separation between the plates. (We will *C* points if it points against \mathbf{dl} ; Γ is oriented to the rate of change of the flux of B , which we have to get by an integration over the shaded area S inside Γ , in Fig. 23-4(a).) The flux through a vertical strip of width dx is $B_x dx$, so the total flux is

$$B_x \int_0^a B_x(x) dx.$$

Setting $-i\omega t$ of the flux equal to the line integral of E_x , we have

$$E_x(x) = -\frac{i}{a} \int_0^x B_x(z) dz. \quad (23.6)$$

Note that $i\omega t$ has been added to the line integral because of the separation of the plates.

Using Eq. (23.5) for $B_x(x)$, we have

$$E_x(x) = \frac{c}{m} \frac{\omega^{1/2}}{4\pi^2} E_0 e^{i\omega t}.$$

The time derivative just brings down another factor $i\omega$, so we get

$$E_x(x) = -\frac{\omega^{1/2}}{2\pi^2} B_0 e^{i\omega t}. \quad (23.7)$$

As we expect, the induced field wants to reduce the electric field, otherwise, the corrected field $E = E_0 + B_0$ is zero.

$$E = E_0 + B_0 = \left(1 - \frac{\omega^{1/2}}{4\pi^2}\right) E_0 e^{i\omega t}. \quad (23.8)$$

The electric field in the capacitor is no longer uniform; it has the parabolic shape shown by the broken line in Fig. 23-5. You see that a simple capacitor is becoming slightly complicated!

We could now use our results to calculate the impedance of the capacitor at high frequencies. Knowing the electric field, we could compute the charges on the plates and then on how the current through the capacitor depends on the frequency ω , but we are not interested in that problem for the moment. We are more interested in seeing what happens as we continue to go up with the frequency ω to see what happens at even higher frequencies. Again, we already find "no". No, because we have corrected the electric field, which means that the magnetic field we have calculated is no longer right. The magnetic field of Eq. (23.5) is approximately right, but it is only a first approximation. So let's call it B_1 . We should then rewrite Eq. (23.5) as

$$B_1 = \frac{i\omega}{2\pi} B_0 e^{i\omega t}. \quad (23.9)$$

You will remember that this field was produced by the variation of E_0 . Now the correct magnetic field will be that produced by the total electric field $E = E_0 + E_1$. If we write the magnetic field as $B = B_1 + B_0$, the second term is just the additional field produced by E_0 . To find B_0 , we can go through the same arguments we have used to find B_1 , the line integral of B_0 around the curve Γ , is equal to the rate of change of the flux of B , through Γ . We will just use Eq. (23.4) again with B replaced by B_0 and A replaced by ΔA :

$$\omega^2 \Delta A / \Delta t = \frac{1}{g} (\text{flux of } B_0 \text{ through } \Gamma).$$

surface inside Γ_1 . Using $2\pi dr$ as the element of area, this integral is

$$\int_{r_0}^r E_2(r') \cdot 2\pi r' dr'$$

So we get for B_{21} :

$$B_{21}(r) = \frac{1}{r_0^2} \frac{\partial}{\partial r} \int_{r_0}^r E_2(r') r' dr'. \quad (23.13)$$

Using $E_2(r)$ from Eq. (23.9), we need the integral of $r^3 dr$, which is, of course, $r^4/4$. The result of the magnetic field becomes

$$B_{21}(r) = -\frac{i\omega^{3/2}}{16\pi^2} E_0 r'^{1/2}. \quad (23.14)$$

But we are still not finished! If the magnetic field B is not the same as we first thought, then we have incorrectly computed E_2 . We must make a further correction to E , which comes from the extra magnetic field B_2 . Let's call this additional correction to the electric field E_3 . It is related to the magnetic field B_2 in the same way that E_2 was related to B_1 . We can use Eq. (23.8) all over again, just by changing the subscripts:

$$E_3(r) = \frac{1}{ir} \int B_2(r') dr'. \quad (23.15)$$

Using our result, Eq. (23.14), for B_2 , the new correction to the electric field is

$$E_3(r) = \pm \frac{i\omega^{1/2}}{64\pi^2} E_0 r'^{1/2}. \quad (23.16)$$

Writing our finally corrected electric field as $E = E_1 + E_2 + E_3$, we get

$$E = E_0 r'^{1/2} \left[1 + \frac{1}{2} \left(\frac{\omega r}{c} \right)^2 + \frac{1}{2} + \frac{1}{48} \left(\frac{\omega r}{c} \right)^4 \right]. \quad (23.17)$$

The variation of the electric field with radius is no longer the simple parabola we drew in Fig. 23-5, but at large radii lies slightly above the curve ($E_1 = E_0$).

We are not quite through yet. The new electric field produces a new correction to the magnetic field, and the newly corrected magnetic field will produce a further correction to the electric field, and on and on. However, we already have all the formulas that we need. For B , we can use Eq. (23.10), changing the subscripts of B and E from 2 to 3.

The next correction to the electric field is

$$E_4 = -\frac{1}{2^2 \cdot 12 \cdot 6^2} \left(\frac{\omega r}{c} \right)^6 E_0 r'^{1/2}.$$

So to this order we have that the complete electric field is given by

$$E = E_0 r'^{1/2} \left[1 + \frac{1}{2} \left(\frac{\omega r}{c} \right)^2 + \frac{1}{2} + \frac{1}{48} \left(\frac{\omega r}{c} \right)^4 - \frac{1}{2^2 \cdot 12 \cdot 6^2} \left(\frac{\omega r}{c} \right)^6 + \dots \right], \quad (23.18)$$

where we have written the corrected coefficients in a way that it is obvious how the series is to be continued.

Our final result is that the starting field between the plates of the capacitor, for any frequency, is given by $E_0 r'^{1/2}$ times the infinite series which contains only the even terms. If we wish, we can define a special function, which we will call $J_0(x)$, as follows (which is precisely in the manner of Eq. (23.18)):

$$J_0(x) = 1 + \frac{1}{(1)^2} \left(\frac{x}{2} \right)^2 + \frac{1}{(2)^2} \left(\frac{x}{2} \right)^4 - \frac{1}{(3)^2} \left(\frac{x}{2} \right)^6 + \dots \quad (23.19)$$

Then we can write our solution as $E_0 e^{i\omega t}$ times this function, with $x = \omega/c$:

$$E = E_0 e^{i\omega t} J_0 \left(\frac{\omega c}{v} \right) \quad (23.17)$$

The reason we have called our special function J_0 is thus, naturally, this is not the first time anyone has ever worked out a problem with oscillations in a cylinder. The function has come up before and is usually called J_0 . It always comes up whenever you solve a problem about waves in a cylindrical geometry. The function J_0 is to cylindrical waves what the cosine function is to waves on a straight line. So it is an important function, invented a long time ago. Then a man named Bessel got his name attached to it. The subscript zero means that Bessel invented a whole lot of different functions and this is just the first of them.

The other functions of Bessel— J_1 , J_2 , and so on—have to do with cylindrical waves which have a variation of their strength with the angle around the axis of the cylinder.

The completely corrected electric field between the plates of our circular capacitor, given by Eq. (23.17), is plotted as the solid line in Fig. 23-5. For frequencies that are not too high, our second approximation was already quite good. The third approximation was even better—so good, in fact, that if we had plotted it, you would not have been able to see the difference between it and the solid curve. You will see in the next section, however, that the complete series is needed to get an accurate description for large radii, or for high frequencies.

23-3 A resonant cavity

We want to look now at what our solution gives for the electric field between the plates of the capacitor as we continue to go to higher and higher frequencies. For large x , the parameter $x = \omega/c$ also gets large, and the first few terms in the series for J_0 of x will increase rapidly. That means that the parabola we have drawn in Fig. 23-3 curves downward more steeply at higher frequencies. In fact, it looks as though the field would fall all the way to zero at some high frequency, perhaps when $c\omega$ is approximately one-half of σ . Let's see whether J_0 does indeed go through zero and become negative. We begin by trying $x = 2$:

$$J_0(2) \approx -1.14 \quad \text{at } \omega = 10^3$$

This value is still not zero, so let's try a higher value of x , say, $x = 2.5$. Putting in our numbers, we write

$$J_0(2.5) \approx -1 - 1.65 - 0.61 - 0.09 \approx -3.04.$$

The function J_0 has already gone through zero by the time we get to $x = 2.5$. Comparing the results for $x = 2$ and $x = 2.5$, it looks as though J_0 goes through zero about half of the way from 2.3 to 2. We would guess that the zero occurs for x approximately equal to 2.3. Let's see what that value of x gives:

$$J_0(2.3) \approx -1 - 1.44 - 0.52 - 0.08 = 3.00$$

We get zero in the same way as our two previous attempts. If we triple the frequency instead of double (or since J_0 is a well-known function, if we took it up a third), we find that it goes through zero at $x = 3.205$. We leave you to try out by hand to show you how you can't have discovered these things without doing some numerical work.

As long as we are looking up J_0 on a table, it is interesting to notice how it goes for large values of x , such as the graph in Fig. 23-6. As longer and longer distances between positive and negative values of x the oscillations damp out of oscillation.

We have given the following interesting test. If we go high enough in frequency, the electric field in the center of our condenser will be negative and the electric field near the edge will point in the opposite direction. For example,

suppose that we take an ω high enough so that $r = \omega/c$ at the outer edge of the capacitor is equal to 4; then the edge of the capacitor corresponds to the a radius $x = 4$ in Fig. 23-6b. This means that our capacitor is being operated at the frequency $\omega = 2\pi c/4$. At the edge of the plates, the electric field will have a rather high magnitude opposite the direction we would expect. That's the terrible thing that can happen to a capacitor at high frequencies. If we go to very high frequencies, the direction of the electric field oscillates back and forth many times as we go out from the center of the capacitor. Also there are the magnetic fields associated with these electric fields. It is not surprising that our capacitor doesn't look like the ideal capacitance for high frequencies. We may even start to wonder whether it looks more like a capacitor or an inductance. We should emphasize that there are even more complicated effects that we have neglected which happen at the edges of the capacitor. For instance, there will be a radiation of waves out past the edges, so the fields are even more complicated than the ones we have computed, but we will not worry about these effects now.

We could try to build an equivalent circuit for the capacitor, but perhaps it is better if we just admit that the capacitor we have designed for low-frequency below is just a lumped element, very much so if the frequency is too high. If we want to treat the equivalent of such an object at high frequencies, we should remember the important features of Maxwell's equations that we have used for creating circuits and return to the complete set of equations which describe completely the fields in space. Instead of dealing with idealized circuit elements, we have to deal with the real conductors as they are, taking into account all the fields in the spaces in between. For instance, if we want a resonant circuit at high frequencies we will not try to design one using a model for a parallel-plate capacitor.

We have already mentioned that the parallel-plate capacitor we have been analyzing has some of the aspects of both a capacitor and an inductance. With the electric field there are charges on the surfaces of the plates, and with the magnetic fields there are rock emfs. Is it possible that we already have a resonant circuit? We do indeed. Suppose we pick a frequency for which the electric field pattern falls to zero at some radius inside the edge of the disc; that is, we choose ω/c greater than 2.405. Everywhere on a circle coaxial with the plates the electric field is to be zero. Now suppose we take a thin metal sheet and cut a strip just wide enough to fit between the plates of the capacitor. Then we bend it into a cylinder that will go around at the radius where the electric field is zero. Since there are no electric fields there, when we put this conducting cylinder in place, no currents will flow in it and there will be no changes in the electric and magnetic fields. We have been able to put a direct short circuit across the capacitor without changing anything. And look what we have: we have a complete cylindrical can with electrical and magnetic fields inside and no connection at all to the outside world. The fields inside won't change even if we turn over the edges of the plates outside our can, and also the capacitor loses. All we have left is a closed can with electric and magnetic fields inside, as shown in Fig. 23-7(a). The electric fields oscillate back and forth at the frequency ω —which, don't forget, determined the diameter of the can. The amplitude of the oscillating E field varies with the distance from the axis of the can, as shown in the graph of Fig. 23-7(b). This curve is just the first arch of the Bessel function of zero order. There is also a magnetic field which goes in circles around the axis and oscillates in time $\sin \omega t$ out of phase $\pi/2$ to the electric field.

We can also write out a series for the magnetic field and plot it, as shown in the graph of Fig. 23-7(c).

How is it that we can have an electric and magnetic field inside a can with no alternating currents? It is because the electric and magnetic fields maintain them selves. In short, if E moves in B and the changing B makes an E , all according to the equations of Maxwell, the magnetic field has its own wave aspect, and the electric field is supported by B ; together they make something like a resonant circuit. Notice that the conditions we have described would only happen if the radius of the can is exactly 2.405 cm. For a can of a given radius, the oscillating electric and magnetic fields will remain in themselves. In the way we have described

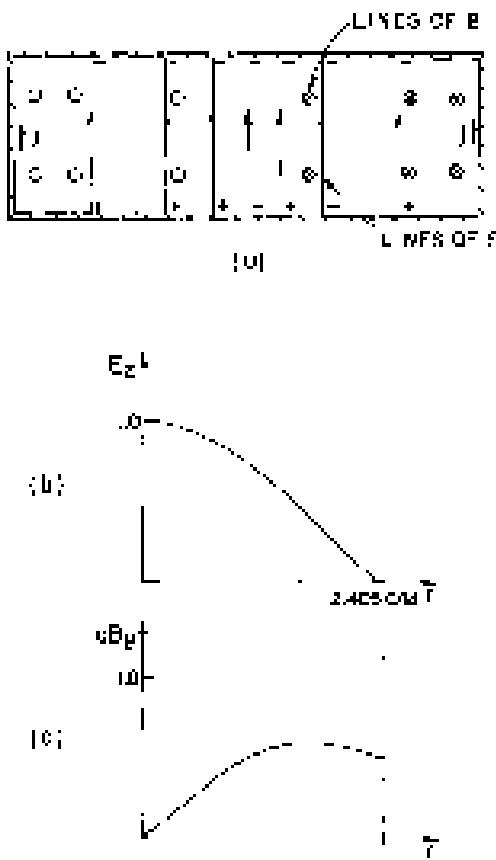


Fig. 23-7. The electric and magnetic fields in an enclosed cylindrical can.

only at that particular frequency. So a cylindrical can of radius a is resonant at the frequency

$$\omega_0 = 2\pi f_0 = \frac{c}{a} \quad (23-18)$$

We have said that the fields continue to oscillate in the same way after the can is completely closed. That is not exactly right. It would be possible if the walls of the can were perfect conductors. For a real can, however, the oscillating currents which exist on the inside walls of the can lose energy because of losses stored in the material. The oscillations of the fields will gradually die away. We can see from Fig. 23-7 that there must be standing waves associated with electric and magnetic field amplitudes in cavity. Because there is real absorption of energy, evidently at the top and bottom plates of the can, it is a logical consequence that there must be positive and negative standing charges on the inner surfaces of the can, as shown in Fig. 23-8(a). When the exterior field reverses, the charges under reversal also reverse, causing an induced dipole moment between the top and bottom plates of the can. These charges will flow in the sides of the can, as shown in the figure. We can also see that an alternating voltage is built up across the cavity by considering what happens to the magnetic field. In a typical Fig. 23-6 it is clear that the magnetic field suddenly drops to zero at the end of the can. Such a sudden change in the magnetic field can happen only if there is a current in the wall. This current is what gives the alternating electric charges on the top and bottom plates of the can.

You may be wondering about that history of currents in the vertical sides of the can. What about our earlier statement that nothing would be changed when we introduce these vertical sides in a region where the electric field was zero? Remember, however, that when we first put in the sides of the can, the top and bottom plates extended out beyond them, so that there were also magnetic fields on the outside of our can. It was only when we threw away the parts of the top/bottom plates beyond the edges of the can that no currents had to appear on the insides of the vertical walls.

Although the electric and magnetic fields in the completely enclosed can will gradually die away because of the energy losses, we can stop this from happening if we make a little hole in the can and put in a little bit of electrical energy to make up the losses. We take a small wire loop through the hole in the side of the can, and fasten it to the inside wall so that it makes a small loop, as shown in Fig. 23-8. If we now connect this wire to a source of high frequency alternating current, this current will couple energy into the electric and magnetic fields of the cavity and keep the oscillations going. This will happen, of course, only if the frequency ω of the drive source is at the resonant frequency of the can. If the source is at the wrong frequency, the electric and magnetic fields will not resonate, and the fields in the can will be very weak.

The resonant behavior can easily be seen by making another small hole in the can and blocking it another coupling loop, as we have also drawn in Fig. 23-8. The changing magnetic field through this loop will generate an induced electromotive force in the loop. If this loop is now connected to some external inductor circuit, the current will be proportional to the strength of the fields in the cavity. Suppose we now connect the input loop of our cavity to an R.F. signal generator, as shown in Fig. 23-9. The signal generator contains a varactor diode, meaning whose frequency can be varied by varying the bias on the diode of the generator. Then we connect the output loop of the cavity to a "detector," which is an instrument that measures the current through the cavity loop. It gives an output signal proportional to this current. If we now increase the current, as a function of the frequency of the signal generator, we find a curve like that shown in Fig. 23-10. The output signal is small for all frequencies except those very near the frequency ω_0 , which is the resonant frequency of the cavity. The resonant frequency ω_0 in fact these we described in Chapter 23 is MHz . The width of the resonance is, however, much narrower than we usually find for resonators, circuits made of resistances and capacitors; that is, the Q of the cavity is very high. It is not unusual to find Q 's as high as 100,000 or more if the inside walls of the cavity are made of some material with a very good constant ϵ_r , such as alumina.

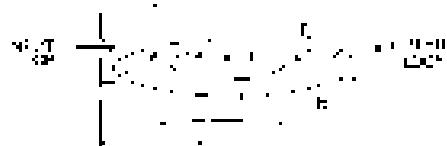


Fig. 23-6. Coupling into and out of a resonant cavity.

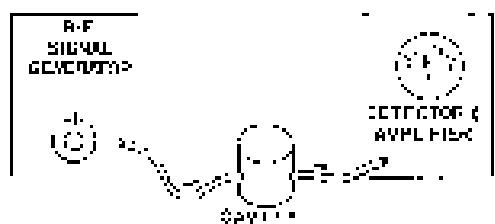


Fig. 23-9. A setup for measuring the cavity resonance.

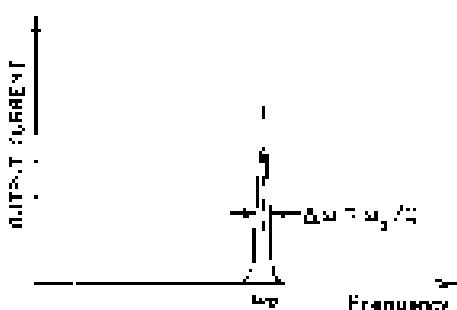


Fig. 23-10. The frequency response curve of a resonant cavity.

Suppose we now try to check our theory by making measurements with an oscilloscope. We take a can which has a diameter of 2.0 inches and a height of about 3.5 inches. The can is filled with an input signal at a frequency as shown in Fig. 23-11. If we take note of the resonant frequency expected for this can according to Eq. 23-10, we get $100\pi f_0 = \omega_0/2\pi = 3016$ megacycles. With the set of ω_0 /frequency of the signal generator at near 3000 megacycles, and vary it slightly up and down the resonance, we observe that the maximum output current occurs for a frequency of 3050 megacycles, which is quite close to the predicted resonant frequency, but not exactly the same. There are several possible reasons for this discrepancy. Perhaps the resonant frequency is changing as a result of some of the voltage or current due to our in the coupling loops. A little thought, however, shows that this does sound likely. The resonant frequency varies a little bit, so that ω_0 and ω the mean. Perhaps there some slight change in frequency of the train of the signal generator or perhaps an expansion of the diameter of the cavity, and oscillatory energy. Any way, the agreement is fairly close.

What more important is something that happens if we vary the frequency of our signal generator somewhat higher than 3000 megacycles. When we do just what we did in Fig. 23-11, we find that, in addition to the resonance we expected near 3000 megacycles, there is another one at 3400 megacycles and one near 3750 megacycles. Why are there two others than one? We might think that from Fig. 23-6, although we have been assuming that the horizontal axis of the Bessel function remains at the edge of the can, it could change due to the second zero of the Bessel function occurring at the edge of the can as the amplitude decreases in the electric field as we move from the center of the can out to the edge as shown in Fig. 23-12. This is another possible cause for this oscillating field. We should again expect the zero to resonate in such a mode. But notice, the second zero of the Bessel function occurs very near which is over 2π in length or over $1/2$ of the radius. The resonant frequency of this mode should therefore be higher than 3000 megacycles. We would not expect this to do, but it does not explain the two more we observe at 3400.

Another possibility is that there may be other boundary conditions or edge which we have considered using only a single parameter, or a portion of the electric field amplitude. We have assumed that the electric fields are vertical and that the magnetic fields remain transverse to the electric field components. The only requirements are that the fields should satisfy Maxwell's equations inside the can, and that the electric field should be zero at the outer right angles. We have considered the case in which the top and bottom of the can are straight, but it is possible that the field boundary conditions at the top and bottom, and which is the case. It is at least possible to show that there is a mode of oscillation of the field's inside the can, in which the electric field is given by the curve of diameter of the can as shown in Fig. 23-13.

It is not too hard to calculate with the natural frequency of this mode should be not very different from the natural frequencies of the first mode we have considered. Suppose the boundary condition along the entire width of the can is that the electric field is zero at the top and bottom. In a mode with the electric field going up and down with constant frequency, the natural frequency is the note in which the electric field was excited right and left. If we now excite the electric field in a mode, we will change the frequency of the note. We would still expect the same. In fact, if we do such, provided we keep the RF boundary of the cavity intact or the top edge, the frequency of the mode of Fig. 23-13 should not be too different from the mode of Fig. 23-4. We could make a rough estimate of the natural frequency of the mode shown in Fig. 23-13, but we do not do that now. When the oscillating amplitude is enough, it is found that for the dimensions we have assumed, the resonant frequency comes out very close to the natural resonance at 3000 megacycles.

By similar reasoning it is possible to show that there should be still another mode at the order $n = 2$. Frequency we found near 3400 megacycles. The ad-

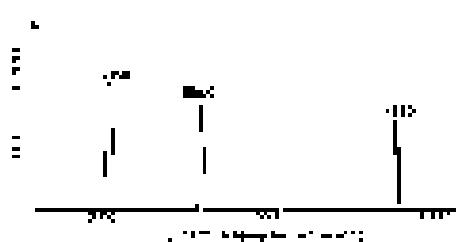


Fig. 23-11. Oscilloscope record for a cavity of cavity 23-1.

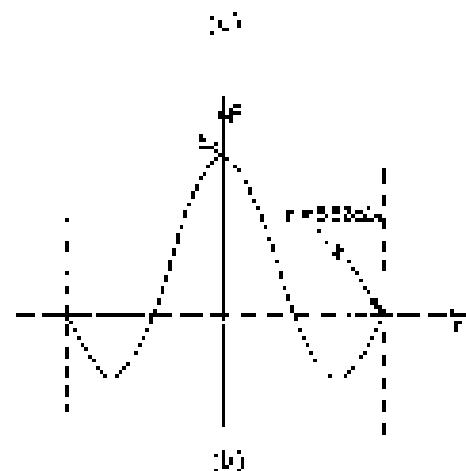
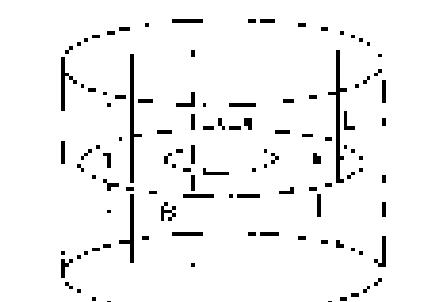


Fig. 23-13. A transverse mode of the cavity 23-1.

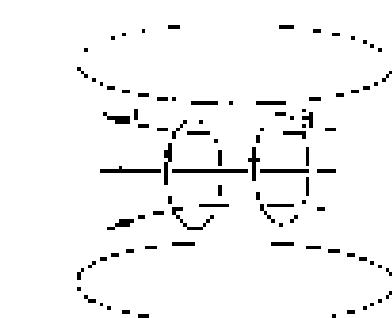


Fig. 23-14. A higher-order mode of the cavity 23-1.

of course, the electric and magnetic fields are as shown in Fig. 23-4. The electric field is zero and Faraday's law gives the tangential flux density. It goes from the axes to the end walls, as shown.

As you will probably know better, if we go higher and higher in frequency we should expect to find more and more resonances. There are many different modes and at a high enough frequency there is a resonance corresponding to some particular spatial placement arrangement of the electric and magnetic fields. Each of these different arrangements is called a resonant mode. The resonance frequency of each mode can be calculated by solving Maxwell's equations for the electric and magnetic fields in the cavity.

When we have a resonance at some particular frequency, how can we know which mode is being excited? One way is to poke a little wire into the cavity through a small hole. If the electric field is along the wire, as in Fig. 23-15(a), there will be relatively large currents in the wire, sapping energy from the field, and the resonance will be suppressed. If the electric field is as shown in Fig. 23-15(b) the wire will have a much smaller effect. We could find which way the field points in this mode by bending the end of the wire, as shown in Fig. 23-15(c). Then, as we move to the wire, there will be a big effect when the end of the wire is parallel to E and a small effect when it is rotated so as to be at 90° to E .

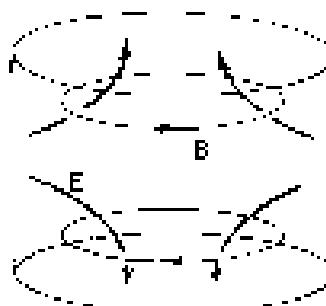


Fig. 23-4. Another view of a cylindrical cavity.

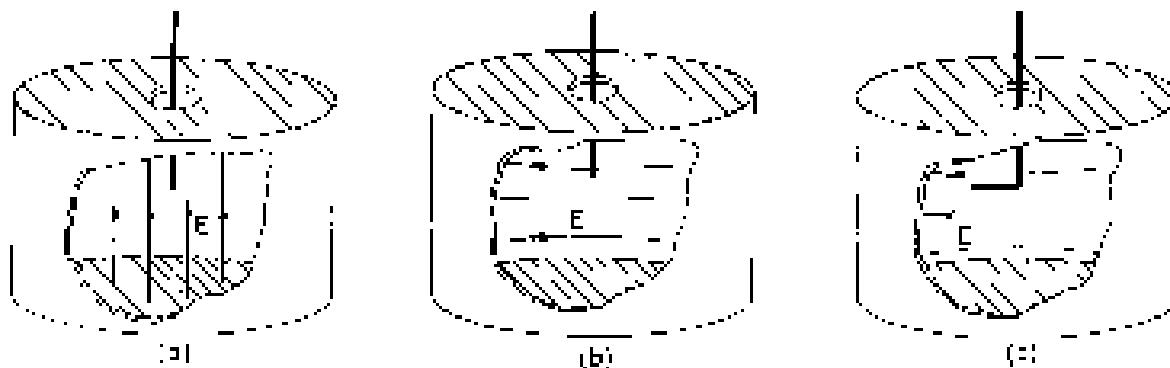


Fig. 23-15. A thin metal wire inserted into a cavity will scrub more power when it is parallel to E than when it is at right angles.

23-5 Cavities and resonant circuits

Although the resonant cavity we have been discussing seems to be quite different from the ordinary resonant circuit consisting of an inductor and a capacitor, the two resonant systems are, of course, closely related. They are both members of the same family; they are just two extreme cases of electromagnetic oscillation—and there are many intermediate cases between these two extremes. Suppose we start by considering the resonant circuit in question in parallel with a capacitor, as shown in Fig. 23-16(a). This circuit will resonate at the frequency $\omega_0 = 1/\sqrt{LC}$. If we want to make the resonant frequency ω_0 this much larger, we can do so by increasing the inductance L . One way is to increase the number of turns in the coil. We can, however, probably see that in the first place. Essentially we will go down to the last term and we will have just a piece of wire, namely, the top and bottom edges of the condenser. We could increase the resonant frequency still further by making the capacitor even smaller; however, we can also continue to increase the inductance by putting several inductances in parallel. Two inductors and another one parallel will triple only half the inductance of each alone. So when we add one we have been reduced to a single turn, but we can add in two more to restore the original inductance by adding other single loops from the top plate to the bottom plate of the condenser. For instance, Fig. 23-16(b) shows the condenser plates separated by six such "single turn inductances"; if we continue to add many more pieces of wire, we can make the transition to the completely enclosed resonant system shown in part (c) of the figure, which is a drawing of the cross section of a cylindrically symmetrical cavity.

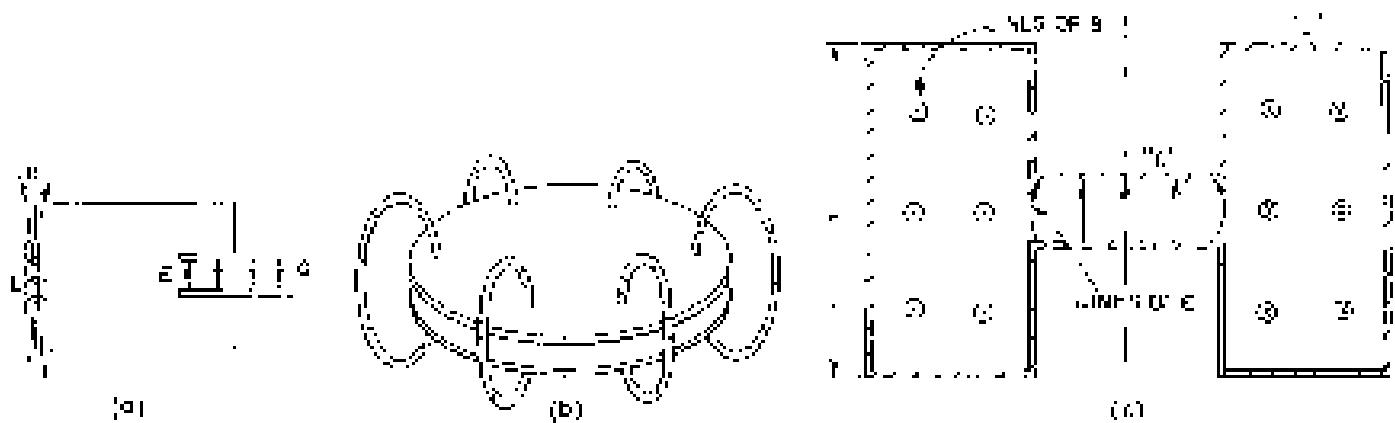


Fig. 23-16. Resonators of progress: (a) higher wave and frequency.

object. Our inductance is now a cylindrical hollow can attached to the edges of the condenser plates. The electric and magnetic fields will be as shown in the figure. Such an object, of course, is resonant. It is called a "hoisted" cavity. But we can still think of it as an $L-C$ circuit in which the capacity section is the region where we find most of the electric field and the inductance section is that region where we find most of the magnetic field.

If we want to make the frequency of the resonator in Fig. 23-16 still higher, we can do so by decreasing its inductance L . To do that, we must decrease the geometric dimensions of the inductance section. For example, by decreasing the dimension a in the drawing. As a is decreased, the resonant frequency will be increased. Eventually, as a comes down to the value a_0 in which the bend B is just equal to the separation between the condenser plates, we shall have just a cylinder or pipe that is resonant at low frequency (resonator of Fig. 24).

You will notice that in the simple $L-C$ circuit of Fig. 13-19, the electric and magnetic fields are quite separate. As we have gradually modified the resonator system by making higher and higher frequencies, the magnetic field has been brought closer and closer to the electric field until, in the cavity resonator, the two are quite intermixed.

Although the cavity resonators we have talked about in this chapter have been cylindrical cans, there is nothing magic about the cylindrical shape. A can of any shape will have resonant frequencies corresponding to various possible modes of oscillations of the electric and magnetic fields. For example, the "cavity" shown in Fig. 23-17 will have its own particular set of resonant frequencies—although they would be rather difficult to calculate.

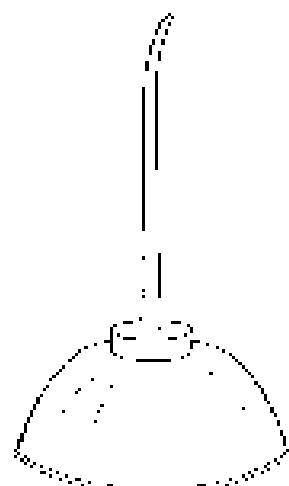


Fig. 23-17. Another resonant cavity.

Waveguides

24-1 The transmission line

In the last chapter we studied what happened to the turned elements of certain resonant circuits when they were operated at very high frequencies, and we were led to see that a resonant circuit could be replaced by a cavity with the fields resonating inside. Another interesting technical problem is the connection of one object to another, so that electromagnetic energy can be transferred between them. At low frequency circuits the connection is made with wires, but this method doesn't work very well at high frequencies because the circuits would radiate energy into all the space around them, and it is hard to control where the energy will go. The fields spread out, are and die away; the currents and voltages are not "guided" very well by the wires. In this chapter we want to look into the ways that objects can be interconnected at high frequencies. At least, that's one way of presenting our subject.

Another way is to say that we have been discussing the behavior of waves in free space. Now it is time to see what happens when oscillating fields are confined in one or more dimensions. We will discover the fascinating new phenomena when the fields are confined in only two dimensions and allowed to go free in the third dimension, they propagate in waves. These are "guided waves"—the subject of this chapter.

We begin by discussing only the general theory of the transmission line. The arbitrary power transmission line that runs from tower to tower over the countryside radiates away some of its power, but the power requirements (50–60 cycles/sec) are so low that this loss is not serious. The radiation could be stopped by surrounding the line with a metal pipe, but this method would not be practical for power lines because the voltage and current used would require a very large, expensive, and heavy pipe. So simple "open lines" are used.

For somewhat higher frequencies—say, a few kilocycles—radiation can already be serious. However, it can be reduced by using "twisted pair" transmission lines, as in telephone transmission. At high frequencies, however, the radiation soon becomes unacceptable, either as a loss of power losses or because the energy appears in other directions where it isn't wanted. Far frequencies from a few kilocycles to some hundreds of megacycles, electronic guitar signals and power are usually transmitted via coaxial lines consisting of a wire inside a cylindrical "outer conductor" or "shield." Although the following treatment will apply to a transmission line of two parallel conductors of any shape, we will carry it out referring to a coaxial line.

We take the simplest coaxial line that has a central conductor, whose top pole is at the top of the long cylinder, and an outer conductor which is another thin cylinder on the same axis as the inner conductor, as in Fig. 24-1. We begin by figuring out approximately how the line behaves at relatively low frequencies. We have already described some of the low-frequency behavior when we said earlier that two such conductors had a certain amount of inductance per unit length and a certain capacity per unit length. We can, in fact, describe the low-frequency behavior of any transmission line by giving its inductance per unit length, L_0 , and its capacity per unit length, C_0 . Then we can analyze the line as the lumped-case of the $L-C$ filter as discussed in Section 22-6. We can make a filter which simulates the line by taking small series elements $L_0 \Delta x$ and small shunt capacities $C_0 \Delta x$, where Δx is an element of length of the line. Using our results for the infinite filter, we see that there would be a propagation of electric

24-1 The transmission line

24-2 The rectangular waveguide

24-3 The cutoff frequency

24-4 The speed of the guided waves

24-5 Observing guided waves

24-6 Waveguide plumbing

24-7 Waveguide modes

24-8 Another way of looking at the guided waves

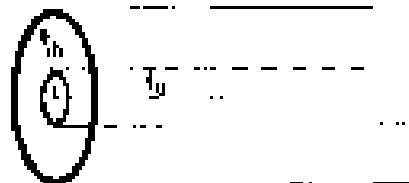


Fig. 24-1. A coaxial transmission line.

signals along the line. Rather than following that approach, however, we would now rather look at the line from the point of view of a differential equation.

Suppose that we see what happens at two neighbouring points along the transmission line, say at the distances x and $x + \Delta x$ from the beginning of the line. Let's call the voltage difference between the two conductors $V(x)$, and the current along the "hot" conductor $I(x)$ (see Fig. 24.2). If the current in the line is varying, the inductance will give us a voltage drop across the small section of line from x to $x + \Delta x$ in the amount

$$\Delta V = V(x + \Delta x) - V(x) = -L_0 \Delta x \frac{dI}{dx}.$$

Or, taking the limit as $\Delta x \rightarrow 0$, we get

$$\frac{dV}{dx} = -L_0 \frac{dI}{dx}. \quad (24.1)$$

The changing current gives a gradient of the voltage.

Referring again to the figure, if the voltage at x is changing, there must be some charge supplied to the capacitor in that region. If we take the small piece of line between x and $x + \Delta x$, the charge on it is $q = C_0 \Delta x V$. The total electric charge of this element is $C_0 \Delta x dV/dx$, but the charge changes only if the current $I(x)$ goes through it, different from the current $I(x + \Delta x)$ just. Calling the difference ΔI , we have

$$\Delta I = -C_0 \Delta x \frac{dV}{dx}.$$

Taking the limit as $\Delta x \rightarrow 0$, we get

$$\frac{dI}{dx} = -C_0 \frac{dV}{dx}. \quad (24.2)$$

So the conservation of charge implies that the gradient of the current is proportional to the time rate-of-change of the voltage.

Equations (24.1) and (24.2) are then the basic equations of a transmission line. If we wish, we could neatly do a to discuss the effects of resistance in the conductors or of leakage of charge through the insulation between the conductors, but for our present discussion we will just stay with the simple example.

The two transmission-line equations can be combined by differentiating one with respect to x and the other with respect to x and differentiating either V or I . Then we have either

$$\frac{\partial^2 V}{\partial x^2} = C_0 L_0 \frac{\partial^2 I}{\partial x^2}, \quad (24.3)$$

or

$$\frac{\partial^2 I}{\partial x^2} = C_0 L_0 \frac{\partial^2 V}{\partial x^2}. \quad (24.4)$$

Once more we recognize the wave equation in x . For a finite transmission line, the voltage (and current) propagates along the line as a wave. The voltage along the line is then of the form $V(x, t) = f(z - vt)$ or $V(x, t) = f(z + vt)$, or a sum of both. Now what is the velocity v ? We know that the coefficient a^2 in the $\partial^2/\partial x^2$ term is just $1/c^2$, so

$$v = \frac{1}{\sqrt{L_0 C_0}}. \quad (24.5)$$

We will leave it for you to show that the voltage for each wave in a line is proportional to the current of that wave and that the constant of proportionality is just the characteristic impedance z_0 . Calling V_0 and I_0 the voltage and current for a wave going in the plus x -direction, you should get

$$V_0 = z_0 I_0. \quad (24.6)$$

Similarly, for the wave going toward minus x the relation is

$$V_{-} = -z_0 f_{-}$$

The characteristic impedance z_0 we found out from our other equations—is given by

$$z_0 = \sqrt{\frac{L_0}{C_0}}, \quad (24.7)$$

and is, therefore, a pure resistance.

To find the propagation speed v and the characteristic impedance z_0 of a transmission line, we have to know the inductance and capacity per unit length. We can calculate them easily for a coaxial cable, so we will see how that goes. For the inductance we follow the ideas of Section 17-8, and set $\oint B \cdot d\ell$ equal to the magnetic energy which we get by integrating $\epsilon_0 r^2 B^2 / 2$ over the volume. Suppose that a coaxial conductor carries the current I . Then we know that $B = I/2\pi r \mu_0$, where r is the distance from the axis. Taking as a volume element a cylindrical shell of thickness dr and of length l , we have for the magnetic energy

$$U = \frac{\mu_0 l^2}{2} \int_{a/2}^{b/2} \left(\frac{I}{2\pi r \mu_0 c^2} \right)^2 / 2 \pi r dr,$$

where a and b are the radii of the inner and outer conductors, respectively. Carrying out the integral, we get

$$U = \frac{l^2}{4\pi \epsilon_0 c^2} \ln \frac{b}{a}. \quad (24.8)$$

Setting the energy equal to $\frac{1}{2} I^2 / R$, we find

$$L = \frac{l}{2\pi \epsilon_0 c^2} \ln \frac{b}{a}. \quad (24.9)$$

It is, as it should be, proportional to the length l of the line, so the inductance per unit length L_0 is

$$L_0 = \frac{\ln(b/a)}{2\pi \epsilon_0 c^2}. \quad (24.10)$$

We have worked out the charge on a cylindrical conductor (see Section 12-2). Now, dividing the charge by the potential difference, we get

$$C_0 = \frac{2\pi \epsilon_0 l}{\ln(b/a)}.$$

The capacity per unit length C_0 is $1/l$. Combining this result with Eq. (24.10), we see that the product $L_0 C_0$ is just equal to $1/c^2$, so $c = \sqrt{L_0 C_0}$ is equal to v . The wave travels down the line with the speed of light. We point out that this result depends on our assumptions: (a) that there are no dielectrics or magnetic materials in the space between the conductors, and (b) that the currents flow all on the surfaces of the conductors (as they would in the perfect conductors). We will see later that for good conductors at high frequencies, and currents distributed themselves on the surfaces as they would form a perfect conductor, so this assumption is then valid.

Now it is interesting that so long as assumptions (a) and (b) are correct, the product $L_0 C_0$ is equal to $1/c^2$ for any parallel pair of conductors—even, say, for a hexagonal inner conductor anywhere inside an elliptical outer conductor. So long as the cross section is constant and the space between has no material, waves are propagated at the velocity of light.

No such general statement can be made about the characteristic impedance. For the coaxial line, it is

$$z_0 = \frac{\pi (b/a)}{2\pi \epsilon_0}, \quad (24.11)$$

The Dielectric Line has the dimensions of a resistance and is equal to Ω ohms. The conductive Factor (I_{conduct}) depends only logarithmically on the dimensions, so for the coaxial line—a $\lambda/2$ waveguide—the characteristic impedance has typical values of the $\pi/2$ ohms up to a few hundred ohms.

24-2 The rectangular waveguide

The next thing we want to talk about seems at first sight to be a striking pleasure memoir: if the central conductor is removed from the waveguide, it can still carry electromagnetic waves. In other words, at high enough frequencies a hollow tube will work just as well as one with wires. It is related to the mysterious way in which a resonator circuit or a condenser and voltage pole replaced by nothing but a string at high frequencies.

A thought may seem to be a reasonable thing when one has been thinking in terms of a transmission line or a distributed inductor and capacitor, we all know that electromagnetic waves can travel along inside a hollow metal pipe. If the pipe is straight, we can see through it. So certainly electromagnetic waves go through a pipe. But we also know that it is not possible to transmit low-frequency waves (power or telephone) through the inside of a single metal pipe. So it must be that high-frequency waves will go through it over a wavelength so short enough. Therefore we want to discuss the limiting case of the longest wavelength for the lowest frequency that can get through a pipe of a given size. Since the pipe is then being used to carry waves, it is called a *waveguide*.

We will begin with a rectangular pipe, because it is the simplest case to analyze. We will first give a mathematical treatment and come back later to look at the problem in a much more elementary way. The more elementary approach, however, can be applied easily only to a rectangular guide. The basic phenomena are the same for a general guide of arbitrary shape, so the mathematical argument is fundamentally more sound.

The problem then is to find what kind of waves can exist inside a rectangular pipe. Let's first choose some convenient coordinates: we take the x -axis along the length of the pipe, and the y - and z -axes parallel to the two sides, as shown in Fig. 24-3.

We know that when light waves go down the pipe, they have a transverse electric field; so suppose we look first for solutions in which E is perpendicular to z , say with only a y -component, E_y . This electric field will have some variation across the guide; in fact, it must go to zero at the sides parallel to the y -axis, because the currents and charges in a conductor always adjust themselves so that there is no tangential component of the electric field at the surface of a conductor. So E_y will vary with y in some such way shown in Fig. 24-4. Perhaps it is the Bessel function we found for a cavity? No, because the Bessel function has to do with cylindrical geometries. For a rectangular geometry, waves are usually simple harmonic functions, so we should try something like $\sin k_y y$.

Since we want waves that propagate down the guide, we expect the field to alternate between positive and negative values as we go along in x , as in Fig. 24-5, and these oscillations will travel along the guide with some velocity v . If we have oscillations at some definite frequency ω , we would guess that the wave might vary with x like $\cos(\omega t - k_x x)$, or to use the more convenient mathematical form, like $\sin(k_x x - \omega t)$. Thus x -dependence represents a wave travelling with the speed $v = \omega/k_x$ (see Chapter 29, Vol. I).

So we might guess that the wave in the guide would have the following mathematical form:

$$E_y = E_0 \sin(k_x x) e^{j(k_y y - \omega t)} \quad (24.12)$$

Let's see whether this guess satisfies the relevant requirements. First, the electric field should have no tangential components at the conductors. Our field satisfies this requirement; it is perpendicular to the top and bottom faces and is zero at the two side faces. Well, it is if we choose k_y so that one-half a cycle of 24-4

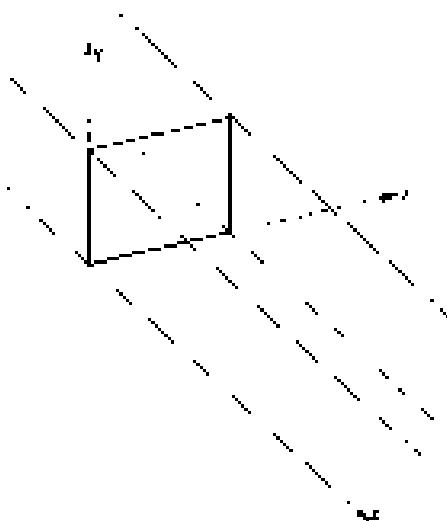


Fig. 24-3. Coordinate system for the rectangular waveguide.

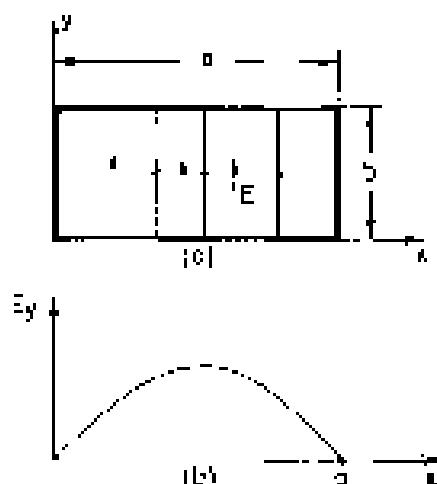


Fig. 24-4. The electric field in the waveguide at some value of z .

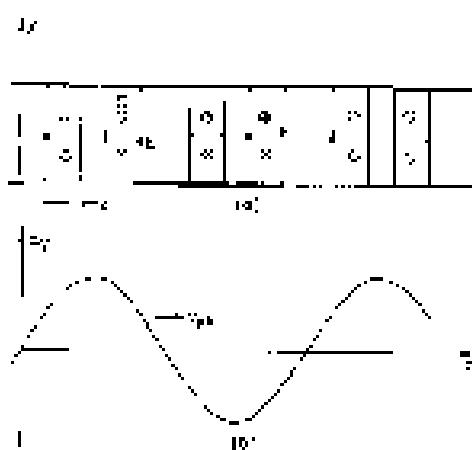


Fig. 24-5. The x -dependence of the field in the waveguide.

sin $k_x z$ just fits in the width of the guide—that is, a^2

$$k_{x0} = \pi/a. \quad (24.15)$$

There are other possibilities, like $k_{x0} = 2\pi, 3\pi, \dots$, or, in general

$$k_{x0} = n\pi, \quad (24.16)$$

where n is any integer. These represent various complicated arrangements of the field, but for now let's take only the smallest one, where $k_x = \pi/a$, where a is the width of the inside of the guide.

Next, the divergence of \mathbf{B} must be zero in the free space inside the guide, since there are no charges there. Our \mathbf{E} has only a y component, and it doesn't change with x , so we do have that $\nabla \cdot \mathbf{E} = 0$.

Finally, our electric field must agree with the rest of Maxwell's equations in the free space inside the guide. That is the same thing as saying that it must satisfy the wave equation:

$$\frac{\partial^2 E_y}{\partial x^2} = \frac{\partial^2 E_y}{\partial y^2} = \frac{\partial^2 E_y}{\partial z^2} = \frac{1}{c^2} \frac{\partial^2 E_y}{\partial t^2} = 0. \quad (24.17)$$

We have to see whether our guess, Eq. (24.12), will work. The second derivative of E_y with respect to x is just $-k_x^2 E_y$. The second derivative with respect to y is zero, since nothing depends on y . The second derivative with respect to z is $-k_z^2 E_y$, and the second derivative with respect to t is $-\omega^2 E_y$. Equation (24.17) then says that

$$k_x^2 E_y + k_z^2 E_y - \frac{\omega^2}{c^2} E_y = 0.$$

Unless E_y is zero everywhere (which is not very interesting), this equation is correct if

$$k_x^2 + k_z^2 - \frac{\omega^2}{c^2} = 0. \quad (24.18)$$

We previously fixed k_x , so this equation tells us that there can be waves of the type we have assumed if k_z is related to the frequency ω so that Eq. (24.18) is satisfied. In other words, if

$$k_z = \sqrt{(\omega^2/c^2) - (\pi^2/a^2)}, \quad (24.19)$$

the waves we have described are propagated in the z direction with this value of k_z .

The wave number k_z we get from Eq. (24.17) tells us, for a given frequency ω , the speed with which the nodes of the wave propagate down the guide. The phase velocity is

$$v = \frac{\omega}{k_z}. \quad (24.20)$$

You will remember that the wavelength λ_0 of a traveling wave is given by $\lambda = 2\pi/v$, so k_z is also equal to $2\pi/\lambda_0$, where λ_0 is the wavelength of the oscillations along the z direction—the guide wavelength. The wavelength in the guide is different, of course, from the free-space wavelength λ_0 of electromagnetic waves of the same frequency. If we call the free-space wavelength λ_0 , which is equal to $2\pi/\omega$, we can write Eq. (24.17) as

$$k_z = \frac{\lambda_0}{\sqrt{1 - (\lambda_0/2a)^2}}. \quad (24.21)$$

Besides the electric fields there are magnetic fields that will travel with the wave, but we will not bother to work out an expression for them right now. Since $c^2 \mathbf{E} \times \mathbf{B} = \mu_0 \mathbf{J}_0$, the lines of \mathbf{B} will circulate around the regions in which $\mu_0 \mathbf{J}_0$ is largest, that is, halfway between the maximum and minimum of \mathbf{E} . The loops of \mathbf{B} will be parallel to the xy plane and between the crests and troughs of \mathbf{E} , as shown in Fig. 24-6.

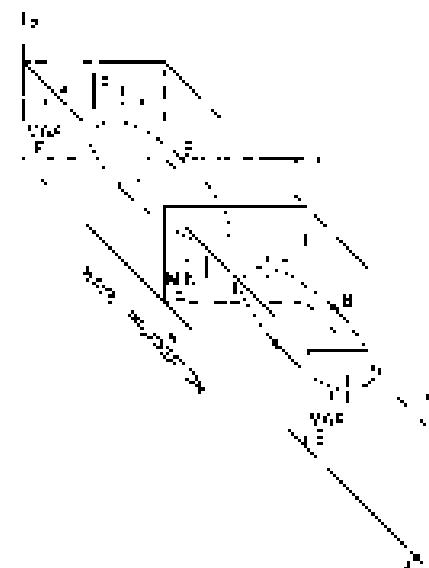


Fig. 24-6. The magnetic field in the waveguide.

24.3 The cutoff frequency

In solving Eq. (24.16) for k_x , there should really be two roots—one plus and one minus. We should write

$$k_x = -\sqrt{(\omega^2/c^2) - (\epsilon_0 \omega^2)} \quad (24.20)$$

The two signs simply mean that there can be waves which propagate with a negative phase velocity (backward $-v$) as well as waves which propagate in the plus v direction in the guide. Naturally, it would be possible for waves to go in either direction. Since both types of waves can be present at the same time, there will be the possibility of standing-wave oscillations.

Our equation for k_x also tells us that higher frequencies give larger values of k_x , and therefore smaller wavelengths. But, in the limit of large ω , k becomes equal to ω/c , which is the value we would expect for waves in free space. The light we "see" through a pipe still travels at the speed c . But now notice that if we go toward low frequencies, something strange happens. At first the wavelength gets longer and longer, but if ω gets too small the quantity inside the square root of Eq. (24.20) suddenly becomes negative. This will happen, as soon as ω gets no less than ω_{c} —that is, when k_x becomes greater than ω/c . In other words, when the frequency gets smaller than a certain critical frequency $\omega_{\text{c}} = \omega_{\text{c}}/\epsilon_0$, the wave number k_x (and also k_y) becomes imaginary and we haven't got a solution any more. Oh dear! Who's it that k has to be real? What if it does come out imaginary? Our field equations are still satisfied. Perhaps an imaginary k also represents a wave?

Suppose ω is less than ω_{c} ; then we can write

$$k_x = \pm ik' \quad (24.21)$$

where k' is a positive real number.

$$k' = \sqrt{\epsilon_0^2/c^2} = (\omega^2/c^2)^{1/2} \quad (24.22)$$

If we now go back to our expression, Eq. (24.13), for E_y , we have

$$E_y = E_0 \sin k_{y0} e^{ik' z_0} e^{-k' z}, \quad (24.23)$$

which we can write as

$$E_y = E_0 \sin k_{y0} e^{ik' z_0} e^{-k' z}. \quad (24.24)$$

This expression gives an E -field that oscillates with time as $e^{i\omega t}$ but which varies with z as $e^{-k' z}$. It decreases or increases with z smoothly as a real exponential. In our derivation we didn't worry about the source that started the waves, but there must, of course, be a source somewhere in the guide. The sign that goes with k' must be the one that makes the field decrease with increasing z away from the source of the waves.

So for frequencies below $\omega = \omega_{\text{c}}$, waves do not propagate down the pipe; the oscillating fields penetrate into the guide only a distance of the order of $1/k'$. For this reason, the frequency ω_{c} is called the "cutoff frequency" of the guide. Looking at Eq. (24.22), we see that for frequencies just a little below ω_{c} , the number k' is small and the fields can penetrate a long distance into the guide. But if ω is much less than ω_{c} , the exponential coefficient, k' , is equal to π/a so the field dies off extremely rapidly, as is shown in Fig. 24-7. The field decreases by 1/e in the distance a/k' , or in only about one-third of the guide width. The fields penetrate very little farther than this distance.

We want to emphasize an interesting feature of our analysis of the guided waves—the appearance of an imaginary wave number k_x . Normally, if we solve an equation in physics and get an imaginary number, it doesn't mean anything physical. For waves, however, an imaginary wave number does mean something. The wave equation is still satisfied; it only means that the solution gives exponentially decaying fields instead of propagating waves. So in any wave problem where k becomes imaginary for some frequency, it means that the form of the wave changes—the sine wave changes into an exponential.

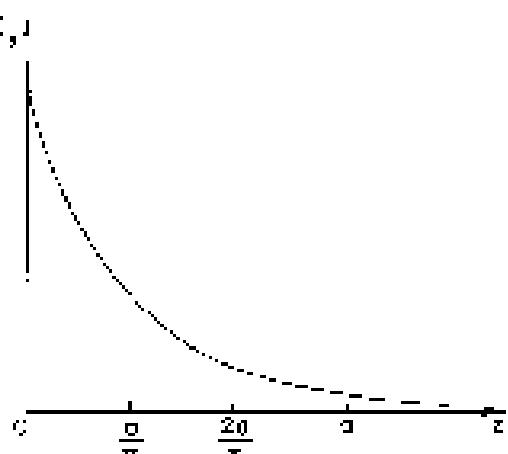


Fig. 24-7. The variation of E_y with z for $\omega < \omega_{\text{c}}$.

24-4 The speed of the guided waves

The wave velocity we have used above is the phase velocity, which is the speed of a mode of the wave; it is a function of frequency. If we combine Eqs. (24.17) and (24.18), we can write

$$v_{\text{phase}} = \frac{c}{\sqrt{1 - (\omega_0/\omega)^2}}. \quad (24.25)$$

For frequencies above cutoff—where travelling waves exist— ω/ω_0 is less than one, and v_{phase} is real and greater than the speed of light. We have already seen in Chapter 25 of Vol. I that phase velocities greater than light are possible, because it is just the locus of the wave, which are moving and not energy or information. In order to know how fast signals will travel, we have to calculate the speed of pulses or modulations made by the interference of a wave of one frequency with one or more waves of slightly different frequencies (see Chapter 18, Vol. II). We have called the speed of the envelope of such a group of waves the group velocity; it is not with our *disks*!

$$v_{\text{group}} = \frac{d\omega}{dk}. \quad (24.26)$$

Taking the derivative of Eq. (24.17) with respect to ω and inverting to $k = \omega/c$, we find that

$$v_{\text{group}} = c \sqrt{1 - (\omega_0/\omega)^2}, \quad (24.27)$$

which is less than the speed of light.

The geometric mean of v_{phase} and v_{group} is just c , the speed of light:

$$\text{geometric mean} = c^{\frac{1}{2}}. \quad (24.28)$$

This is curious, because we have seen a similar relation in quantum mechanics. For a particle with any velocity— even relativistic—the momentum p and energy E are related by

$$E^2 = p^2c^2 + m^2c^4. \quad (24.29)$$

Now, in quantum mechanics the energy is $\hbar\omega$, and the momentum is \hbar/k , which is equal to $\hbar k$; so Eq. (24.29) can be written

$$\frac{\omega^2}{c^2} - k^2 = \frac{m^2c^2}{\hbar^2}. \quad (24.30)$$

$$k = \sqrt{(\omega^2/c^2) - (m^2c^2/\hbar^2)}, \quad (24.31)$$

which looks very much like Eq. (24.17). . . interesting!

The group velocity of the waves is c as the speed at which energy is transported along the guide. If we want to find the energy flow down the guide, we can get it from the energy density times the group velocity. If the root-mean-square electric field is E_0 , then the average density of electric energy is $\epsilon_0 E_0^2/2$. There is also some energy associated with the magnetic field. We will not prove it here, but in any cavity or guide the magnetic and electric energies are equal, so the total electromagnetic energy density is $\epsilon_0 E_0^2$. The power *down* a section of the guide is then

$$\frac{dU}{dx} = \epsilon_0 E_0^2 v_{\text{group}}. \quad (24.32)$$

(We will see later another, more general way of stating the *conservation law*.)

24-5 Observing guided waves

Energy can be coupled into a waveguide by some kind of an "antenna." For example, a little vertical wire or "stub" will do. The presence of the guided waves can be observed by picking up some of the electromagnetic energy with a little receiving "antenna," which again can be a little stub of wire or a small loop.

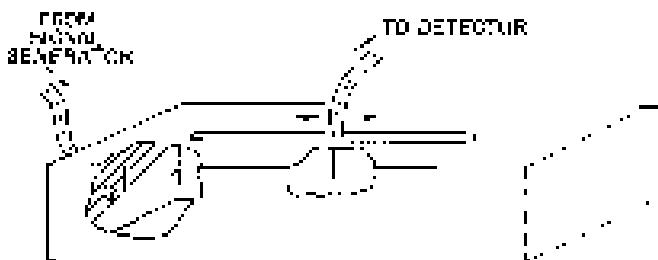


Fig. 24-8. A waveguide with a driving stub and a pickup probe.

In Fig. 24-8, we show a guide with some arrows to show a driving stub and a pickup "probe". The driving stub can be connected to a signal generator via a coaxial cable, and the pickup probe can be connected by a similar cable to a detector. It is usually convenient to insert the pickup probe via a long thin slot in the guide, as shown in Fig. 24-8. Then the probe can be moved back and forth along the guide to sample the fields at various positions.

If the signal generator is set at some frequency ω greater than the cutoff frequency ω_c , there will be waves propagated down the guide from the driving stub. There will be the only waves present if the guide is infinitely long, which can effectively be arranged by terminating the guide with a carefully designed absorber in such a way that there are no reflections from the far end. Then, since the detector measures the time average of the fields near the probe, it will pick up a signal which is independent of the position along the guide; its output will be proportional to the power being transmitted.

If now the far end of the guide is terminated off in some way that produces a reflected wave—say, an extreme example, if we close it off with a metal plate—there will be a reflected wave in addition to the original forward wave. These two waves will interfere and produce a standing wave in the guide similar to the standing waves on a string which we discussed in Chapter 9 of Vol. 2. Each as the pickup probe is moved along the line, the detector reading will rise and fall periodically, showing a maximum in the field at each node of the standing wave and a minimum at each anti-node. The distance between two successive nodes (or crests) is just $\lambda/2$. This gives a convenient way of measuring the guide wavelength. If the frequency is now moved closer to ω_c , the distances between nodes increase, showing that the guide wavelength increases as predicted by Eq. (24-19).

Suppose now the signal generator is set at a frequency just a little below ω_c . Then the detector output will decrease gradually as the pickup probe is moved down the guide. If the frequency is set somewhat lower, the field strength will fall rapidly, following the curve of Fig. 24-7, one showing that waves are not propagated.

24-6 Waveguide plumbing

An important practical use of waveguides is for the transmission of high-frequency power, e.g., for example, in coupling the high-frequency oscillator or output amplifier of a radar to an antenna. In fact, the antenna itself usually consists of a parabolic reflector fed at its focus by a waveguide. Part of it at the end is made a "horn," that radiates the waves coming down the guide. Although high frequencies can be transmitted along a coaxial cable, a waveguide is better for transmitting large amounts of power, first, the maximum power that can be transmitted along a line is limited by the breakdown of the insulation between the conductors. For a given amount of power, the field strengths in a guide are usually less than they are in a coaxial cable, so higher powers can be transmitted before breakdown occurs. Second, the power losses in the coaxial cable are usually greater than in a waveguide. In a coaxial cable there must be insulating material to support the central conductor, and there is an energy loss in this material—particularly at high frequencies. Also, the current densities on the central conductor are quite high, and since the losses go as the square of the current density, the lower currents that appear on the walls of the guide result in lower

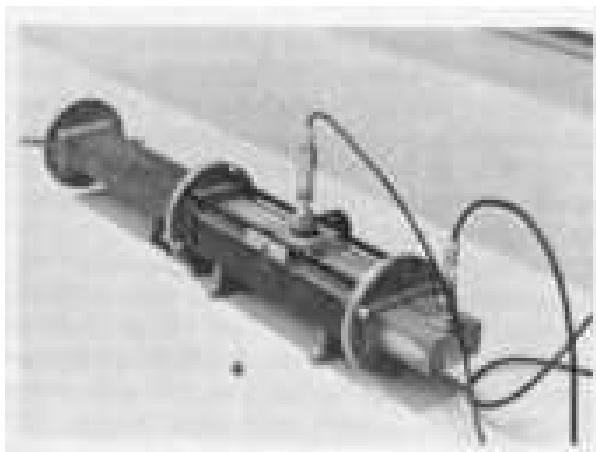


Fig. 24-9. Section of waveguide connected with flanges.

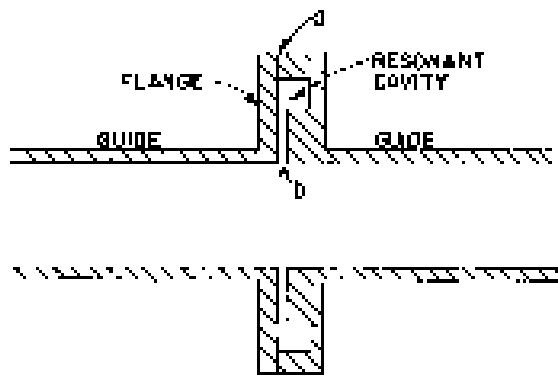


Fig. 24-10. A low-loss connection between two sections of waveguide.

energy losses. To keep these losses as minimum, the inner surfaces of the guide are often plated with a material of high resistivity, such as silver.

The problem of connecting a "circuit" with waveguides is quite different from the corresponding circuit problem at low frequencies, and is usually called *waveguide "ply along"*. Many special devices have been developed for this purpose. For instance, two sections of waveguide are usually connected together by means of flanges, as can be seen in Fig. 24-9. Such connections can, however, cause serious energy losses, because the surface currents must flow across the joint which may have a relatively high resistance. One way to avoid such losses is to make the flanges as shown in the cross-section chart in Fig. 24-10. A small gap is left between the adjacent sections of the guide, and a groove is cut in the face of one of the flanges to make a small cavity of the type shown in Fig. 24-10(a). The dimensions are chosen so that this cavity is resonant at the frequency being used. This resonant cavity presents a high "impedance" to the currents, so no relatively little current flows across the metallic joints (as is in Fig. 24-10). The high guide currents simply charge and discharge the "capacitance" of the gap (or *d* in the figure), where there is little dissipation of energy.

Suppose you want to stop a waveguide in a way that won't result in reflected waves. Then you must put something at the end that imitates an infinite length of guide. You need a *terminator*, i.e., something that absorbs the arriving waves without making reflections. Then the guide will act as though it went on forever. Such terminations are made by putting inside the guide some wedges of resistance material carefully designed to absorb the wave energy while generating almost no reflected waves.

If you want to connect three things together—for instance, one source to two different antennas—then you can use a "T" like the one shown in Fig. 24-11. Power fed into the center section of the "T" will be split and go into the two side arms (and there may also be some reflected waves). You can see qualitatively from the sketches in Fig. 24-12 that the fields would spread out when they get to the end of the input section and make electric fields that will start waves going out the two arms. Depending on whether electric fields in the guides are parallel or perpendicular to the "top" of the "T," the fields at the junction would be roughly as shown in (a) or (b) of Fig. 24-12.

Finally, we would like to measure how much energy "leaks out" of a waveguide, which is very useful for telling what is going on after you have connected a complicated arrangement of waveguides. Suppose you want to know which way the waves are going in a particular section of guide; you might be wondering, for instance, whether or not there is a strong reflected wave. The total reflected power is taken as a small fraction of the power of a guide if there is a wave going one way, but none of the wave is going the other way. By connecting the output of the coupler to a detector, you can measure the "one-way" power in the guide.



Fig. 24-11. A waveguide "T". [The flanges have plastic end caps to keep the inside clean while the "T" is not being used.]

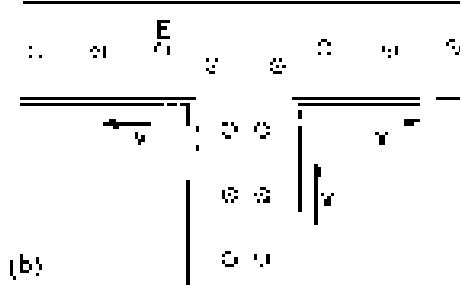
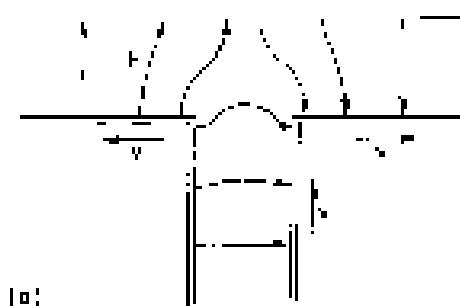


Fig. 24-12. The electric fields in a waveguide "T" for two parallel field orientations.

Figure 24-12 is a drawing of a unidirectional coupler. A piece of waveguide AB has another piece of waveguide CD soldered to it along one face. The guides CD is moved away so that there is room for the connecting flanges. Before the guides can solder together, two (or more) holes have been drilled in each guide (not let up) and then so that some of the fields in the main guide AB can be coupled out to the secondary guide CD. Each of the holes acts like a little antenna that generates a wave in the secondary guide. If there were only one hole, waves would be sent in both directions and would be the same no matter which way the wave was going in the primary guide. But when there are two holes with a separation equal to one-quarter of the guide wavelength, they will make two sources *out of phase*. Do you remember that we considered in Chapter 29 of vol. I the interference of the waves from two antennas spaced $\lambda/4$ apart and excited 90° out of phase at time? We found that the waves interfere in one direction and add in the opposite direction. The same thing will happen here. The wave produced in the guide CD will be going in the same direction as the wave in AB.

If the wave in the primary guide is travelling from A toward B, there will be a wave at the output D of the secondary guide. If the wave in the primary guide goes from B toward A, there will be a wave going toward the end C of the secondary guide. This end is equipped with a terminal or, so that this wave is absorbed and there is no wave at the output of the coupler.

24-7 Waveguide modes

The wave we have chosen to analyze is a special solution of the field equations. There are many more. Each solution is called a waveguide "mode." For example, if the dependence of the field was just one-half a cycle of a sine wave, then a $\pi/2$ equally good solution with a full cycle, then the variation of E_y with x is as shown in Fig. 24-14. The k_x for such a mode is twice as large, so the cutoff frequency is much higher. Also, in the wave we studied, E has only a y -component, but there are other modes with more complicated electric fields. If the electric field has components only in y and z , so that the total electric field is always at right angles to the x -direction, the mode is called a "transverse electric" (TE) mode. The magnetic field of such modes will always have a z -component. If I look at that if E has a component in the x -direction (along the direction of propagation), then the magnetic field will always have only transverse components. So such fields are called transverse magnetic (TM) modes. For a rectangular guide, all the other modes have a single cutoff frequency, but in the simple TE mode we were described, it is, therefore, possible—and usual—to use a guide with a frequency just above the cutoff for the lowest mode but below the cutoff frequency for all the others, so that just the one mode is propagated. Of course, the behavior gets complicated and difficult to control.

24-8 Another way of looking at the guided waves

We want now to show you another way of understanding why a waveguide attenuates its signals rapidly for frequencies below the cutoff frequency ω_c . Then you will have a more "physical" idea of why the behavior changes so drastically between low and high frequencies. We can do this for the rectangular guide by visualizing the fields in terms of cylindrical harmonics in the walls of the guide. This approach only works for rectangular guides, however, that's why we started with the round, rather than flat, cylinders which works, in principle, for guides of any shape.

In the mode we have described, longitudinal charges are moving but no x -field, so we can ignore the top and bottom of the guide and imagine that the guide is extended indefinitely in the vertical direction. We may then set the guide just as easily in two vertical parts with no separation.

Let's say that the current in the field is a vertical wire placed in the middle of the guide, with the wire carrying a current that oscillates at the frequency ω . In the absence of the guide walls such a wave would produce cylindrical waves.

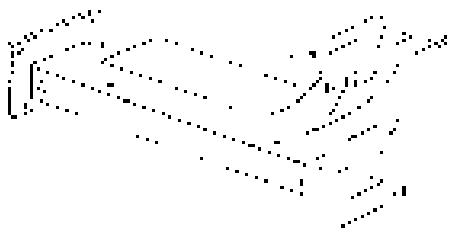


Fig. 24-12. A unidirectional coupler.

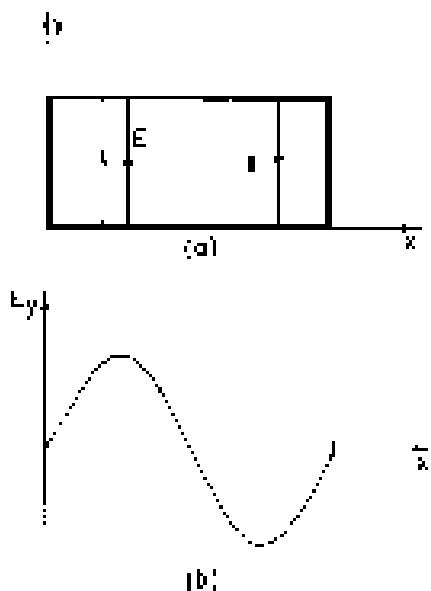


Fig. 24-14. Another possible wave for $\omega < \omega_c$.

Now we consider that the guide walls are perfect conductors. Then, just as in electrostatics, the assumptions of the source will be correct if we add to the field of the wire the field of one or more suitable image wires. The image idea works just as well for electrodynamics as it does for electrostatics, provided, of course, that we also include the retardations. We know that is true because we have often seen a mirror producing an image of a light source. And a mirror is just a "perfect" conductor for electric and magnetic waves with infinite frequencies.

Now let's take a horizontal cross section, as shown in Fig. 24-15, where W_1 and W_2 are the two guide walls and S_1 is the source wave. We call the direction of the current in the wire positive. Now if there were only one wall say W_1 , we could remove it if we placed an image source (with opposite polarity) at the position marked S_1' . But with both walls in place there will also be an image of S_1 in the wall W_2 , which we show as the image S_2 . This source, too, will have an image in W_1 , which we call S_3 . Now both S_1 and S_3 will have images in W_2 at the positions marked S_4 and S_5 , and so on. For our two plane conductors with the source halfway between, the fields are the same as those produced by an infinite line of sources, all separated by the distance a . (It is, in fact just what you would see if you looked at a wire placed half way between two parallel mirrors.) For the fields to be zero at the walls, the polarity of the currents in the images must alternate from one image to the next. In other words, they oscillate 180° out of phase. The waveguide field is, then, just the superposition of the fields of such an infinite set of line sources.

We know that if we are close to the sources, the field is very much like the static fields. We considered in Section 7-5 the static field of a grid of line sources and found that it is like the field of a charged plate except for terms that decrease exponentially with the distance from the grid. Here the average source strength is zero, because the sign alternates from one source to the next. Any field will start slowly but fall exponentially with distance. Close to the source, we see the field mainly of the nearest source; at large distances, many sources contribute and their average offset is zero. So now we see why the waveguide (below cutoff frequency) gives an exponentially decreasing field. At low frequencies, in particular, the static approximation is good, and it predicts a rapid attenuation of the fields with distance.

Now we are faced with the opposite question: Why are waves propagated at all? Just is the *in-phase* part! The reason is that at each k -multiple of the retardation of the fields contained in additional changes in phase will not change the fields of the out-of-phase sources to add instead of canceling. In fact, in Chapter 29 of Vol. I we have already studied just this problem: the fields generated by a row of sources in a uniform posting. There we found that when several sources are linearly arranged, they can give an interference pattern that has a strong signal in some direction but no signal in another.

Suppose we go back to Fig. 24-15 and look at the fields which arrive at a large distance from the array of image sources. The fields will be strong only in certain directions which depend on the frequency, only in those directions for which the fields from all the sources add in phase. At a reasonable distance from the sources the field propagates in these directions as sinusoidal waves. We have sketched such a wave in Fig. 24-16, where the solid lines represent the wave crests and the dashed lines represent the troughs. The wave direction will be the one for which the difference in the retardation for two neighboring sources to the crest of a wave corresponds to one-half a period of oscillation. In other words, the difference between ϵ_1 and ϵ_2 in the figure is one-half of the free space wavelength:

$$\epsilon_2 - \epsilon_1 = \frac{\lambda_0}{2}.$$

The angle θ is then given by

$$\sin \theta = \frac{\lambda_0}{2a}. \quad (24.33)$$

There is, of course, another set of waves traveling downward at the same phase angle with respect to the array of sources. The complete waveguide field (at 100%

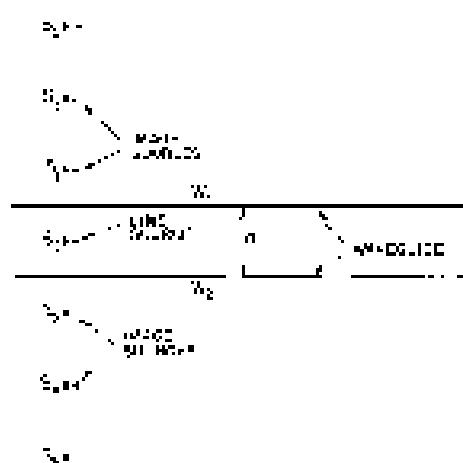


Fig. 24-15. The line source S_1 between the conducting plates W_1 and W_2 . The wave can be replaced by the infinite succession of image sources.



show to the source; & the superposition of these two sets of waves, as shown in Fig. 24-17. The optical fields are easily like this, if one goes only between the two walls of the waveguide.

At points like A and C, the fields of the two waves interfere, and the light will have an amplitude at points like B, both waves having peak negative voltage, and the light has its minimum (or zero) negative voltage. As time goes on, the light in the guide appears to be traveling along the guide with a wavelength λ_m , which is the distance from A to C. This distance is related to θ by

$$\cos \theta = \frac{\lambda_m}{\lambda_0} \quad (24.14)$$

Using Eq. (24.13) for θ , we get that

$$k_m = \frac{\lambda_0}{\cos \theta} = \frac{\lambda_0}{\sqrt{1 - (\lambda_0/\lambda_0)^2}} \quad (24.15)$$

which is just what we found in Eq. (24.10).

Now we see why there is only wave propagation above the "cut-off" frequency ω_c . If the free-space wavelength is longer than λ_0 , there is no angle where the waves shown in Fig. 24-16 can appear. The necessary constructive interference disappears suddenly when λ_0 drops below $2a$, or when ω rises above $\omega_c = \omega_0^2/a$.

If the frequency is high enough, there can be two or more possible directions in which the waves will appear. For our case, this will happen if $k_m < k_0$. In general, however, it could also happen when $k_m > k_0$. These additional waves correspond to the higher guide modes we have mentioned.

It has also been made evident, by our analysis why the phase velocity of the guided waves is greater than c and why this velocity depends on ω . As ω is changed, the angle of the two waves of Fig. 24-16 changes, and therefore so does the velocity along the guide.

Although we have described the guided wave as the superposition of the fields of an infinite array of line sources, you can see that we would arrive at the same result if we imagined two sets of free-space waves being continually reflected back and forth between two perfect mirrors—remembering that a reflection means a reversal of phase. These sets of reflecting waves would all cancel each other unless they were going in just the angle θ given in Eq. (24.13). There are many ways of looking at the same thing.

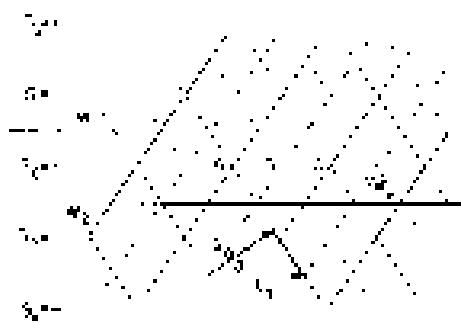


Fig. 24-17. The wave-guide field can be viewed as the superposition of two sets of plane waves.

Electrodynamics in Relativistic Notation

24-1 Four-vectors

We now show the application of the special theory of relativity to electrodynamics. As we have already studied the special theory of relativity in Chapters 15 through 17 of Vol. I, we will just review quickly the basics here.

It is found experimentally that the laws of physics are not coupled if we move with uniform velocity. You can tell if you are moving at speed v by moving with uniform velocity in a straight line, unless you know outside the spaceship, or at least make an observation having to do with the world outside. Any rule law of physics we write down must be arranged so that this fact of nature is built in.

The relationship between the space and time of two systems of coordinates, one S , in uniform motion in the direction with speed v relative to the other, S' , is given by the *Lorentz transformation*:

$$\begin{aligned} t' &= \frac{t - vx}{\sqrt{1 - v^2}}, & x' &= y, \\ x' &= \frac{x - vt}{\sqrt{1 - v^2}}, & z' &= z. \end{aligned} \quad (24.1)$$

The laws of physics must be such that after a Lorentz transformation, the new form of the laws looks just like the old form. This is just like the principle that the laws of physics can't depend on the *coordinates* of our coordinate system. In Chapter 14 of Vol. I, we saw that it was necessary to describe the mathematics of the invariance of physics with respect to rotations was to write our equations in terms of tensors.

For example, if we have two vectors

$$A = (A_x, A_y, A_z) \quad \text{and} \quad B = (B_x, B_y, B_z),$$

we found that the combination

$$A \cdot B = A_x B_x + A_y B_y + A_z B_z,$$

was not changed if we changed over to a rotating coordinate system. So we know that if we want a scalar product like $A \cdot B$ on both sides of an equation, we require it will have exactly the same form in all rotated coordinate systems. We also discussed an operator (see Chapter 2),

$$\nabla = \left(\frac{\delta}{\delta x}, \frac{\delta}{\delta y}, \frac{\delta}{\delta z} \right),$$

which, when applied to a scalar function, gave three quantities which transform just like a vector. With this operator we defined the gradient, and in combination with other vectors, the divergence and the Laplacian. Finally we discovered that by taking sums of certain products of parts of the components of two vectors, we could get three new quantities which behaved like new vectors. We called the cross product of two vectors. Using the cross product with our operator ∇ we then defined the curl of a vector.

Since we will be referring back to what we have done in vector analysis, we have put in Table 24.1 a summary of all the important vector operations in three dimensions that we have used in the past. The point is that it might be possible to write the equations of physics so that both sides transform the same way under

24-2 Four-vectors

24-2 The scalar product

24-3 The four-dimensional gradient

24-4 Electrodynamics in four-dimensional notation

24-5 The four-potential of a moving charge

24-6 The invariance of the equations of electrodynamics

In this chapter: $c = 1$

Ref. to: Chapter 15, Vol. I, *The Special Theory of Relativity*
Chapter 16, Vol. I, *Relativistic Energy and Momentum*

Chapter 17, Vol. I, *Space-Time*

Chapter 15, Vol. II, *Mechanics*

Table 25.1

The important quantities and operations of vector analysis in three dimensions

Definition of:

vector	$\mathbf{A} = (A_x, A_y, A_z)$
scalar product	$A \cdot B$
Differential vector operator:	∇
Gradient	∇V
Divergence	$\nabla \cdot \mathbf{A}$
Laplacian	$\nabla \cdot \nabla = \nabla^2$
Cross product	$\mathbf{A} \times \mathbf{B}$
Curl	$\nabla \times \mathbf{A}$

orientation. One side is a vector, the other side must also be a vector, and both sides will change length in exactly the same way if we rotate our coordinates system. But, finally, if a vector is a vector, the other side must also be a vector, so that both sides change when we rotate coordinates, and so on.

Now in the case of special relativity, time and space are intertwined, since, and we must do the analogous things for four dimensions. We would not expect this to happen just for rotations but also for other invariant frames. That means that the equations will be invariant under the Lorentz transformation, or, in equation (25.1). To give you a taste, this chapter is to show you how that can be done. Before we get started, however, we ought to do some relativity, and guess what? we work in four dimensions, not just three-dimensional space. And that is to choose a unit of length in time so that the speed of light is constant. You might think about us taking that unit of time to be the time that it takes light to go one meter (which is about 3×10^{17} sec). We can then call this time unit "the meter". Using this definition of our equation will allow it to be covariant, i.e., space-time covariant. Also, all the $\partial/\partial x^\mu$ will disappear. They are now absolute derivatives. (If this bothers you, you can always add the c back in by multiplying every derivative, in general, by c , though you know c is needed to make the dimensions of the equations of motion right.) Well, this program works, we are ready to begin. Our program is to do the first dimensionless "homogeneous" of the "Laplace operator" exercises for three dimensions. It is only quite a simple problem, we just work by analogy. The only real complication is the notation, since there is used up the vector symbol for three dimensions and one short list of signs.

First, by analogy with exercises in three dimensions, we define a four-vector to consist of the four quantities $(\rho_0, \rho_1, \rho_2, \rho_3)$, which "travel" like (x_0, x_1, x_2, x_3) when we change to a moving coordinate system. There are several different ways that people use for a four-vector; we will write ρ_μ , by which we mean a group of four numbers: $(\rho_0, \rho_1, \rho_2, \rho_3)$. In other words, the subscripts will take all the four "values" $0, 1, 2, 3$. It will also be convenient, at times, to include the three space components by a three vector, like this: $\rho_1 = (\rho_1, \rho_2, \rho_3)$.

We have already encountered one four-vector, which consists of the energy and momentum of a particle (Chapter 14, Sec. 10). In our new notation we write

$$\rho_1 = (E, \mathbf{p}) \quad (25.2)$$

which means that the "bottom" (ρ_0) is called E , or "energy". From the three components of the three-vector \mathbf{p} of a particle,

it looks as though the game is really very simple. But, and three-vector is physical and we have to consider what the remaining component should be, and we have a four-vector. So see that this is not the case, consider the velocity vector with components

$$v_\mu = \frac{d\rho_\mu}{dt} = \rho_1 \cdot \frac{dx^\mu}{dt} = \rho_1 \cdot \frac{ds}{dt},$$

The question is: What is the time component? (Insure that give the right answer. If the components are like v_x, v_y, v_z , we would guess that the time component is

$$v_0 = \frac{ds}{dt} = 1.$$

This is wrong. The reason is that the s in each denominator is not an invariant when we make a Lorentz transformation. The numerators have the right behavior to make a four-vector, but the s in the denominators, which is the distance in 4-dimensions, and is not the same in two different systems.

To figure out that the four "velocity" components, which we have written down, will become the components of a four-vector, if we just divide by $\sqrt{1 - v^2/c^2}$. We can see that that is true because, if we start with the momentum four-vector

$$\rho_1 = (E, \mathbf{p}) = \left(\frac{mc}{\sqrt{1 - v^2/c^2}}, v^0, \frac{mv^1}{\sqrt{1 - v^2/c^2}}, v^2, \frac{mv^3}{\sqrt{1 - v^2/c^2}} \right), \quad (25.3)$$

the ratio is by the Lorentz transform which is an invariant scalar in four dimensions, we have

$$\frac{p_2}{m} = \left(\sqrt{1 - \frac{v^2}{c^2}}, \frac{v}{\sqrt{1 - v^2}} \right). \quad (25.4)$$

which must still be a four-vector. (Dividing by an invariant scalar doesn't change the transformation properties.) So we can define the "velocity four-vector" \mathbf{v}_v by

$$v_v = \frac{1}{\sqrt{1 - v^2}}, \quad v_v^\mu = \frac{v^\mu}{\sqrt{1 - v^2}}, \quad (25.5)$$

$$v_v^\mu = \frac{v^\mu}{\sqrt{1 - v^2}}, \quad v_v^\mu = \frac{v^\mu}{\sqrt{1 - v^2}}.$$

This four-vector is a useful quantity; we can, for instance, write

$$p_\mu = m v_\nu g^{\mu\nu}. \quad (25.6)$$

This is the typical sort of relation in an equation which is relativistically covariant, just like v is a four-vector. (The right-hand side is an invariant times a four-vector, which is itself a four-vector.)

25-2 The scalar product

It is an accident of fate, if you wish, that under coordinate rotations the distance of a point from the origin does not change. This means mathematically that $x^2 + y^2 + z^2$ is an invariant. In other words, after a rotation $x'^2 + y'^2 + z'^2 = x^2 + y^2 + z^2$

$$x'^2 + y'^2 + z'^2 = x^2 + y^2 + z^2.$$

Now the question is: Is there a similar quantity which is invariant under the Lorentz transformation? There is. From Eq. (25.4) you can see that

$$x'^2 + y'^2 + z'^2 = v^2. \quad (25.7)$$

That is pretty nice except that it's not a point. For choice of the reference. We can fix that up by multiplying x^2 and v^2 . The Lorentz transformation or else a rotation will leave the quantity unchanged. So the quantity which is analogous to $x^2 + y^2 + z^2$ for four dimensions, in three dimensions is

$$v^2 = x^2 + v^2 - z^2.$$

It is an invariant under rot so it's called the "complete Lorentz group", which means the transformation of both translations at constant velocity and rotations.

Now since this invariance is an algebraic matter depending only on the transformation rule of Eq. (25.4) (no rotations), it is true for any four vector (by definition it is a 4D transform the same). So for a four-vector a , we have that

$$a_1^2 + a_2^2 + a_3^2 + a_4^2 = a_1'^2 + a_2'^2 + a_3'^2 + a_4'^2.$$

We will call this quantity the square of "the length" of the four vector a . (Some other people change the sign of all the terms and call the length $a_1^2 + a_2^2 + a_3^2 - a_4^2$, so you'll have to watch out.)

Now if we have two vectors a and b , their corresponding components transform in the same way, so the combination

$$a_\mu b_\nu = a_1 a_1 + a_2 a_2 + a_3 a_3 + a_4 a_4$$

is also an invariant, scalar quantity. (We have in fact already proved this in Chapter 17 of Vol. I.) Clearly this expression is quite analogous to the dot product for vectors. We will, in fact, call it the dot product or scalar product of two four vectors. It would seem logical to write this as $a_\mu b^\mu$, but that would look like a dot product. Unfortunately, it's not done that way; it's usually written without the μ .

So we will follow the convention and write the dot product simply as $a_i b_i$. So, by definition,

$$a_i b_i = a_1 b_1 + a_2 b_2 + a_3 b_3 + a_4 b_4 \quad (25.7)$$

Whenever you see two identical subscripts (not i,j) you will occasionally have to use α, β, γ or some other letter instead of j). It means that you are to take the four products and sum, remembering the same sign for the products of the space components. With this convention the transpose of the scalar product under a Lorentz transformation would be written as

$$a'_i b'_j = a_i b_j.$$

Since the last three terms in (25.7) are just the scalar dot product in three dimensions, it is often more convenient to write

$$a_i b_i = a_1 b_1 + a_2 b_2$$

It is also obvious that the four-dimensional length we described above can be written as $a_i a_i$,

$$a_i a_i = a_1^2 + a_2^2 + a_3^2 + a_4^2 = a^2 = \sigma^2 \quad (25.8)$$

It will also be convenient to sometimes write this quantity as a_i^2

$$a_i^2 = a_i a_i.$$

We will now give you an illustration of the usefulness of four-vector dot products. Antiprotons (\bar{p}) are produced in large accelerators by the reaction



That is, an energetic proton collides with a proton at rest (for example, in a hydrogen target placed in the beam), and if the incident atom has enough energy, a proton-antiproton pair may be produced, in addition to the two original protons.* The question is: How much energy must be given to the incident proton to make this reaction energetically possible?

The easiest way to get the answer is to consider what the reaction looks like in the center-of-mass (CM) system (see Fig. 25-1). We'll call the incident proton a and its laboratory momentum p_a . Similarly, we'll call the target proton b and its lab momentum

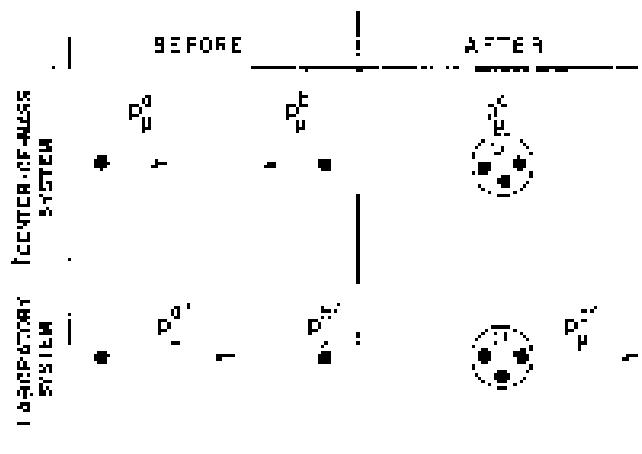


Fig. 25-1. The reaction $p + p \rightarrow p + p + \bar{p} + \bar{p}$ viewed in the laboratory and CM systems. The incident proton is supposed to have just barely enough energy to make the reaction go. Protons are denoted by solid circles; antiprotons, by open circles.

* You may well ask: Why not consider the reaction

$$p + p \rightarrow p + p + \bar{p}$$

or even

$$p + p \rightarrow p + p$$

which clearly looks simpler? The answer is that a process like collision conservation of baryon tells us the quantity "one less proton minus number of antiprotons" cannot change. This quantity is 2 on the left side of our reaction. Therefore, if we want an antiproton on the right side, we must "cancel the two protons" (as in the last eq.).

momentum p_1^4). If the nuclear system has *precisely* enough energy to make the reaction go, the final state (the situation after the collision) will consist of a glob containing three protons and an antiproton at rest in the CM system. If the incident energy were slightly higher, the final state particles would have some kinetic energy and be moving apart; if the incident energy were slightly lower, there would not be enough energy to make the four particles.

If we call ρ^4 the total four-momentum of the whole glob in the final state, conservation of energy and momentum tells us that

$$\rho^0 = p_1^0 + p_2^0 + p_3^0 + p_4^0$$

and

$$E^0 = p_1^0 + p_2^0 + p_3^0 + p_4^0$$

Combining these two equations, we can write that

$$p_1^0 + p_2^0 + p_3^0 = p_4^0. \quad (25.9)$$

Now the important thing is that this is an equation among four-vectors, and is therefore true in any inertial frame. We can use this fact to simplify our calculations. We start by taking the "length" of each side of Eq. (25.9); they are, of course, also equal. We get

$$(p_1^0 + p_2^0 + p_3^0)^2 = p_4^0 p_4^0 \quad (25.10)$$

Since $p_4^0 p_4^0$ is invariant, we can evaluate it in *any* coordinate system. In the CM system, the time component of p_4^0 is the rest mass of four protons, namely $4M$, and the space part \vec{p} is zero; so $p_4^0 = (4M, 0)$. We have used the fact that the rest mass of an antiproton equals the rest mass of a proton, and we have called this combined mass M .

Thus, Eq. (25.10) becomes

$$p_1^0 p_1^0 + p_2^0 p_2^0 + p_3^0 p_3^0 = 16M^2. \quad (25.11)$$

Now $p_1^0 p_1^0$ and $p_2^0 p_2^0$ are very easy to evaluate: the "length" of the momentum four-vector of any particle is just the mass of the particle squared:

$$p_1^0 p_1^0 = E^2 - M^2.$$

This can be shown by direct calculation or, more elegantly, by noting that for a particle at rest $p_1^0 = (M, 0)$, so $p_1^0 p_1^0 = M^2$, but since E is an invariant, it is equal to M^2 in any frame. Using these results in Eq. (25.11), we have

$$2p_1^0 p_1^0 = 14M^2$$

or

$$p_1^0 p_1^0 = 7M^2. \quad (25.12)$$

Now we can also evaluate $p_1^0 p_1^0$ in the laboratory system. The four-vector p_1^0 can be written (E^0, \vec{p}^0) , while $p_1^0 = (M, 0)$, since it describes a proton at rest. Thus, $p_1^0 p_1^0$ must also be equal to M^2 , and since we know the scalar product is an invariant this must be numerically the same as what we found in (25.12). So we have that

$$E^2 = 7M^2,$$

which is the result we were after. The *total* energy of the initial proton must be at least $7M$ (about 6 GeV since $M = 938$ MeV) in order that the rest mass M , the kinetic energy must be at least $6M$ (about ± 6 GeV). The Bevatron accelerator at Berkeley was designed to give about ± 2 GeV of kinetic energy to the protons it accelerates, in order to be able to make antiprotons.

Since scalar products are invariant, they are always interesting to evaluate. What about the "length" of the four-velocity (v, \vec{v}) ?

$$v \cdot v = v^2 = v^0 v^0 + \frac{1}{c^2} \vec{v} \cdot \vec{v} = \frac{c^2}{1 - v^2/c^2} - 1.$$

Thus, v_c is the *light speed factor*.

25.3. The four-dimensional gradient

The next thing that we have to discuss is the four-dimensional analog of the gradient. We recall (Chapter 14) that in three dimensions there are three operators $\partial/\partial x^i$, $\partial/\partial y^i$, $\partial/\partial z^i$ that can be called the three partial derivatives or the gradient. The same scheme ought to work in four dimensions; that is, we might guess that the four-dimensional gradient should be $(\partial/\partial t)$, $(\partial/\partial x)$, $(\partial/\partial y)$, $(\partial/\partial z)$. That is wrong.

To see the error, consider a scalar function ϕ which depends only on x and t , the storage of ϕ . If we make a small change Δx in x while holding t constant, it is

$$\Delta\phi = \frac{\partial\phi}{\partial x} \Delta x \quad (25.13)$$

On the other hand, according to a moving observer,

$$\Delta\phi = \frac{\partial\phi}{\partial x'} \Delta x' = \frac{\partial\phi}{\partial t'} \Delta t'.$$

We can express $\Delta x'$ and $\Delta t'$ in terms of Δx by using Eq. (25.1). Remembering that we are holding t constant, so that $\Delta v = 0$, we write

$$\Delta x' = -\frac{v}{\sqrt{1-v^2}} \Delta x \quad \Delta t' = \frac{\Delta t}{\sqrt{1-v^2}}$$

Thus,

$$\begin{aligned} \frac{\partial\phi}{\partial x'} &= \frac{\partial\phi}{\partial x} \left(-\frac{v}{\sqrt{1-v^2}} \frac{\Delta t}{\Delta x} \right) = \frac{\partial\phi}{\partial t'} \left(\frac{\Delta t}{\sqrt{1-v^2}} \right) \\ &= \left(\frac{\partial\phi}{\partial t'} + v \frac{\partial\phi}{\partial x'} \right) \frac{\Delta t}{\sqrt{1-v^2}} \end{aligned}$$

Comparing this result with Eq. (25.13), we learn that

$$\frac{\partial\phi}{\partial t'} = \frac{1}{\sqrt{1-v^2}} \left(\frac{\partial\phi}{\partial t} + v \frac{\partial\phi}{\partial x} \right) \quad (25.14)$$

A similar calculation gives

$$\frac{\partial\phi}{\partial x'} = -\frac{v}{\sqrt{1-v^2}} \left(\frac{\partial\phi}{\partial x} + v \frac{\partial\phi}{\partial t} \right). \quad (25.15)$$

Now we can see that the gradient is rather strange. The formulas for x and t in terms of x' and t' [obtained by solving Eq. (25.2)] are:

$$t = \frac{t' + vx'}{\sqrt{1-v^2}}, \quad x = \frac{x' + vt'}{\sqrt{1-v^2}}.$$

This is the way a four-vector must transform. But Eqs. (25.14) and (25.15) have a couple of signs wrong!

The answer is that instead of the operator $\partial/\partial t$ (∇), we must define the four-dimensional gradient operator, which we will call ∇_4 , by

$$\nabla_4 = \left(\frac{\partial}{\partial t}, -\nabla \right) = \left(\frac{\partial}{\partial t}, -\frac{\partial}{\partial x}, -\frac{\partial}{\partial y}, -\frac{\partial}{\partial z} \right) \quad (25.16)$$

With this definition, the sign difficulties mentioned above go away, and ∇_4 behaves like a four-vector should. (It's rather awkward to have these minus signs, but that's the way the world is.) Of course, what I mean to say is that ∇_4 behaves like a four-vector because the four-gradient of a scalar is a four-vector. If α is a time-constant invariant field (i.e., α is invariant); then $\nabla_4 \alpha$ is a four-vector field.

All right, now that we have vectors, gradients, and dot products, the next thing is to look for an invariant which is analogous to the divergence of three-dimensional vector analysis. Clearly, the analog is to form the expression $\nabla_4 \cdot b$, where b is a four-vector field whose components are functions of space and time.

We define the divergence of the four-vector $b_\mu = (\nu, \mathbf{b})$ as the dot product of ∇_μ and b_μ :

$$\begin{aligned}\nabla_\mu b_\mu &= \frac{\partial}{\partial t} b_0 + \left(-\frac{\partial}{\partial x}\right) b_x + \left(-\frac{\partial}{\partial y}\right) b_y + \left(-\frac{\partial}{\partial z}\right) b_z \\ &= \frac{\partial}{\partial t} b_0 - \nabla' \cdot \mathbf{b},\end{aligned}\quad (25.17)$$

where $\nabla' \cdot \mathbf{b}$ is the value of three-dimensional divergence of the three-vector \mathbf{b} . Note that one has to be careful with the signs. Some of the minus signs come from the definition of the scalar product, Eq. (15.4); the others are required because the space components of ∇_μ are $(\partial/\partial t, -\partial/\partial x, -\partial/\partial y, -\partial/\partial z)$, as in Eq. (25.9). The divergence as defined by (25.17) is not the same as the three-dimensional divergence system ∇' after a Lorentz transformation.

Let's look at a physical example in which the four-divergence shows up. We can start to solve the problem of the fields around a moving wire. We have already seen (Section 13.7) that the electric charge density ρ and the current density j form a four vector $j_\mu = (\rho, \mathbf{j})$. If an uncharged wire carries the current \mathbf{j} , then in a frame moving parallel with velocity v along x_0 , the wire will have the charge and current density obtained from the Lorentz transformation, Eqs. (25.1), as follows:

$$\rho' = \frac{\rho}{\sqrt{1 - v^2}}, \quad j'_\mu = \frac{j_\mu}{\sqrt{1 - v^2}}.$$

These are just what we found in Chapter 12. We can now use the Lorentz-Maxwell's equation in the moving system to find the fields.

The charge conservation law, Section 13.7, also takes on a simple form in the relativistic situation. Consider first the divergence of j_μ :

$$\nabla_\mu j_\mu = \frac{\partial \rho}{\partial t} + \nabla' \cdot \mathbf{j}. \quad (25.18)$$

The law of the conservation of charge says that the outflow of charge per unit volume τ ms. equals the negative rate of increase of charge density. In other words, that

$$\nabla' \cdot \mathbf{j} = -\frac{d\rho}{dt}.$$

Putting this in to Eq. (25.18), the law of conservation of charge takes on the simple form:

$$\nabla_\mu j_\mu = 0. \quad (25.19)$$

Since $\nabla_\mu j_\mu$ is an invariant scalar, it is zero in one frame if it is zero in all frames. We have the result that if charge is conserved in one coordinate system, it is conserved in all coordinate systems moving with uniform velocity.

As our last example we want to consider the scalar product of the gradient operator ∇_μ with itself. In three dimensions, such a product gives the Laplacian

$$\nabla^2 = \nabla \cdot \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

What do we get in four dimensions? That's new. Following on rules for dot products and gradients, we get

$$\begin{aligned}\nabla_\mu \nabla_\mu &= \frac{\partial}{\partial t} \frac{\partial}{\partial t} + \left(-\frac{\partial}{\partial x}\right) \left(-\frac{\partial}{\partial x}\right) + \left(-\frac{\partial}{\partial y}\right) \left(-\frac{\partial}{\partial y}\right) + \left(-\frac{\partial}{\partial z}\right) \left(-\frac{\partial}{\partial z}\right) \\ &= \frac{\partial^2}{\partial t^2} - \nabla'^2.\end{aligned}$$

This quantity, which is the summing of the three-dimensional Laplacian, is called

the D'Alembertian and has a special notation:

$$\square = \nabla_\mu \nabla^\mu - \frac{\partial^2}{\partial t^2} \quad (25.20)$$

From its definition it is an invariant scalar operator; it operates on a four-vector field, it produces a new four-vector field. (Some people define the D'Alembertian with the opposite sign to Eq. (25.20), so you will have to be careful when reading the literature.)

We have now found four-dimensional equivalents of most of the three-dimensional quantities we also listed in Table 25-1. (We do not yet have the equivalents of the cross product and the curl operator, we won't get to them until the next chapter.) It may help you remember how they go if we put all the important definitions and results from 25-1 in one place, so we have made such a summary in Table 25-2.

Table 25-2

The important quantities of vector analysis in three and four dimensions.

	Three dimensions	Four dimensions
Vectors	$\mathbf{A} = (A_x, A_y, A_z)$	$\mathbf{a}_\mu = (a_0, a_1, a_2, a_3) = (a_0, \mathbf{a})$
Scalar product	$A \cdot B = A_x B_x + A_y B_y + A_z B_z$	$a_0 b_0 = a_0 b_0 - a_1 b_1 - a_2 b_2 - a_3 b_3 = a_0 b_0 - \mathbf{a} \cdot \mathbf{b}$
Vector operator	$\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$	$\nabla_\mu = (\partial/\partial x^0, \partial/\partial x^1 - \partial/\partial y^1, \partial/\partial x^2 - \partial/\partial z^2, \partial/\partial x^3 - \nabla)$
Gradient	$\nabla \psi = \left(\frac{\partial \psi}{\partial x}, \frac{\partial \psi}{\partial y}, \frac{\partial \psi}{\partial z} \right)$	$\nabla_\mu \psi = \left(\frac{\partial \psi}{\partial x^0} + \frac{\partial \psi}{\partial x^1} - \frac{\partial \psi}{\partial y^1} - \frac{\partial \psi}{\partial z^1}, \frac{\partial \psi}{\partial x^2} + \frac{\partial \psi}{\partial x^3} \right) = \left(\frac{\partial \psi}{\partial x} + \nabla_\perp \right)$
Divergence	$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$	$\nabla_\mu \cdot \mathbf{a}_\mu = \frac{\partial a_0}{\partial x^0} + \frac{\partial a_1}{\partial x^1} + \frac{\partial a_2}{\partial x^2} + \frac{\partial a_3}{\partial x^3} = \frac{\partial a_0}{\partial x^0} - \nabla \cdot \mathbf{a}$
Laplacian and D'Alembertian	$\nabla \cdot \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$	$\nabla_\mu \nabla_\mu = \frac{\partial^2}{\partial x^{02}} + \frac{\partial^2}{\partial x^{12}} - \frac{\partial^2}{\partial y^{12}} - \frac{\partial^2}{\partial z^{12}} + \frac{\partial^2}{\partial x^{22}} + \frac{\partial^2}{\partial x^{32}} - \nabla^2 = \square$

25-4 Electrodynamics in four-dimensional notation

We have already considered the D'Alembertian equation, without going to that much, in Section 18-6. The differential equations we found there for the potentials can be written in the new notations as:

$$\square^2 \psi = \frac{q}{\epsilon_0 c} \quad \square^2 M = \frac{j}{c} \quad (25.21)$$

The four quantities on the right-hand side of the two equations in (25.21) are proportional to each other by q/c , which is a universal constant which will be the same in all coordinate systems if the same unit of charge is used in all frames. So the four quantities $q/\epsilon_0 c$, j/c , ψ , M , also transform as a four-vector. We can write them as \mathbf{A}_μ . The D'Alembertian doesn't change when the coordinate system is changed, so the quantities ψ , A_0 , A_1 , A_2 and A_3 also transform like a four-vector, which means that they are the components of a four-vector. In short

$$\mathbf{A}_\mu = (\phi, \mathbf{A})$$

is a four-vector. Why, we call the scalar and vector potentials are really different aspects of the same physical thing. They belong together. And if they are kept together the elegant structure of the world is obvious. We call \mathbf{A}_μ the four-potential.

In the four-vector notation Eqs. (25.21) become simply

$$\nabla^{\mu} A_{\mu} = \frac{\partial \phi}{\partial t}. \quad (25.22)$$

The physics of this equation is just the same as Maxwell's equations. But there is some pleasure in being able to write them in an elegant form. The pretty form is also meaningful, i.e., shows directly the invariance of electrodynamics under the Lorentz transformation.

Remember that Eqs. (25.21) could be deduced from Maxwell's equations only if we imposed the gauge condition

$$\frac{\partial \phi}{\partial t} = \nabla \cdot \mathbf{A} = 0, \quad (25.23)$$

which just says $\partial_{\mu} A_{\mu} = 0$, the gauge condition says that the divergence of the four-vector A_{μ} is zero. This condition is called the *Lorentz condition*. It is very convenient because it is an invariant condition and therefore Maxwell's equations stay in the form of Eq. (25.23) for all frames.

25-8 The four-potential of a moving charge

Although it is implicit in what we have already said, let us write down the transformation laws which give ϕ and \mathbf{A} in a moving system in terms of ϕ and \mathbf{A} in a stationary system. Since $A_0 = -q\phi/dt$ is a four-vector, the equation must look just like Eqs. (25.21), except that ϕ is replaced by ϕ' and a_0 is replaced by A_0 plus

$$\begin{aligned} \phi' &= \frac{\phi + v A_0}{\sqrt{1 - v^2}}, & A_0 &= A_{00}, \\ A'_0 &= \frac{A_0 - v \phi'}{\sqrt{1 - v^2}}, & A'_0 &= A_0. \end{aligned} \quad (25.24)$$

This assures that the primed coordinate system is moving with speed v in the positive x -direction, as measured in the unprimed coordinate system.

We will consider one example of the usefulness of the idea of the four-potential. What are the vector and scalar potentials of a charge q moving with speed v along the x -axis? The problem is easy in a coordinate system keeping with the charge, since in this system the charge is standing still! Let's say that the charge is at the origin of the S' -frame, as shown in Fig. 25.7. The vector potential in the moving system is then given by

$$A' = \frac{q}{2\pi r v}, \quad (25.25)$$

r being the distance from q to the field point as measured in the moving system. The scalar potential ϕ' is, of course, zero.

Now it is straightforward to find ϕ and A_0 , the potentials as measured in the stationary coordinate system. The answer (due to Eqs. (25.24)) is

$$\begin{aligned} \phi &= \frac{q'}{\sqrt{1 - v^2}}, & A_0 &= A_{00}, \\ A_0 &= \frac{A'_0 + v \phi'}{\sqrt{1 - v^2}}, & A'_0 &= A_0. \end{aligned} \quad (25.26)$$

Using the ϕ' given by Eq. (25.25), and $A'_0 = 0$, we get

$$\begin{aligned} \phi &= \frac{q}{4\pi c} \frac{1}{\sqrt{1 - v^2}} \\ &= \frac{q}{4\pi c} \frac{1}{\sqrt{1 - \beta^2}} \sqrt{\gamma^2 + \gamma^2 \beta^2 + \gamma^2 \beta^2} \end{aligned}$$



Fig. 25.7. The frame S' moves with velocity v (in the x -direction) with respect to S . A charge at rest at the origin of S' is at $x = vt$ in S . The potentials at P can be computed in either frame.

This gives us the scalar potential ϕ we would see in S, but, unfortunately, expressed in terms of the S' coordinates. We can get things in terms of r, θ, ϕ, z by substituting for x', y', z' , and t' using (25.1). We get

$$\phi = \frac{q}{4\pi c} \frac{1}{\sqrt{1 - v^2}} \frac{1}{\sqrt{(1 - v^2) \sin^2 \theta + v^2 + 1 - v^2}}. \quad (25.27)$$

Following the same procedure for the components of A , you can show that

$$A = v\phi. \quad (25.28)$$

These are the same formulae we derived in a different method in Chapter 21.

25.4. The invariance of the equations of electromagnetism

We have found that the potentials ϕ and A taken together form a four-vector which we call A_μ , and that the wave equations—the full equations which determine the A_μ in terms of the v 's—can be written as in Eq. (25.22). This equation, together with the conservation law (Eq. (25.19)), gives us the **fundamental law** of the electromagnetic field:

$$\nabla_\mu A_\nu - \frac{1}{c} L_{\mu\nu} = V_{\mu\nu} = 0. \quad (25.29)$$

There, in one tiny space on the page, are all of the Maxwell equations—beautiful and simple. Did we learn anything from writing the equations this way, besides that they are beautiful and simple? In the first place, is it anything different from what we had before when we wrote everything out in all the various components? Can we from this equation deduce something that could not be deduced from the seven equations for the potentials in terms of the charges and currents? The answer is definitely no. The only thing we have been doing is changing the name of things.

Using a new notation, we have written a square symbol to represent the derivatives, but it still means nothing more nor less than the second derivative with respect to t , minus the second derivative with respect to x , minus the second derivative with respect to y , minus the second derivative with respect to z . And the difference that we have four equations, one each for $\omega = t, x, y$, or z . What does this say about the fact that the equations can be written in this simple form? From the point of view of calculating anything directly, it doesn't mean anything. Perhaps, though, the simplicity of the equations means that nature also has a certain simplicity.

Let us show you something interesting that we have recently discovered: all of the laws of physics can be contained in one equation. That equation is

$$U = 0. \quad (25.30)$$

With a simple example. Of course, it is necessary to know what the symbol means. U is a physical quantity which we will call the "universality" of a situation. And we have a formula for it. Here is how you calculate the universality. You take all of the known physical laws and write them in a special form. For example, suppose you take the law of mechanics, $F = ma$, and rewrite it as $F = m\ddot{x} = 0$. Then you can call $(F = m\ddot{x})$ —which should, of course, be zero—the "universality" of mechanics. Next, you take the laws of thermodynamics and call it U_1 , which can be called the "universality of nonelementary effects." In other words, you take

$$U_1 = (F - m\ddot{x})^2. \quad (25.31)$$

Now you write another physical law, say, $\nabla \cdot H = \rho/c_0$, and define

$$U_2 = \left(\nabla \cdot H - \frac{\rho}{c_0} \right)^2,$$

which you might call "the universal nonexistence of electricity." You continue to write U_3 , U_4 , and so on—and for every physical law there is

Finally you add the individual fluxes \mathbf{U} of the world to sum of the various unworld fluxes \mathbf{U}_i , from all the i phenomena that are available, that is, $\mathbf{U} = \sum \mathbf{U}_i$. Then the great "law of calculus" is

$$\boxed{\mathbf{U} = 0} \quad (25.2)$$

This "law" means, of course, that the sum of the squares of all the individual mismatchings is zero, and the only way the sum of a set of squares can be zero is for each one of the terms to be zero.

So the "beautifully simple" law in Eq. (25.2) is equivalent to the combination of equations that you originally wrote down. It is therefore what many believe that a simple notation that just hides the complexity in the definitions of symbols is more satisfying. At your risk, the beauty that appears in Eq. (25.2) is just from the fact that several equations are hidden within it, is no more than a trick. When you unwrap the whole thing, you get back where you were before.

However, there is more to the simplicity of the laws of electromagnetism written in the form of Eq. (25.2). It means more, just as a theory of vector analysis means more. The fact that the electromagnetic equations can be written in a very particular notation which is convenient for the four-dimensional geometry of the Lorentz transformations—in other words, as a vector equation in the four-space—means that it is invariant under the Lorentz transformations. This is because the Maxwell equations are invariant under those transformations, but they can be written in a beautiful form.

It is not obvious that the equations of electrodynamics can be written in the beautiful form of Eq. (25.2). The theory of relativity was developed however at the same time experimentally that the phenomena predicted by Maxwell's equations were the same in all inertial systems. And it was precisely by studying the transformation properties of Maxwell's equations that Lorentz discovered his transformation as the one which left the equations invariant.

There is, however, another reason for writing our equations this way. It has been discovered, after Einstein guessed that it might be so—that all of the laws of physics are invariant under the Lorentz transformation. That is the principle of relativity. Therefore, if we invent a notation which shows immediately when a law is written down whether it is invariant or not, we can be sure that in trying to make new theories we will write only equations which are consistent with the principle of relativity.

The fact that the Maxwell equations are simple in this particular notation is not a miracle, because the notation was invented with them in mind. But the interesting physical thing is that a certain of physics—the propagation of meson waves or the behavior of neutrinos in beta decay, are as Earth—centric here, as some invariance under the same transformation. Even when you are moving at a uniform velocity in a spaceship, all of the laws of nature remain together in such a way that no new phenomena will show up. It is because the principle of relativity is a fact of nature that in the notation of four-dimensional vectors the equations of the world will look simple.

Lorentz Transformations of the Fields

26-1 The four-potential of a moving charge

We saw in the last chapter that the potential $A_\mu = (\phi, \mathbf{A})$ is to be invariant. The time component is the scalar potential ϕ , and the three spatial components \mathbf{A} the vector potential. We also worked out the potentials of a particle moving with uniform speed on a straight line by using the Lorentz transformation. (We had already found them by another method in Chapter 21.) For a point charge whose position at the time t is (x, y, z) , the potentials at the point (x', y', z') are

$$\begin{aligned} \phi' &= -\frac{\eta}{4\pi\epsilon_0 c^2 \sqrt{1-v^2}} \left[\frac{(x'-vt)^2}{1-v^2} + y'^2 + z'^2 \right]^{1/2} \\ A_x' &= -\frac{v}{2\pi\epsilon_0 c \sqrt{1-v^2}} \left[\frac{y'^2}{1-v^2} + \frac{z'^2}{1-v^2} \right]^{1/2} \quad (26.1) \\ A_y' &= A_z' = 0 \end{aligned}$$

Equations (26.1) give the potentials at x', y', z' at the time t' for a charge whose "present" position (by which we mean the position at the time t) is $x = vt$. Notice that the equations are in terms of $(x = vt), y$, and z , which are the coordinates measured from the current position P' of the moving charge (see Fig. 26-1). The only influence we know really travels at the speed c , or v , is the behavior of the charge back at the retarded position P'' that really counts. The point P'' is at $t = t''$ (where $t'' = t - r/c$ is the retarded time). But, we said that the charge was moving with uniform velocity in a straight line, so naturally the behavior at P'' at the time of position are directly related. In fact, if we make the added assumption that the potentials depend only upon the position and the velocity at the current moment, we can do in equations (26.1) a complete reversal. For the point (x', y', z') a charge moving very fast, it would be very difficult to imagine, in some arbitrary fashion, say $t = t'$ the temporary in Eq. 26.1, and you are trying to find the potentials at the point (x', y', z') . First, you find the retarded position P'' and the velocity v at that point. Then you realize that the charge would keep on moving with the velocity cutting the easy time $t'' = t'$, so that it would then appear at an temporary position P_{temp} , which we call the "projected position," and would move there with the velocity v . (Of course, it doesn't do that; its real position at this t'' is P'' .) Then the potentials at (x', y', z') are just what equations (26.1) would give for the imaginary charge at the projector position P_{temp} . What we are saying is that since the potentials depend only on what the charge is doing at the retarded time, the potentials will be the same whether the charge continued moving at a constant velocity or whether it changed its velocity after t'' —that is, after the potentials that were going to appear at (x', y', z') at the time t'' were already determined.

You know, of course, that the moment that we have the formula for the potentials from a charge moving in any manner whatsoever, we have the complete electrodynamics: we can get the potentials of any charge distribution by super-

[†] The prime, and leading to the subscript μ , means that this does not refer to the primes referring to a Lorentz-transformed frame in the preceding chapter.

26-2 The four-potential of a moving charge

26-3 The fields of a point charge with a constant velocity

26-4 Relativistic transformation of the fields

26-5 The equations of motion in relativistic mechanics

In this chapter: $c = 1$

Review: Chapter 20, Vol. II, Solution of Maxwell's Equations in Free Space

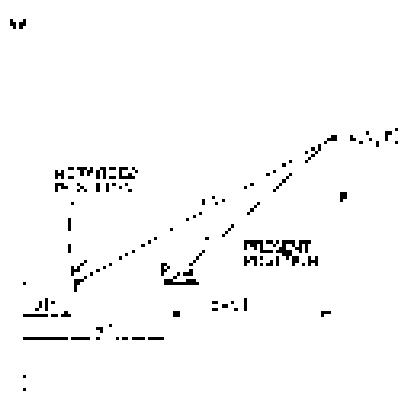


Fig. 26-1. Field of the fields of a charge moving along the x -axis with the constant speed v . The field "now" at the point (x, y, z) can be expressed in terms of the "present" position P , as well as in terms of P' , the "retarded position" ($t' = t - r/c$).

position. Therefore we can summarize all the phenomena of electrodynamics either by writing Maxwell's equations or by the following series of remarks. (These, like them, have you ever been even on a desert island? From them, all can be extrapolated.) As well, of course, knowing the Lorentz transformation; you will never forget where to look if you get lost in a system elsewhere.)

First, ϕ_0 is a function. Again, the Coulomb potential for a stationary charge is such that ϕ_0 , the potential produced by a charge moving in any way depends only upon the velocity and position at the retarded time. With these things being so, we have $\phi_0 = \phi_0(\mathbf{r}, t)$. From this and that \mathbf{A}_0 is a four-vector, we transform the Coulomb potential, which we know, and get the potentials for a constant velocity. Then, by the first statement that potentials depend only upon the past velocity at the retarded time, we can use the projected motion again to find the actual motion systematically, as I was doing earlier. But it is interesting to show that the laws of physics can be given stationary in different ways.

It is sometimes said, by people who are careless, that all of electrodynamics can be reduced solely from the Lorentz invariance and Coulomb's law. Of course, that is completely false. First, we have to suppose that there is a scalar potential ϕ_0 and a vector potential, just kept for book-keeping sake. That tells us how the potentials transform. Then why is it that the effects at the retarded time are the ones "long ago"? Well, just as it is that the potential's depend only on the position and the velocity won't, for instance, on the acceleration? To justify A and B do depend on the acceleration. If you try to make the calculation of a further win respect to them, you would say that they depend only upon the position and velocity at the retarded time. But then the fields from an accelerated charge would be the same as the fields from a charge at the projected position—which is false. The fields depend not only on the position and the velocity along the path, but also on the acceleration. So there are several additional tacit assumptions in this first statement that everything can be deduced from the Lorentz invariance. (Whence you see a sweeping statement that a tremendous amount of work can be done with very small number of assumptions, your always had the type in mind. It can be equally a large amount of impeded assumptions that, as far as I can see, you think about them sufficiently carefully.)

26.2 The fields of a point charge with a constant velocity

Now that we have the potentials from a point charge moving at constant velocity, we ought to find the fields—for practical reasons. There are many cases where we have uniformly moving particles—for instance, cosmic rays going through a cloud chamber, or even slow-moving electrons in a wire. So let's at least see what the fields actually do look like for any speed—even the speeds nearly that of light—assuming only that there is no acceleration. It is an uninteresting question.

We get the fields from the potentials by the usual rules:

$$\mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A}.$$

First for E_x ,

$$E_x = -\frac{\partial\phi}{\partial x} = \frac{\partial A_x}{\partial t}.$$

But if we do so, differentiating ϕ in equations (26.1), we get

$$E_x = \frac{q}{4\pi\epsilon_0 v^2(1-v^2)} \left[\frac{x - vx^2}{1 - v^2 + v^2 + z^2} \right] \hat{x}. \quad (26.2)$$

Similarly, for E_y ,

$$E_y = \frac{q}{4\pi\epsilon_0 v^2(1-v^2)} \left[\frac{y - vyt^2}{1 - v^2 + y^2 + z^2} \right] \hat{y}. \quad (26.3)$$

The z -axis project is a little more work. The derivative of ϕ is now complicated

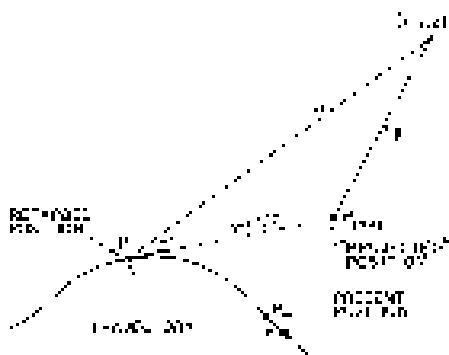


Fig. 24-3. A charge moves on an arbitrary trajectory. The potential's delayed at the time t are determined by the position \mathbf{r}' and velocity v at the retarded time $t' = t - r/c$. They are conveniently expressed in terms of the coordinates over the "projected" position \mathbf{r}_p . (The total position \mathbf{r} is fixed.)

and A_1 is not zero. First,

$$\frac{\partial A_1}{\partial t} = \frac{q}{c^2 r_0 \sqrt{1 - v^2/c^2}} \cdot \frac{(v - ct)(1 - v^2)}{\left[1 - \frac{(v - ct)^2}{c^2} + \frac{y^2}{r_0^2} + \frac{z^2}{r_0^2}\right]^{3/2}}. \quad (26.2)$$

Then, differentiating A_1 with respect to x , we find

$$\frac{\partial A_1}{\partial x} = \frac{q}{c^2 r_0 \sqrt{1 - v^2/c^2}} \cdot \frac{x^2(v - ct)(1 - v^2)}{\left[1 - \frac{(v - ct)^2}{c^2} + \frac{y^2}{r_0^2} + \frac{z^2}{r_0^2}\right]^{5/2}}. \quad (26.3)$$

And finally, taking the sum,

$$E_x = \frac{q}{4\pi \epsilon_0 c^2 r_0 \sqrt{1 - v^2/c^2}} \cdot \frac{x^2(v - ct)(1 - v^2)}{\left[1 - \frac{(v - ct)^2}{c^2} + \frac{y^2}{r_0^2} + \frac{z^2}{r_0^2}\right]^{5/2}}. \quad (26.4)$$

We'll look at the physics of E in a minute, let's first find B . For the z component,

$$B_z = \frac{\partial A_2}{\partial z} = \frac{\partial A_3}{\partial z}.$$

Since A_3 is zero, we have just one derivative to do. Notice, however, that A_2 is, and $\partial A_2/\partial z$ or $\partial A_3/\partial z$ of A_2 is just $-v E_x$. So

$$B_z = -v E_x. \quad (26.5)$$

Similarly,

$$B_y = \frac{\partial A_2}{\partial y} = \frac{\partial A_3}{\partial y} = -1 \cdot \frac{\partial A_3}{\partial z}$$

and

$$B_y = -v E_x. \quad (26.6)$$

Finally, B_x is zero, since A_1 and A_2 are both zero. We can write the magnetic field simply as

$$B = v \times E. \quad (26.7)$$

Now let's see what the fields look like. We will try to draw a picture of the field at various positions around the present position of the charge. It is true that the influence of the charge comes, in a certain sense, from the retarded position, but because the motion is exactly specified, the retarded position is uniquely given in terms of the present position. For uniform velocities, it's nice to relate the fields to the current position, because the field components at (x, y, z) depend only on $(v - ct)$, y , and z (which are the components of the displacement r_0 from the present position to (x, y, z)) (see Fig. 26-2).

Consider first a point with $v = 0$. Then E has only y and z components. From Eqs. (26.4) and (26.6), the ratio of these components is just equal to the ratio of the y - and z -components of the displacement. That means that E is in the same direction as r_0 , as shown in Fig. 26-3. Since E_r is also proportional to v , it is clear that this result holds in three dimensions. In short, the electric field is radial to the charge, and the field lines radiate directly out of the charge, just as they do for a stationary charge. Of course, the field isn't exactly the same as for the stationary charge, because v will be a factor of $(1 - v^2/c^2)$. But we can show something rather interesting: the difference is just v/c ! You know, if you were to have the constant field E in a position r_0 of coordinates in which the static field was expressed by the factor $\sqrt{1 - v^2/c^2}$. If you do that, the field lines will be curved and shifted and bent so the charge will be squeezed together around the ends, as shown in Fig. 26-4.

If we retrace the strength of E to the density of the field lines in the conventional way, we see a stronger field on the sides and a weaker field ahead and behind, which is just what the equations say. Furthermore, if we look at the strength of the field at right angles to the line of motion, that is, for $(v - ct) = 0$, the distance from

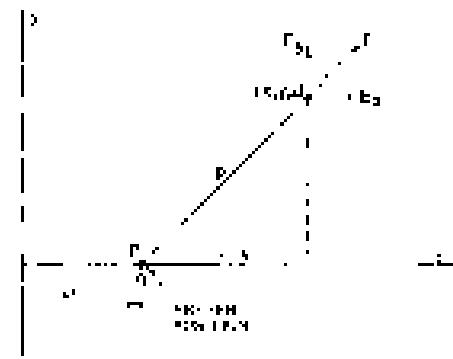


Fig. 26-2. For a charge moving with constant speed, the electric field points radially from the "present" position of the charge.

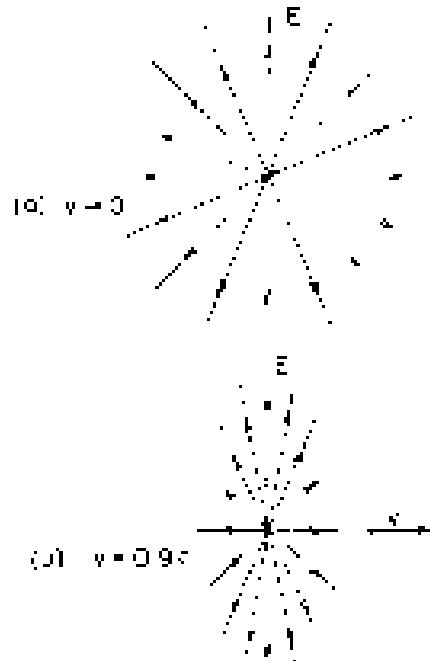


Fig. 26-4. The electric field of a charge moving with the constant speed $v = 0.9c$, part (b), compared with the field of a charge at rest, part (a).

the charge is $(1 - \gamma^{-1})\mathbf{e}$). Here the total field strength is $\sqrt{E_0^2 + E_1^2}$, which is

$$E = \frac{q}{4\pi\epsilon_0\gamma^2} + \frac{1}{c^2}\frac{\mathbf{v} \times \mathbf{B}}{\gamma^2}. \quad (26.10)$$

The field is proportional to the inverse square of the distance—just like the Coulomb field except increased by the constant extra factor $1/\gamma^2 = 1/\gamma$, which is always greater than one. So at the rest of a moving charge, the electric field is stronger than you get from the Coulomb law. In fact, the field in the sidewise direction is bigger than the Coulomb potential by the ratio of the energy of the particle to its rest mass.

Ahead of the charge (and behind), v and c are zero and

$$E = E_0 = \frac{q(1 - \gamma^2)}{4\pi\epsilon_0 c^2 \gamma^2}. \quad (26.11)$$

The field again varies as the inverse square of the distance from the charge but is now reduced by the factor $(1 - \gamma^2)$ in agreement with the picture of the field lines. If γ is small, γ^2/c^2 is still smaller, and the effect of the $1/(1 - \gamma^2)$ terms is very small; we get back to Coulomb's law. But if a particle is moving very close to the speed of light, the field in the forward direction is enormously reduced, and the field in the sidewise direction is enormously increased.

Our results for the electric field of a charge can be put this way: Suppose you were to draw on a piece of paper the field lines for a charge at rest, and then let the particle be travelling with the speed v . Then, of course, the whole picture would be compressed by the Lorentz contraction; that is, the career granules on the paper would appear in different places. The mistake of v , is that the picture you would see as the page lies by would still represent the field lines of the point charge. The contraction moves them closer together at the sides and spreads them out ahead and behind, just in the right way to give the correct line densities. We have emphasized because field lines are not real but are only one way of representing the field. However, here they almost seem to be real. In this particular case, if you make the mistake of thinking that the field lines are somehow really there in space, and transform them, you get the correct field. That doesn't, however, mean the field lines are *more* real. All you must do is remind yourself that they are *not* real, so that's what the electric fields perceived by a charge together will be unequal when the charges interact; new electric fields are produced, and *you* may feel it! I hope not. So the best idea of the contracting picture doesn't work if you act as if v , however, is the only way to remember what the fields *are*; it's better to think of them as like

The magnetic field is $\mathbf{v} \times \mathbf{E}$ [Eq. (26.9)]. If you take the velocity crossed only with the E -field, you get a B whose circles around the line of motion is shown in Fig. 26.5. If we put back E_0 , v , you will see that (in the same result we had for low-velocity charges). A good way to see where the v 's must go is to refer back to the Feynman diagram:

$$F = q(\mathbf{B} - \mathbf{v} \times \mathbf{E}).$$

You see that in velocity terms the magnetic field has the same dimensions as an electric field. So the right-hand side of Eq. (26.9) must have a factor $1/c^2$:

$$\mathbf{B} = \frac{q}{4\pi\epsilon_0 c^2} \frac{\mathbf{v} \times \mathbf{r}}{r^3}. \quad (26.12)$$

For a slow moving charge ($v \ll c$) we can take for E the Coulomb field (from

$$\mathbf{B} = \frac{q}{4\pi\epsilon_0 c^2} \frac{\mathbf{v} \times \mathbf{r}}{r^3} \quad (26.13)$$

this formula corresponds exactly to equations for the magnetic field of a current (see your book, Section 24.7).

We would like to point out, in passing, something interesting for you to think about. (We will come back to discuss it again later.) Imagine two electrons with velocities at right angles, so that one will cross over the path of the other, but in front of it, so they don't collide. At first, instead, the velocities would say will be as in Fig. 26-5(a). We look at the force on q_2 due to q_1 and vice versa. On q_2 , there is only the electric force from q_1 , since q_2 makes no magnetic field along its line of motion. On q_1 , however, there is again the electric force but, in addition, a magnetic force, since it is moving in a B -field made by q_2 . The forces are as shown in Fig. 26-5(b). The electric forces on q_1 and q_2 are equal and opposite. However, the E is a distance (magnetic) force on q_1 and an electric force on q_2 . Does action ever equal reaction? We leave it for you to worry about.

26-3 Relativity in Transformation of the Fields

In Fig. 26-2 we calculated the electric and magnetic fields from the non-relativistic potentials. The fields are important, of course, in spite of the arguments given earlier that there is physical meaning and reality to the potentials. The fields, too, are real. It would be nice just for many purposes to have a way to compute the fields in a moving system if you already know the fields in some "rest" system. We see the transformation laws for ϕ and A , because A is a four-vector. Now we would like to know the transformation laws of E and B . Given E and B in one frame, how do they look in another frame moving past? It is a concrete and rather transparent subject. We could always work back through the potentials, but it is much easier to be able to transform the fields directly. We will now see how that goes.

How can we find the transformation laws of the fields? We know the transformation laws of the ϕ and A , and we know how the fields are given in terms of ϕ and A ; it should be easy to find the transformation for the E and B . (You might think that with every vector there should be something to make it a four-vector, in which E there's got to be something else we can use for the fourth component. And also for B . But it's not so. It's quite different from what you would expect.) To begin with, let's take just a magnetic field B , which is, of course, $\nabla \times A$. Now we know that the vector potential has x , y , and z components; it's only a piece of something; there's also a t component. And we know that, for derivatives like ∇ , besides the x , y , z parts, there's also a derivative with respect to t . So let's try to figure out what happens if we replace a " x " by a " t ", or a " y " by a " t ", or something like that.

First, notice the form of the terms in $\nabla \times A$ when we write out the components:

$$B_x = \frac{\partial A_y}{\partial z} - \frac{\partial A_z}{\partial y}, \quad B_y = \frac{\partial A_z}{\partial x} - \frac{\partial A_x}{\partial z}, \quad B_z = \frac{\partial A_x}{\partial y} - \frac{\partial A_y}{\partial x}. \quad (26.14)$$

The t component is equal to a couple of terms that involve only y and z components. Suppose we call this combination of derivatives and components a "ything" and put it in a structure, E_{yz} . We simply mean that

$$E_{yz} = \frac{\partial A_x}{\partial y} - \frac{\partial A_y}{\partial z}. \quad (26.15)$$

Similarly, B_y is equal to the same kind of "thing," but this time it's an "x thing." And B_z is, of course, the corresponding z "thing." We have

$$B_x = E_{yz}, \quad B_y = E_{xz}, \quad B_z = E_{xy}. \quad (26.16)$$

Now what happens if we multiply by to convert all the "y" type things, like E_{yz} , and E_{xz} , (since none should be zero and symmetric in x , y , z , and t). For instance, what is E_{yt} ? If v_y , of course,

$$\frac{\partial A_t}{\partial z} = \frac{\partial A_z}{\partial t}.$$

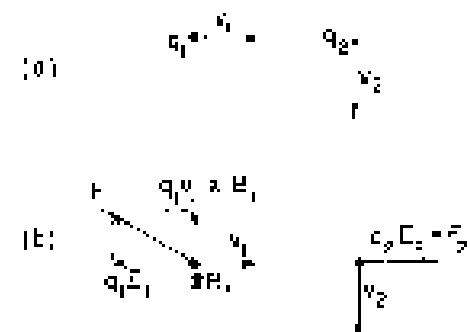


Fig. 26-5. The forces between two moving charges are not always equal and opposite. It appears that "action" is not equal to "reaction."

But remember that $A_0 = \phi$, so it is also

$$\frac{\partial A_0}{\partial t} = \frac{\partial \phi}{\partial t}.$$

You've seen that before: ϕ is the *asymptotic* part of A . Well, indeed, the sign is a sign wrong. But we forget that in the frame associated with the source charge comes with the opposite sign from x , y , and $-z$; so we should really have taken the more convenient expression of A_0 as

$$A_{0x} = \frac{\partial A_0}{\partial x} + \frac{\partial A_x}{\partial t}. \quad (26.17)$$

Then E_x is exactly equal to $-E_0$. Using also E_{0y} and E_{0z} , we find that the three possibilities are

$$E_{0x} = -E_{0y} = E_{0z} = -E_{0x} = E_0 = -E_0. \quad (26.18)$$

What happens if both subscripts are j ? Or, for that matter, if both are k ? We get things like

$$E_{0j} = \frac{\partial A_j}{\partial t} - \frac{\partial A_0}{\partial t},$$

and

$$E_{0k} = \frac{\partial A_k}{\partial x} - \frac{\partial A_x}{\partial k},$$

which give nothing but zero.

We have then six of these E -things. There are six more which you get by reversing the subscripts, but they give nothing really new, either:

$$E_{0j} = -E_{0k}.$$

So, we see, 8 out of sixteen possible combinations of the E_{0j} -subscripts taken in pairs, we get only six different physical objects; and they are the components of B and E .

To represent the *general* form of A , we will use two general subscripts just as we did in section 16.1, 1, 2, or j — meaning in our case four vector indices (x, y, z, t) . Also, everything will be suppressed with one superscript dot. Let us define E_{0j} as

$$E_{0j} = \nabla_j A_0 - \nabla_0 A_j. \quad (26.19)$$

It is now clear that $E_{0j} = (\partial A_0 / \partial x_j) - \partial A_0 / \partial t$ and that $A_0 = (t, A_x, A_y, A_z)$.

What we have found is that there are six quantities that build up an E — the six different aspects of the same thing. The electric and magnetic fields — which we have considered as separate entities in our intervening work (where we didn't worry about the speed of light) are not vectors in flat-space. They are parts of what "slings." Our physical "slings" is really the six-dimensional object E_0 . That is the way we must look at it, for relativity. We summarize our results on E_{0j} in Table 26-1.

Now, the first task we have to do here is to understand the *one* physical E . We begin with the *one* operation, and, in particular, the transformation properties of the E 's are the same as the transformation properties of our vector \mathbf{x} . In ordinary three-dimensional space we studied the position operator, which we typically denote like a vector. Let's focus for a moment on the ordinary cross product. This is a measure, for example, the angle between two vectors. When an object is moving in a plane, the quantity (v_x, v_y) is important. For motion in three dimensions, there are three such independent quantities, which we call the angular momentum:

$$L_{xy} = m v_x p_y - m v_y p_x, \quad L_{xz} = m v_x p_z - m v_z p_x, \quad L_{yz} = m v_y p_z - m v_z p_y.$$

Over all through you may have forgotten, we discussed in Chapter 20 of Vol. 1 the principle that these three quantities can be identified with the components

products of two vectors. In order to do so, we have to make an artificial rule with a right-hand rule again. It may not work. It was luck because ϵ_{ij} is zero if i and j equal to each other. This was our first symmetric identity:

$$E_{ij} = -E_{ji}, \quad E_{ii} = 0.$$

Of the nine possible quantities, there are only three independent numbers. And it just happens that when you change coordinate systems these three objects transform in exactly the same way as the components of a vector.

The same thing happens to represent an element of surface as a vector. A surface element has two parts, say $d\sigma$ and $d\hat{n}$, which we can represent by the vector \hat{n} normal to the surface. But we can't do that in four dimensions. What is the "normal" to a surface? Is it along \hat{n} or along $-\hat{n}$?

In short, for three dimensions it happens by luck that after you've taken a combination of two vectors like E_{ij} , you can represent it again by another vector because there are just three terms that happen to transform like the components of a vector. But in four dimensions that's evidently impossible, because there are six independent terms, and you can't represent six things by four things.

Even in three dimensions it's possible to have combinations of vectors that won't be represented by vectors. Suppose we take any two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ and make six various possible combinations of components, like $a_1 b_1, a_1 b_2, a_1 b_3, \dots$. There would be nine possible quantities:

$$\begin{aligned} a_1 b_1 &= a_2 b_2 &= a_3 b_3 \\ a_1 b_2 &= a_2 b_1 &= a_3 b_1 \\ a_1 b_3 &= a_2 b_3 &= a_3 b_2 \end{aligned}$$

We might call these quantities T_{ij} .

Let's now get the original coordinate system (by rotating about the x_3 axis), the new products of a and b unchanged. In the new system, a_3 , for example, gets replaced by

$$a'_3 = a_3 \cos \theta + a_1 \sin \theta,$$

and b_3 gets replaced by

$$b'_3 = b_3 \cos \theta + b_1 \sin \theta.$$

And similarly for other components. The nine components of the product quantity T_{ij} we have mentioned are all changed, too, of course. For instance, $T_{11} = a_1 b_1$ gets renamed to

$$T'_{11} = a'_1 b'_1 (\cos^2 \theta) + a'_1 b_2 (\cos \theta \sin \theta) + a'_1 b_3 (\sin \theta \cos \theta) = a'_1 b'_1 (\sin^2 \theta).$$

Or

$$T'_{11} = T_{11} \cos^2 \theta + T_{12} \cos \theta \sin \theta + T_{13} \sin \theta \cos \theta + T_{22} \sin^2 \theta.$$

Each component T_{ij} is a linear combination of components of T_{ij}' .

So we discover that it's not only possible to have a "vector" product like $a \cdot b$ at one time, it's still possible that a and b are basis vectors, but we end up with a six-vector (another kind of "product") of two vectors T_{ij} , with nine components. But translation under a rotation by a simple law of rules (that we could figure out). Such a thing is called a tensor, and it's the most basic type of tensor. It's called a tensor of the second rank, "because you can play the game with three vectors and get a tensor or the elements, or with four to get a tensor of the fourth rank, and so on." A tensor of the first rank is a vector.

The aim of all this is that our electric magnetic quantity E_{ij} is also a tensor of the second rank, because it has two indices, i and j . It is, however, a tensor in four dimensions. It transforms in a special way which we will work out in more detail. It's just the way a product of vectors transforms. If $a \cdot b$, it happens that if you change the index orientation, E_{ij} changes sign. That's a special case. In a

an antisymmetric tensor. So we say: the electric and magnetic fields are part of an antisymmetric tensor of the second rank in four dimensions.

You've come a long way. Remember why such things were called an "electric circuit"? Now we are talking about "an antisymmetric tensor of the second rank in four dimensions."

Now we have to find the law of the transformation of $E_{\mu\nu}$. It isn't at all difficult to do; it's just tedious—the matrix involved is real, but the work is not. What we want is the Lorentz transformation of $E_{\mu\nu} = E_{\nu\mu}$. Since $E_{\mu\nu}$ is just a special case of a vector, we will work with the general solution of the vector transformation, which we can call $G_{\mu\nu}$:

$$G_{\mu\nu} = a_\mu b_\nu - a_\nu b_\mu. \quad (26.20)$$

(For this purpose, a_0 will eventually be replaced by ∇_j and b_0 will be replaced by the potential A_{j0} .) The components of a_μ and b_μ transform like the Lorentz basis, which is:

$$\begin{aligned} a^0 &= \frac{a_0 + \omega a_\perp}{\sqrt{1 - v^2}}, & b^0 &= \frac{b_0 - vb_\perp}{\sqrt{1 - v^2}}, \\ a^\perp &= \frac{a_\perp - \omega a_0}{\sqrt{1 - v^2}}, & b^\perp &= \frac{b_\perp + vb_0}{\sqrt{1 - v^2}}, \\ a'_0 &= a_0, & b'_0 &= b_0, \\ a'_\perp &= a_\perp, & b'_\perp &= b_\perp. \end{aligned} \quad (26.21)$$

Now let's transform the components of $G_{\mu\nu}$. We start with $a_0 b_0 - G_{00}$:

$$\begin{aligned} G_{00} &= a_0 b_0 - a'_0 b'_0 \\ &= \left(\frac{a_0 + \omega a_\perp}{\sqrt{1 - v^2}} \right) \left(\frac{b_0 - vb_\perp}{\sqrt{1 - v^2}} \right) - \left(\frac{a_\perp - \omega a_0}{\sqrt{1 - v^2}} \right) \left(\frac{b_\perp + vb_0}{\sqrt{1 - v^2}} \right) \\ &= a_0 b_0 - a'_0 b'_0. \end{aligned}$$

But $a_0 = b_0$, so we have the simple result

$$G_{00} = a_0 b_0.$$

We take the same steps

$$G_{0\perp} = \frac{a_0 + \omega a_\perp}{\sqrt{1 - v^2}} b_\perp - a'_0 \frac{b_0 - vb_\perp}{\sqrt{1 - v^2}} = (a_0 b_\perp - a'_0 b'_\perp) \frac{\sqrt{1 - v^2}}{\sqrt{1 - v^2}} = a_0 b_\perp - a'_0 b'_\perp.$$

So we get the:

$$a'_0 b'_\perp = \frac{a_0 b_\perp - a'_0 b'_\perp}{\sqrt{1 - v^2}}.$$

And, of course, at the same way,

$$G_{\perp 0} = \frac{a_0 b_\perp - a'_0 b'_\perp}{\sqrt{1 - v^2}}.$$

It is clear how the rest will go: with the help of induction it is obvious, only now we may as well write them for $E_{\mu\nu}$:

$$\begin{aligned} E'_{00} &= E_{00}, & E'_{0\perp} &= \frac{E_{0\perp} - vE_{\perp 0}}{\sqrt{1 - v^2}}, \\ E'_{\perp 0} &= \frac{E_{\perp 0} + vE_{0\perp}}{\sqrt{1 - v^2}}, & E'_{\perp\perp} &= E_{\perp\perp}, \\ E'_{0\perp} &= \frac{E_{0\perp} - vE_{\perp 0}}{\sqrt{1 - v^2}}, & E'_{\perp 0} &= E_{\perp 0}, \\ E'_{\perp\perp} &= \frac{E_{\perp\perp} + vE_{0\perp}}{\sqrt{1 - v^2}}, & E'_{00} &= \frac{E_{00} - vE_{0\perp}}{\sqrt{1 - v^2}}. \end{aligned} \quad (26.22)$$

From this we will have $E'_0 = -E'_0$ and $E'_0 = 0$.

So we have the transformation of electric and magnetic fields. All we have to do is look at Table 26-1 to convert what our given notation in terms of E , terms of terms of E and B , into just one other of substitution. So that we can see how it looks in the ordinary symbols, and finally our transformation of the field components in Table 26-2.

Table 26-2

The Lorentz transformation of the electric and magnetic fields (Note: $c = 1$)

$E_x' = E_x$	$B_z' = B_z$
$E_y' = \frac{E_y - vB_z}{\sqrt{1 - v^2}}$	$B_y' = \frac{B_y + vE_z}{\sqrt{1 - v^2}}$
$E_z' = \frac{E_z + vB_y}{\sqrt{1 - v^2}}$	$B_x' = \frac{B_x - vE_y}{\sqrt{1 - v^2}}$

The equations in Table 26-2 tell us how E and B change if we go from one inertial frame to another. If we know E and B in one system, we can find what they are in another that moves by with the speed v .

We can write these equations in a form that is easy to remember. We notice that since v is in the x -direction, all the terms will be related to components of the y -axis produce $v \times E$ and $v \times B$. So we can rewrite the equations, which are shown in Table 26-3.

Table 26-3

An alternative form for the field transformations (Note: $c = 1$)

$E_x' = E_x$	$B_z' = B_z$
$E_y' = \frac{(E_y + vB_z)}{\sqrt{1 - v^2}}$	$B_y' = \frac{(B_y - vE_z)}{\sqrt{1 - v^2}}$
$E_z' = \frac{(E_z - vB_y)}{\sqrt{1 - v^2}}$	$B_x' = \frac{(B_x + vE_y)}{\sqrt{1 - v^2}}$

It is now easier to remember which components go where. In fact, the transformation can be written even more simply if we define the field components along x as the "parallel" components E_x and B_x (because they are parallel to the relative velocity of S and S'), and the total transverse components (the vector sum of the y and z components) as the "perpendicular" components E_y and B_y . Then we get the equations in Table 26-4. (We have just put back the $\sqrt{1 - v^2}$, as it will be more convenient when we want to refer back to it.)

Table 26-4

Still another form for the Lorentz transformation of E and B

$E_x' = E_x$	$B_z' = B_z$
$E_y' = \frac{(E_y - vB_z)}{\sqrt{1 - v^2}}$	$B_y' = \frac{\left(B_y + \frac{vE_z}{c}\right)}{\sqrt{1 - v^2}}$

The field transformation part is another way of solving some problems we have done before. For instance, in finding the fields of a moving point charge. We have used and the fields before by calculating the potentials. But we could now do it by transforming the Lorentz field. If we have a point charge at rest in the S -frame, then there is only the unperturbed E -field. In the S' -frame we will see a point charge moving with the velocity v , if the S' -frame moves by the

A frame with the speed $v \ll c$. We will let you show that the transformations of Table 26-3 and 26-4 give the same electric and magnetic fields we got in Section 26-2.

The transformation of Table 26-2 gives us an interesting and simple answer for what we see if we move past a system of fixed charges. For example, suppose we want to know the fields in our frame S' if we are moving along between the plates of a condenser, as shown in Fig. 26-7. (It is, of course, the same thing if we say that a charged condenser is moving past us.) What do we see? The transformation is easy in this case because the B field in the original system is zero. Suppose, first, that our motion is perpendicular to E . Then we will see an E' : $E'_{\parallel} = v/c E_{\perp}$, which is still completely transverse. We will see, in addition, a magnetic field $B'_{\parallel} = v/c E_{\perp}$. (The $\sqrt{1 - v^2/c^2}$ doesn't appear in our formula for B' because we wrote it in terms of E' rather than E , but it's the same thing.) So when we move along perpendicular to a static electric field, we see a reduced E and an added transverse B . If our motion is not perpendicular to E , we break E into E_{\parallel} and E_{\perp} . The parallel part is unchanged, $E'_{\parallel} = E_{\parallel}$, and the perpendicular component does as just described.

Let's take the opposite case, and imagine we are moving through a pure static magnetic field. This time we would see an electric field E' equal to $v \times B$, and the magnetic field changed by the factor $\sqrt{1 - v^2/c^2}$ (assuming it is transverse). So long as v is much less than c , we can neglect the change in the magnetic field, and the main effect is that an electric field appears. As one example of this effect, consider this once famous problem of determining the speed of an airplane flying eastward, since radar can now be used to determine the air speed (it's no longer necessary, since radar can now be used to determine the air speed from ground references), but for many years it was very hard to find the speed of an airplane in bed weather. You could not see the ground, and you didn't know which way was up, and so on. Yet it was important to know how fast you were moving relative to the earth. How can this be done without seeing the earth? Many who knew the transformation formula thought of the idea of using the fact that the airplane moves in the magnetic field of the earth. Suppose that an airplane is flying, where there is a magnetic field, more or less known. Let's just take the simplest case where the magnetic field is vertical. If we were flying, straight west at a constant velocity v , according to our formula, we should see an electric field which is $v \times B$, i.e., perpendicular to the line of motion. If we hang an ordinary wire across the airplane, this electric field will induce charges on the ends of the wire. That's nothing new. From the point of view of someone on the ground, we are moving now to the left, and the $v \times B$ from static charges induces no charges on the ends of the wire. The transformation formulas just say the same thing in a different way. (The fact that we can say this thing more than one way doesn't mean that one way is better than another. We are getting so many different methods and tools that we can usually get the same result in a different way.)

So to measure v , all we have to do is measure the voltage between the ends of the wire. We can't do it with a voltmeter because the same fields will act on the wires in the voltmeter, but there are ways of measuring such fields. We talked about some of them when we discussed atmospheric electricity in Chapter 2. So it should be possible to measure the speed of the airplane.

The important problem was, however, never solved this way. The reason is that the electric field that is developed is of the order of millionths per meter. It is possible to measure such fields, but the trouble is that these fields are, unfortunately, not very different from any other electric fields. The field that is produced by motion through the magnetic field can't be distinguished from static electric fields that was already in the air from another cause, say from electrostatic charges in the air, or in the clouds. We described in Chapter 9 that there are, typically, electric fields above the surface of the earth with strengths of about 100 volt per meter. But they are quite irregular. So as the airplane flies through the air, it sees fluctuations of atmospheric electric fields which are enormous in comparison to the tiny field produced by the $v \times B$ term, and it turns out for practical reasons to be impossible to measure speeds of an airplane by its motion through the earth's magnetic field.

26.4 The equations of motion in relativistic notation*

I doesn't do much good to find electric and magnetic fields from Maxwell's equations unless we know what the fields do when we have them. You may remember that the fields are required to find the forces on charges, and that those forces determine the motion of the charge. So, of course, part of the theory of electrodynamics is the relation between the motion of charges and the forces.

For a single charge in the fields E and B , the force is

$$\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B}. \quad (26.23)$$

This force is equal to the mass times the acceleration (in low velocities), but the correct law for any velocity is that the force is equal to dp/dt . Writing $p = m\gamma\sqrt{1 - v^2/c^2}$, we find that the relativistically correct equation of motion is

$$\frac{d}{dt} \left(\frac{mv}{\sqrt{1 - v^2/c^2}} \right) = \mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B}. \quad (26.24)$$

We would like now to discuss this equation from the point of view of relativity. Since we have our four Maxwell equations in relativistic form, it would be interesting to see what the equations of motion would look like in relativistic form. Let's see whether we can rewrite the equation in a four-vector notation.

We know that the momentum is part of a four-vector \mathbf{p} , whose time component is the energy $m\gamma$, $\gamma = 1/\sqrt{1 - v^2/c^2}$. So we might think to replace the left-hand side of Eq. (26.24) by $d\mathbf{p}/dt$. Then we need only find a fourth component to go with F . This fourth component must equal the rate-of-change of the energy, or the rate of doing work, which is $F \cdot v$. We would then like to write the right-hand side of Eq. (26.24) as a fourvector like (F_0, F_x, F_y, F_z) . But this does not make a four-vector.

The rate derivative of a fourvector is no longer a fourvector, because the *drift* respects the cause of zero spatial frame for measuring it. We got into that trouble before when we tried to make v into a fourvector. Our first guess was that the new components would be $v \partial/\partial t - v$. But the quantities

$$\left(\gamma, \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \right) = (\gamma, \mathbf{v}) \quad (26.25)$$

are not the components of a fourvector. We found that they could be made into one by multiplying each component by $1/\gamma^2 = v/c^2$. The "four-velocity" η_μ is the fourvector

$$\eta_\mu = \left(\gamma, \frac{c}{\gamma}, \frac{v_x}{\gamma}, \frac{v_y}{\gamma}, \frac{v_z}{\gamma} \right). \quad (26.26)$$

So it appears that the trick is to multiply $d\mathbf{p}/dt$ by $1/\gamma^2 = 1/v^2$, if we want the derivatives to make a fourvector.

Our second guess then is that

$$\frac{1}{\sqrt{1 - v^2/c^2}} \frac{d}{dt} (\mathbf{p}_\mu) \quad (26.27)$$

should be a fourvector. But, wait! It is the velocity of the particle — v_μ , of a coordinate t, x, y, z . Then the quantity \mathcal{F}_μ defined by

$$\mathcal{F}_\mu = \left(\frac{\mathbf{F} \cdot \mathbf{v}}{\sqrt{1 - v^2/c^2}}, \frac{\mathbf{F}}{\sqrt{1 - v^2/c^2}} \right) \quad (26.28)$$

is the extension into four dimensions of a force—we can call it the "four-force." It is indeed a fourvector, and its space components are not the components of \mathbf{F} but of $\mathbf{F}/\sqrt{1 - v^2/c^2}$.

...

* In this section we neglect all of the c^2 s.

The question is: "why is γ a four-vector?" It would be nice to get a little understanding of that fact. After all, since it has come up before now, it is time to see why the γ 's can always be fixed by the same factor. The answer is in the following: When we take the time derivative of some function x , we compute the increment Δx in a small interval Δt in the variable t . But in another frame, the interval Δt might correspond to a change in both t' and x' , so if we vary only t' , the change in x will be different. We have to find a variable for our differentiation that is a measure of an "interval" in space-time, which will then be the same in all coordinate systems. When we take Δx for that interval, it will be the same for all coordinate frames. When a particle "moves" in four-space, choose these ranges $\Delta t, \Delta x, \Delta p, \Delta s$. Can we make an invariant interval out of them? Well, they are the components of the four vector $v_\mu = (c, \mathbf{x}, \mathbf{p})$ so if we choose a quantity Δs by

$$(\Delta s)^2 = \frac{1}{c^2} \Delta t_\mu \Delta s_\mu = \frac{1}{c^2} (c^2 \Delta t^2 - \Delta x^2 - \Delta p^2 + \Delta s^2) \quad (26.29)$$

—which is a four-dimensional dot product—now we have a good four scalar to use as a measure of a four-dimensional interval. For the Δs in its turn ds , we can define a parameter $s = \int ds$. And a derivative with respect to s , ds/ds , is a like four-dimensional operation, because it is invariant with respect to a Lorentz transformation.

It is easy to relate ds to dt for a moving particle. For a moving point particle,

$$dx = v_x dt, \quad dv = v_x' dt, \quad ds = v_x dt, \quad (26.30)$$

and

$$ds = \sqrt{(v_x)^2/c^2} dx^2 = v_x^2 - v_y^2 - v_z^2 = v_x^2 - v^2 = d\tau v \quad = v^2/c. \quad (26.31)$$

So the operator

$$\frac{d}{ds} = \frac{1}{\sqrt{1 - v^2/c^2}} \frac{d}{dt}$$

is an "invariant operator". If we operate it on any function with it, we get another four-vector. For instance, if we operate on $(1, \mathbf{x}, \mathbf{p}, \mathbf{f})$, we get the four velocity v_μ :

$$\frac{dv_\mu}{ds} = u_\mu.$$

We see now why the factor $\sqrt{1 - v^2/c^2}$ fixes things up.

The invariant variable s is a useful physical quantity. It is called the "proper time" along the path of a particle, because ds is always an interval of time for some s that is moving with the particle at any particular instant. (That is, $\Delta s = \Delta \tau = \Delta x = 0$, and $\Delta t = \Delta s$.) If you can imagine some "clock" whose rate doesn't depend on the acceleration, such a clock carried along with the particle would show the time s .

We can now go back and write Newton's law (as corrected by Einstein) in the next form:

$$\frac{dp_\mu}{ds} = f_\mu \quad (26.32)$$

where f_μ is given in Eq. (26.28). Also, the momentum p_μ can be written as

$$p_\mu = m u_\mu = m u_\mu \frac{ds}{dt}, \quad (26.33)$$

where the coordinates $x_\mu = (ct, \mathbf{x})$ now describe the trajectory of the particle. Finally, the four-dimensional notation gives us this very simple form of the equations of motion:

$$f_\mu = m u_\mu \frac{d^2 x_\mu}{ds^2}. \quad (26.34)$$

which is reminiscent of $\ddot{x} = a_0$. It is important to notice that Eq. (26.34) is not the same as $\ddot{x} = a_0$, because the transverse formula Eq. (26.14) has m in the Eq. 12

non-relativistic mechanics which are different from Newton's. Now the high velocities fit in with the rest of Maxwell's equations, where we were able to rewrite the equations in the relativistic form *almost*, except for the *massless* result, or with just a change of notation.

Now let's return to Eq. (26.24) and see how we can write the right-hand side in four-vector notation. The three components, when divided by $\sqrt{1 - v^2/c^2}$, are the components of $v^2 f_0$, so

$$f_0 = \frac{q(E + v \times B)_x}{\sqrt{1 - v^2/c^2}} = q \left[\frac{E_x}{\sqrt{1 - v^2/c^2}} + \frac{v_x B_y}{\sqrt{1 - v^2/c^2}} - \frac{v_y B_x}{\sqrt{1 - v^2/c^2}} \right]. \quad (26.35)$$

Now we must put all quantities in their relativistic notation. First, $v/\sqrt{1 - v^2/c^2}$ and $v_x/v = v_x/c$ and $v_y/v = v_y/c$ are the x , y , and z -components of the four velocity v_0 . And the components of E and B are components of the second rank tensor of the fields $F_{\mu\nu}$. Looking back in Table 26-1 for the components of $F_{\mu\nu}$ that correspond to E_x , B_x , and B_y , we get

$$f_0 = q(v_0 E_{xx} + v_0 F_{xy} - v_0 F_{yz}),$$

which begins to look impressive. Every term has the subscript x , which is reasonable, since we're finding the x -component. Then all the others appear in pairs: y , z , xy , yz , except that the xy term is missing. So we just add it in, and write

$$f_0 = q(v_0 E_{xx} + v_0 F_{xy} + v_0 F_{yz} - v_0 F_{yy}). \quad (26.36)$$

We haven't changed anything because F_{yy} is antisymmetric, and F_{yy} is zero. The reason for wanting to put in the xy terms is so that we can write Eq. (26.36) in the shorthand form

$$f_0 = q v_0 F_{xx}. \quad (26.37)$$

This equation is the same as Eq. (26.36) if we make the rule that whenever any subscript appears twice, it's understood that you automatically sum over terms in the same way as the scalar product, using the same convention for the signs.

You can easily believe that (26.37) works correctly with the $v = c$, $c^2 = \infty$, but what about $v = c^2$? Let's see, for fun, what it says:

$$f_0 = q v_0 F_{xx} + v_0 F_{yy} - v_0 F_{yz} + v_0 F_{xy}.$$

Now we have to translate each term in $F_{\mu\nu}$ s and $B_{\mu\nu}$ s. We get

$$f_0 = q \left(0 - \frac{v_x}{\sqrt{1 - v^2/c^2}} E_x - \frac{v_y}{\sqrt{1 - v^2/c^2}} E_y - \sqrt{1 - v^2/c^2} B_z \right),$$

or

$$f_0 = \frac{qv \cdot E}{\sqrt{1 - v^2/c^2}}. \quad (26.38)$$

But from Eq. (26.26), f_0 is supposed to be

$$\frac{F \cdot v}{\sqrt{1 - v^2/c^2}} = \frac{q(E + v \times B) \cdot v}{\sqrt{1 - v^2/c^2}},$$

That's the same thing as Eq. (26.13), since $(v \times B)$ is zero. So everything comes out all right.

Summarizing, our equation of motion can be written in the elegant form

$$m_0 \frac{d^4 x_\mu}{dt^4} = f_0 = q v_0 F_{xx}. \quad (26.39)$$

Although it is nice to see that the equations can be written that way, this form is not particularly useful. It's usually more convenient to solve for particle motion by using the four-graviton equations (26.24), and that's what we will usually do.

Field Energy and Field Momentum

27-1 Local conservation

It is clear that the energy of matter is not conserved. When an object radiates light it loses energy. However, the energy lost is possibly recoverable in some other way, say in the light. Therefore the theory of the conservation of charge is incomplete without a consideration of the energy which is associated with the light, or, in general, with the electromagnetic field. We take up now the law of conservation of energy and a set of momentum, or the light. Certainly, we cannot treat one without the other, because in the relativity theory they are different aspects of the same four-vector.

Very early in Volume I, we discussed the conservation of energy; we said then merely that the total energy in the world is constant. Now we want to extend the idea of the energy conservation law in an important way—in a way that says something in detail about how energy is conserved. The new law will say that if energy goes away from a region, it's because it flows away through the boundaries of that region. It is a somewhat stronger law than the conservation of energy without such a restriction.

To see what the statement means, let's look at how the law of the conservation of charge works. We described the conservation of charge by saying that, there is a current density j and a charge density ρ , and that when the charge decreases at some place there must be a flow of charge away from that place. We call that the conservation of charge. The mathematical form of the conservation law is

$$\nabla \cdot j = -\frac{\partial \rho}{\partial t}, \quad (27.1)$$

This law has the consequence that the total charge in the world is always constant—there is never any net gain or loss of charge. However, the total charge in the world could be constant in another way. Suppose that there is some charge Q_1 near some point (1) while there is no charge near some point (2) some distance away (Fig. 27-1). Now suppose that, as time goes on, the charge Q_1 were to gradually fade away and that simultaneously with the decrease of Q_1 some charge Q_2 would appear near point (2), and in such a way that at every instant the sum of Q_1 and Q_2 was a constant. In other words, at any one measurable state the amount of charge lost by Q_1 would be added to Q_2 . Then the total amount of all charge in the world would be conserved. That is a "world-wide" conservation, but not what we will call a "local" conservation, because in order for the charge to get from (1) to (2) it didn't have to appear somewhere in the space between point (1) and point (2). Locally, the charge was just "lost."

There is a difficulty with such a "world-wide" conservation law, even in the theory of relativity. This concept of "global charge" means that a distant particle is one which is not experienced in different systems. Two events that are simultaneous in one system are not simultaneous in another system moving past. For "world-wide" conservation of the kind described, it is necessary that the charge lost from Q_1 should appear simultaneously in Q_2 . Otherwise there would be some moments when the charge was not conserved. There seems to be no way to make the law of charge conservation relativistically invariant without making it a "local" conservation law. As a matter of fact, the requirement of the Lorentz relativity seems to restrict the possible laws of nature in surprising ways. In modern quantum field theory, for example, people have often wanted to alter the theory by allowing what we call a "nonlocal" interaction—where something here

27-1 Local conservation

27-2 Energy conservation and electromagnetism

27-3 Energy density and energy flow in the electromagnetic field

27-4 The ambiguity of the field energy

27-5 Examples of energy flow

27-6 Field momentum

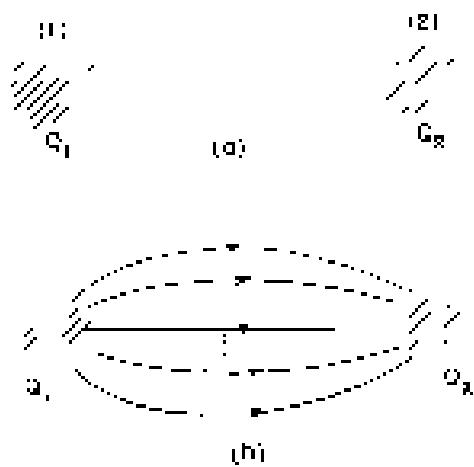


Fig. 27-1. Two ways to conserve charge: (a) $Q_1 + Q_2$ is constant; (b) $dQ_1/dt = i_j \cdot n \hat{n} = -dQ_2/dt$.

has a direct effect on working there—but we just introduced with the relativity principle.

"Local" conservation involves another part. It says that a charge can get from one place to another only if there is something moving in the space between. To describe this law we need not only the density of charge ρ , but also another kind of quantity, namely j , a vector giving the rate of flow of charge across a surface. Then the law is related to the rate of change of charge by Eq. (27.1). This is the more extensive kind of a conservation law. It says that charge is conserved in a special way—what we "locally."

It turns out that energy can have fluxes like a fluid process. There is not only energy density in a given point of space but also a vector to represent the rate of flow of the energy through a surface. For example, when light source radiates, we can find the rate of energy flow out from the source. If we imagine some mathematical surfaces around the light source, the energy lost from inside the surface is equal to the energy that flows out through the surface.

27-2 Energy conservation and electromagnetism

We want now to write quantitatively the conservation of energy for electromagnetism. To do that, we have to describe how much energy there is in any volume element of space, and also the rate of energy flow. Suppose we think first only of the electromagnetic field energy. We will let ϵ represent the energy density in the field (that is, the amount of energy per unit volume in space) and the vector S represent the energy flux of the field (that is, the flow of energy per unit time across a unit area perpendicular to the flow). Then, in parallel analogy with the conservation of charge, Eq. (27.1), we can write the "local" law of energy conservation in the field as

$$\frac{du}{dt} = -\nabla \cdot S. \quad (27.2)$$

Of course, this law is not true in general; it is not true that the field energy is conserved. Suppose you are in a dark room, and then turn on the light, say left. All of a sudden, the room is full of light, so there is energy in the light, although there wasn't any energy there before. Equation (27.2) is not the complete conservation law, because the field energy alone is not conserved; only the total energy in the world—matter is also the energy of matter. The field energy *leaves* if there is some work being done by matter on the field or by the field on matter.

However, if there is matter inside the volume of interest, we know how much energy it has. Each particle has the energy $m c^2 \sqrt{1 - v^2/c^2}$. The total energy of the matter is just the sum of all the particle energies, and the flow of this energy through a surface is just the sum of the energy carried by each particle that crosses the surface. We want now to take only account of the energy in the electromagnetic field. So we must write an equation which says just the total field energy in a given volume decreases either because field energy flows out of the volume or because the field loses energy to matter (or gains energy, which is just a negative loss). The field energy inside a volume V is

$$\int_V u dV,$$

and its rate of decrease is minus the time derivative of this integral. The law of field energy out of the volume V is the integral of the normal component of S over the surface Σ that encloses V ,

$$\int_{\Sigma} S \cdot n d\sigma,$$

so

$$-\frac{\partial}{\partial t} \int_V u dV = \int_{\Sigma} S \cdot n d\sigma + (\text{work done on matter outside } V) \quad (27.3)$$

We have seen before that the field does work on each unit volume δV in the field at the rate $E \cdot j$. [The flux density a satisfies $a \cdot E = q(E \cdot a - B \times B)$, and the rate of doing work is $E \cdot a = qE \cdot a$.] If there are N particles per unit volume, the rate of doing work per unit volume is $NqE \cdot a$, but $Nqz = j$. So the quantity $E \cdot j$ must be equal to the loss of energy per unit time and per unit volume by the field. Equation (27.3) then becomes

$$\frac{\partial}{\partial t} \int_V a \cdot dV = \int_V S \cdot a \, dz + \int_V E \cdot j \, dV. \quad (27.4)$$

It is an interesting thought for energy in the field. We can convert it into a different equation like Eq. (27.2) if we can change the second term to a volume integral. That is, easy to do with Green's theorem. The surface integral of the normal component of S is the integral of its divergence over the volume inside. So Eq. (27.4) is equivalent to

$$\int_V \frac{da}{dt} \, dV = \int_V \nabla \cdot S \, dV + \int_V E \cdot j \, dV,$$

where we have put the time derivative of the first term inside the integral. Since this equation is true for any volume, we can take away the integrals and we have the energy equation for the electromagnetic field:

$$-\frac{\partial u}{\partial t} = \nabla \cdot S + E \cdot j. \quad (27.5)$$

Now this one you doesn't expect a lot of good unless we know what u and S are. And say we should just tell you who they are in terms of E and B , because all we really want is the result. However, we would rather show you the kind of argument that was used by Feynman in 1944 to obtain formulas for S and u , so you can see where they came from. (You won't, however, need to learn this derivation for all other work.)

27-5 Energy density and energy flux in the electromagnetic field

It is often useful to suppose that the e is a field energy density u and the A that depends only upon the fields E and B . (For example, we know that in electrostatics, A and the energy density u are written $\frac{1}{2}\epsilon_0 A \cdot E$.) Otherwise, the E and B might depend on the potentials or something else, but let's see what we can work out. We can try to write the quantity $E \cdot j$ in such a way that it becomes the sum of two terms, one that is the time derivative of one quantity and another that is the divergence of a second quantity. The first quantity would then be u and the second would be S (with suitable signs). Both quantities must be written in terms of the fields only; that is, we want to write our equality as

$$E \cdot j = -\frac{\partial u}{\partial t} - \nabla \cdot S. \quad (27.6)$$

The problem is that j must be expressed in terms of the fields only. How can we do that? By using Maxwell's equations, of course. Then Maxwell's equation for the curl of B ,

$$j = \epsilon_0 c^2 \nabla \times B + \epsilon_0 \frac{\partial B}{\partial t},$$

Substituting this in (27.6) we will have only E 's and B 's.

$$E \cdot j = \epsilon_0 c^2 E \cdot (\nabla \times B) - c \cdot E \frac{\partial B}{\partial t}. \quad (27.7)$$

We are almost partly finished—the last term is a time derivative; it is $(\partial/\partial t)(c \cdot E \times B)$. So $c \cdot E \times B$ is at least one part of u . It's the same thing we found in electrostatics. Now, all we have to do is to make the other term into the divergence of something.

Notice that the first term in the right-hand side of (27.7) is the same as

$$(\nabla \times \mathbf{B}) \cdot \mathbf{E}. \quad (27.8)$$

And, as you know from earlier lectures, if $a \geq 0$, a is the same as $a \cdot (b \times c)$. You can then justify the equation

$$\nabla \cdot (\mathbf{B} \times \mathbf{E}) \quad (27.9)$$

and we have the divergence of "something," just as we wanted. Only that's wrong. We warned you before that, ∇ is "like" a vector, but not "exactly" the same. The reason it is not is because there is an additional *convention* here, concerning what a derivative operator is in front of a product. It works on everything to the right. In Eq. (27.7), the ∇ operates only on \mathbf{B} , not on \mathbf{E} . But, in fact, in (27.9), the normal convention would say that ∇ operates on both \mathbf{B} and \mathbf{E} . So it's not the same thing. In fact, if we take out the components of $\nabla \cdot (\mathbf{B} \times \mathbf{E})$ we can see that it is equal to $\mathbf{E} \cdot (\nabla \times \mathbf{B})$ plus some other terms. It's like when you multiply two vectors and then swap them around. For instance,

$$\frac{\partial}{\partial x}(y\mathbf{z}) = \frac{\partial y}{\partial x}\mathbf{z} + y\frac{\partial \mathbf{z}}{\partial x}$$

Rather than working out all the components of $\nabla \cdot (\mathbf{B} \times \mathbf{E})$ we would like to show you a trick that is very useful for this kind of problem. It is a trick that allows you to use all the rules of vector algebra on expressions with the ∇ operator, without getting into trouble. The trick is to throw out—for a while at least—the rule of the calculus notation about what the derivative operator works on. You see, obviously, the order of terms is used for two separate purposes. One is for derivatives of f (where f is not the same as $\text{rot}(g)\mathbf{f}$), and the other is for vectors: $\mathbf{a} \times \mathbf{b}$ is differentiation, $\mathbf{b} \times \mathbf{a}$. We cannot we want choose or abandon incrementarily the \mathbf{a} -derivative rule. Instead of saying that a derivative operates on everything to its right, we make a new rule that it doesn't depend on the order in which terms are written down. Then we can jumble terms around without worrying.

Here is our new convention: we know, by a subscript, what a well-defined operator works on: the *variables* are missing. Suppose we let the operator D_1 stand for $\partial/\partial x_1$. Then D_1 means \mathbf{i} . Only the derivative of the variable quantity i is taken. Then

$$D_1 f = \frac{\partial f}{\partial x_1}$$

But if we have $D_1 g$, it means

$$D_1 g = \left(\begin{matrix} 0 \\ 1 \end{matrix} \right) g$$

But notice now that according to our new rule, $/ D_1$ means the same thing. We can write the same thing any which way.

$$D_1 f \mathbf{i} = \mathbf{i} D_1 f = f D_{1i} = f_i D_1$$

You see, f is D_1 , an even更深 after everything. It's surprising that such a fancy notation is necessary just for a few basic mathematics or physics.

You may wonder: What if I ever to write the derivative of fg ? I mean the derivative of f with respect to g . That's easy, you just say so, you write $D_g(fg) + D_f(g)$. That is just $g D_g(f) + f D_g(g)$ which is what you mean in the old notation by $D_g(fg)$.

You will see that it is now going to be very easy to work out a new expression for $\nabla \cdot (\mathbf{B} \times \mathbf{E})$. We start by changing to the new notation: we write

$$\nabla \cdot (\mathbf{B} \times \mathbf{E}) = \nabla_x \cdot (\mathbf{B} \times \mathbf{E}) + \nabla_y \cdot (\mathbf{B} \times \mathbf{E}) \quad (27.10)$$

The important we note that we won't need to keep the order straight anymore. We always know that ∇_x operates on \mathbf{B} only, and ∇_y operates on \mathbf{B} only. In these circumstances, we can use ∇ as though it were an ordinary vector. (Of course, \mathbf{E} is)

when we are finished, we will want to return to the "standard" notation that everybody usually uses). So now we can do the various things like interchanging dots and crosses and making other kinds of rearrangements of the terms. For instance, the initially form of Eq. (27.10) can be rewritten as $E \cdot \nabla_B \times B$. (You remember that $a \cdot b \times c = b \cdot c \times a$.) And the last term is the same as $B \cdot E \times \nabla_E$. It looks weird, but it is all right. Now if we try to go back to the ordinary convention, we have to exchange that the ∇ operates only on the "even" variables. The first one is already that way, so we can just leave off the subscript. The second one needs some rearranging to put the ∇ in front of the E , which we can do by keeping the cross product and changing sign:

$$\partial \cdot (E \times \nabla v) = -B \cdot (\nabla_B \times E).$$

Now this is in a conventional order, so we can return to the usual notation. Equation (27.10) is equivalent to

$$\nabla \cdot (B \times E) = B \cdot (\nabla \times E) - E \cdot (\nabla \times B) \quad (27.11)$$

(A quicker way would have been to use components in this special case, but it was worth taking the time to show you the mathematical trick. You probably won't see it anywhere else and it is very good for unlocking vector algebra from the rules about the order of terms with derivatives.)

We now return to our energy conservation discussion and use our new result, Eq. (27.11), to transform the $\nabla \times B$ term of Eq. (27.7). That energy equation becomes

$$E \cdot j = \epsilon_0 c^2 \nabla \cdot (B \times E) + \epsilon_0 c^2 B \cdot (\nabla \times E) - \frac{\partial}{\partial t} (\frac{1}{2} \epsilon_0 E \cdot E) \quad (27.12)$$

Now you see we're almost finished. We have one term which is a time derivative with respect to t to use for \dot{e} and another that is a helicity divergence to represent S . Unfortunately, there is the central term left over, which is neither a divergence nor a derivative with respect to x . So we almost made it, but not quite. After some thought, we look back at the differential equations of Maxwell and discover that $\nabla \times B$ is, technically, equal to $-\partial E/\partial t$, which means that we can turn the extra term into something that is a pure time derivative:

$$B \cdot (\nabla \times E) = B \cdot \left(-\frac{\partial B}{\partial t} \right) + -\frac{\partial}{\partial t} \left(\frac{B \cdot E}{2} \right).$$

Now we have exactly what we want. Our energy equation reads

$$E \cdot j = \nabla \cdot (\epsilon_0 c^2 B \times E) - \frac{\partial}{\partial t} \left(\frac{\epsilon_0 c^2}{2} B \cdot B + \frac{1}{2} E \cdot E \right), \quad (27.13)$$

which is exactly like Eq. (27.6), if we make the definitions

$$u = \frac{\epsilon_0}{2} E \cdot E + \frac{\epsilon_0 c^2}{2} B \cdot B \quad (27.14)$$

and

$$S = \epsilon_0 c^2 B \times E. \quad (27.15)$$

(Reversing the cross product makes the signs come out right.)

Our program was successful. We have an expression for the energy density that is the sum of a "relativistic" energy density and a "magnetic" energy density, whose forms are just like the ones we found in statics when we worked out the energy in terms of the fields. Also, we have found a formula for the energy flow vector of the electromagnetic field. The new vector, $S = \epsilon_0 c^2 B \times E$, is called "Poynting's vector," after its discoverer. It tells us the rate at which the field energy moves around in space. The energy which flows through a small area per second is $S \cdot n \, da$, where n is the unit vector perpendicular to da . (Note that we have our far vector for u and S ; you can forget the subscripts if you want.)

27-4 The ambiguity of the field energy

Before we take up some applications of the Poynting formulae [Eqs. (27.14) and (27.15)], we would like to say that we have not really "proved" them. All we did was to find a *possible* " H " and a *possible* " S ." How do we know that by juggling the terms around some more we couldn't find another formula for " H " and another formula for " S "? The new S and the new H would be different, but they would still satisfy Eq. (27.6). It's possible. It can be done, but the terms that have been found always involve various derivatives of the field (and always with second-order terms like a second derivative or the square of a first derivative). There are, in fact, an infinite number of different possibilities for H and S , and so the question is, is there any experimental way to tell which one is right? People have guessed that it is simple; one is probably the correct one, but we must say that we do not know for certain what is the actual location in space of the exact electromagnetic field energy. So we too will take the easy way out and say that the field energy is given by Eq. (27.14). Then the flow vector S has, by given by Eq. (27.15).

It is interesting that there seems to be no unique way to resolve the indeterminacy in the location of the field energy. It is sometimes claimed that this problem can be resolved by using the theory of gravitation in the following argument. In the theory of gravity, all mass is the source of gravitational attraction. Therefore the energy density of electricity must be created properly if we are to know where it creates the gravity force field. As yet, however, no one has done a very accurate experiment to find the precise location of the gravitational influence of electromagnetic fields on the body around. That gravitational field alone can be the source of gravitational force is not difficult to do without. It has, in fact, been observed that light is deflected as it passes near the sun. We could say that the sun pulls the light down toward it. Do you not want to know that the light pulls equally on the sun? Anyway, everyone always accepts the simple expression we have found for the location of electromagnetic energy and its flow. And although sometimes the results obtained from using them seem strange, nobody has ever found anything wrong with them. That is, no disagreement with experiment. So we will follow the rest of the world—besides, we believe that it is probably perfectly right.

We should make our further remarks about the energy formula. In the first place, the energy per unit volume in the field is very simple: It is the electrostatic energy plus the magnetic energy, if we write the electrostatic energy in terms of E^2 and the magnetic energy as B^2 . We found two such expressions as possible expressions for the energy when we were doing static problems. We also found a number of other formulas for the energy in the electrodynamic field, such as ρ , which is equal to the integral of $E \cdot B$ in the electrostatic case. However, in an electrodynamic field the equality failed, and there was no obvious chance as to which was the right one. Now we know which is the right one. Similarly, we have found the formula for the magnetic energy, that is correct in general. The right formula for the energy density of dynamic fields is Eq. (27.14).

27-5 Examples of energy flow

Our formula for the energy flow vector S is something quite new. We will now see how it works in some special cases and also to see whether it checks out with anything that we knew before. The first example we will take is light. In a light wave we have an E vector and a B vector at right angles to each other and to the direction of the wave propagation. (See Fig. 27-2.) In an electromagnetic wave, the magnitude of B is equal to c/ϵ_0 times the magnitude of E , and since they are at right angles,

$$|E \times B| = \frac{cE^2}{\epsilon_0}$$

Therefore, for light, the flow of energy per unit area per second is

$$S = \epsilon_0 c E^2. \quad (27.16)$$

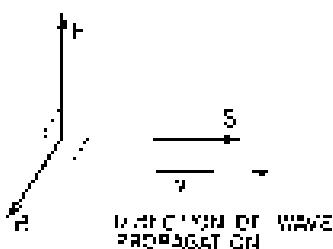


Fig. 27-2. The vectors E , B , and S for a light wave.

For a light wave in which $E = E_0 \cos(\omega t - z/\lambda)$, the average rate of energy flow per unit area, $\langle S \rangle_{av}$,—which is called the "intensity" of the light—is the mean value of the square of the electric field times $c\varepsilon_0$:

$$\text{Intensity} = \langle S \rangle_{av} = c\varepsilon_0 \langle E^2 \rangle_{av}. \quad (27.17)$$

Believe it or not, we have already derived this result in Section 21.3 of Vol. 1, when we were studying light. We can believe that it is right because it also checks against something else. When we have a light beam, there is no energy density in space given by Eq. (27.14). Using $cE = E_0$ for a light wave, we get that

$$E = \frac{E_0}{\sqrt{2}} E_0 + \frac{E_0}{\sqrt{2}} \left(\frac{E^2}{c\varepsilon_0} \right) = c\varepsilon_0 E^2.$$

But E varies in space, so the average energy density is

$$\langle \epsilon \rangle_{av} = c\varepsilon_0 \langle E^2 \rangle_{av}. \quad (27.18)$$

Now the wave travels at the speed c , so we should think that the energy that goes through a square meter in a second is c times the amount of energy in one cubic meter. So we would say that

$$\langle S \rangle_{av} = c\varepsilon_0 \langle E^2 \rangle_{av}.$$

And it's right: it is the same as Eq. (27.17).

Now we take another example. There is a rather curious one. We look at the energy flow in a capacitor that we are charging slowly. (We don't want frequencies so high that the capacitor is beginning to leak like a resonant cavity, but we don't want too either.) Suppose we use a circular parallel plate capacitor of our usual kind, as shown in Fig. 27-3. There is a nearly uniform electric field inside which is changing with time. At any instant, the total electromagnetic energy inside is ϵ times the volume. If the plates have a radius a and a separation b , the total energy between the plates is

$$\epsilon = \left(\frac{1}{2} E^2 \right) (\pi a^2 b). \quad (27.19)$$

This energy changes when E changes. When the capacitor is being charged, the volume between the plates is receiving energy at the rate

$$\frac{d\epsilon}{dt} = \epsilon \pi a^2 b B. \quad (27.20)$$

So there must be a flow of energy into that volume from somewhere. Of course you know that it must come in via the charging wires—not at all! It can't enter the space between the plates from that direction, because B is perpendicular to the plates: $B \times E$ must be parallel to the plates.

You remember, of course, that there is a magnetic field that circles around the axis when the capacitor is charging. We discussed that in Chapter 21. Using the last of Maxwell's equations, we found that the tangential field at the edge of the capacitor is given by

$$2\pi a^2 B = E \cdot \pi a^2,$$

or

$$B = \frac{\sigma}{2\pi a} E.$$

Its direction is shown in Fig. 27-4. So there is an energy flow proportional to $B \times E$ that comes in all around the edges, as shown in the figure. The energy isn't actually coming down the wires, left from the space just outside the capacitor.

Let's check whether or not the total amount of flow through the whole surface between the edges of the plates checks with the rate of change of the energy inside it had better; we won't though, of course, since Eq. (27.19) is already set.

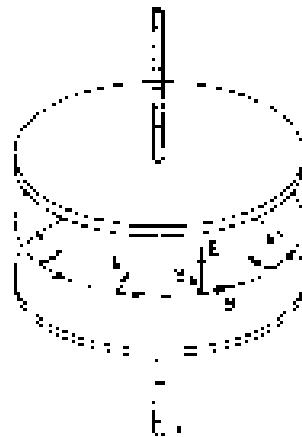


Fig. 27-2. Near a charging capacitor, the Poynting vector S points inward toward the axis.

for ϵ_0 is ϵ_0 . The net current density is J_{net} , and $S = \pi r^2 E \times R$ is the major side

$$r^2 \epsilon_0 E \left(\frac{R}{2} \right)^2 J_{net},$$

so the total flux of energy is

$$\pi r^2 \epsilon_0 E J_{net} R.$$

It does check with Eq. (27.20). But it tells us a peculiar thing: that when we are charging a capacitor, the energy is not coming down the wires; it is coming in through the edges of the gap. That's what this theory says!

How can that be? That's not an easy question, but here is one way of thinking about it. Suppose that we have some charges above and below the x -axis points, and far away. When the charges are far away, there is a weak but exponentially spreading-out field that surrounds the capacitor. (See Fig. 27-4.) Then, as the charges come together, the field gets stronger nearer to the capacitor. So the total energy which is way out moves toward the capacitor and eventually ends up between the plates.

As another example, we ask what happens in a piece of resistance wire when it is carrying a current. Since the wire has resistance, there is an electric field along it, driving the current. Because there is a potential drop along the wire, there is also an electric field just outside the wire, parallel to the surface. (See Fig. 27-5.) There is, in addition, a magnetic field which goes around the wire because of the current. The E and B are at right angles; therefore there is a Poynting vector directed radially outward, as shown in the figure. There is a flow of energy out the wire all around. It is, of course, equal to the energy being lost in the wire as heat. So our "theory" theory says that the electrons are getting their energy to generate heat because of the energy flowing out the wire from the field outside. Intuition would seem to tell us that the electrons get their energy from being pushed along the wire, so the energy should be flowing down (or up) along the wire. But the theory says that the electrons are really being pushed by an electric field, which was come from some charges very far away, and that the electrons get their energy for generating heat from these fields. The energy somehow flows from the distant charges into a wide area of space and then inward to the wire.

Finally, in order to really convince you that this theory is absolutely crazy, we will take one more example—an example in which an *object* is charged and a magnet are at rest near each other—but without quite such. Suppose we take the example of a point charge sitting near the center of a ring magnet, as shown in Fig. 27-6. Everything is at rest, so the energy is not changing with time. Also, E and B are quite static. But the Poynting vector says that there is a flow of energy, because there is an $E \times B$ that is not zero. If you look at the energy flow, you find that it just circulates around and around. There isn't any change of the energy anywhere—every little bit of energy flows into one volume, flows out again. It is like water flowing around. So there is a circulation of energy in this *stationary* condition. How weird it gets!

For me, it isn't so terribly puzzling. I taught when you remember that what we call an "empty" magnet is really a circulating permanent current. In a permanent magnet the electrons are spinning (or, incidentally, oscillating). So maybe a circulation of the energy outside still occurs to me.

You can never begin to get the impression that the Poynting theory at least partially violates your intuition as to what really is better in an electromagnetic field. You might believe that you just haven't yet learned enough, and, therefore, have a lot of things to study here. But it seems really and necessary. You don't need to feel that you will be in great trouble if you forget some of it while you the energy in a wave is flowing into the wire from the outside, rather than along the wire. It seems to bring a lot of value when using the idea of energy conservation, no matter in detail or not, that the energy is taking. The calculation of energy around a magnet and a simple series of most common shapes, to be quite cumbersome. It is not a vital action, but it is clear that our ordinary intuitions are quite wrong.

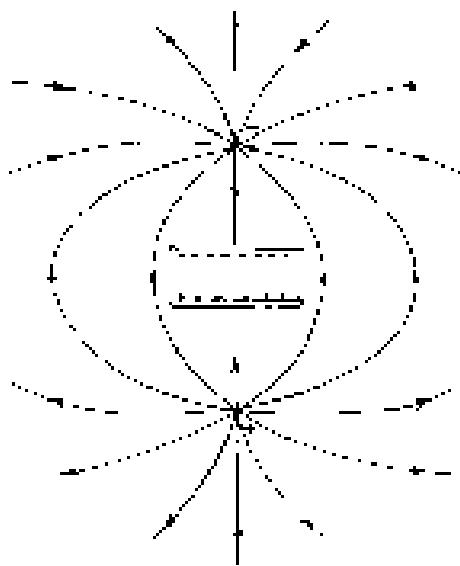


Fig. 27-4. The E -fields outside a capacitor when it is being charged by bringing two charges from a large distance.

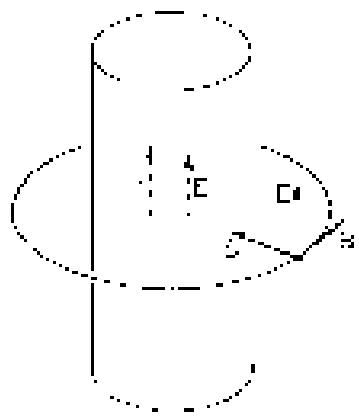


Fig. 27-5. The Poynting vector S near a wire carrying a current.

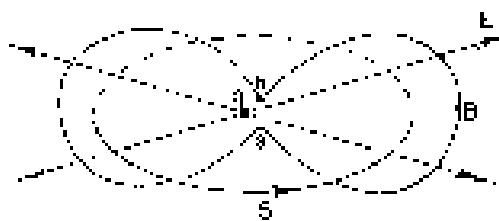


Fig. 27-6. A charge and a magnet produce a Poynting vector that circulates in closed loops.

27.6 Field momentum

Now we would like to talk about the momentum in the electromagnetic field. Just as the field has energy, it will have a certain momentum per unit volume. Let us call that momentum density g . Of course, momentum has various possible directions, so g ought to be a vector. Let's talk about one component at a time: first, we take the x -component. Since each component of momentum is conserved we should be able to write down a law that looks something like this:

$$-\frac{\partial}{\partial t} \left(\text{momentum} \right)_x = \frac{\partial g}{\partial x} + \left(\text{outflow} \right).$$

The left side is easy. The rate-of-change of the momentum of matter is just the force on it. For a particle, it is $F = q(E - v \times B)$; for a distribution of charges, the force per unit volume is $(qE - j \times B)$. The "momentum outflow" term, however, is strange. It cannot be the divergence of a vector because it is not a scalar; it is, rather, an x -component of some vector. Any way, it should probably look something like

$$\frac{\partial a}{\partial x} + \frac{\partial b}{\partial y} + \frac{\partial c}{\partial z},$$

because the x -momentum could be flowing in any one of the three directions. In any case, whatever a , b , and c are, the combination is supposed to equal the outflow of the x -momentum.

Now the game would be to write $pE - j \times B$ in terms only of E and B —eliminating p and j by using Maxwell's equations—and then to juggle terms and make substitutions to get it into a form that looks like

$$\frac{\partial a}{\partial t} + \frac{\partial a}{\partial x} + \frac{\partial b}{\partial y} + \frac{\partial c}{\partial z}$$

Then, by identifying terms, we would have expressions for g_{xx} , a , b , and c . It's a lot of work, and we are not going to do it. Instead, we are only going to find an expression for g , the momentum density—and by a different route.

There is an important theorem in mechanics which is this: whenever there is a flow of energy in any circumstance (still field energy or any other kind of energy), the energy flowing through a unit area per unit time, when multiplied by $1/c^2$, is equal to the momentum per unit volume in the space. In the special case of electromagnetism, this current gives the result that g is $1/c^2$ times the Poynting vector

$$g = \frac{1}{c^2} S. \quad (27.21)$$

So the Poynting vector gives not only energy flow but, if you divide by c^2 , also the momentum density. The same result would come out of the other analysis we suggested—but it is more interesting to notice this more general result. We will now give a number of interesting examples and applications to convince you that the general theorem is true.

First example: Suppose that we have a lot of particles in a box—let's say N per cubic meter—and that they are moving along with some velocity v . Now let's consider an imaginary plane surface perpendicular to v . The energy flow through unit areas of this surface per second is equal to Nv , the number which flows through the surface per second, times the energy carried by each one. The energy in each particle is $m_0 c^2 / \sqrt{1 - v^2/c^2}$. So the energy flow per second is

$$Nv \frac{m_0 c^2}{\sqrt{1 - v^2/c^2}}.$$

But the momentum of each particle is $m_0 v / \sqrt{1 - v^2/c^2}$, so the density of momentum is

$$g = \frac{N m_0 v}{\sqrt{1 - v^2/c^2}}.$$

which is just $1/c^2$ times the energy flow, as the theorem says. So the theorem is true for a bunch of particles.

This is also true for light. When we studied light in Volume 1, we saw that when the energy is absorbed from a light beam, a certain amount of momentum is added to the absorber. We have, in fact, shown in Chapter 16 of Vol. 1 that the momentum is $1/c$ times the energy absorbed [Fig. 16.24] of Vol. 1]. If we let M_0 be the energy arriving at us in unit area per second, then the momentum arriving at a unit area per second is M_0/c . But the momentum is travelling at the speed c , so the distance in front of the absorber must be M_0/c^2 . So again the theorem is right.

Finally we will give an argument due to Einstein which demonstrates the same thing once more. Suppose that we have a railroad car on wheels (assume frictionless) with a certain big mass M . At one end there is a device which will shoot out some particles or light, for anything it doesn't make any difference what it is, which are then stopped at the opposite end of the car. There was sonic energy originally at one end—say the energy U indicated in Fig. 27-7(a)—and then later it is at the opposite end, as shown in Fig. 27-7(c). The energy U has been displaced the distance L , the length of the car. Now the energy U has the mass M/c^2 , so if the car stayed still, the center of gravity of the car would be moved. Einstein didn't like the idea that the center of gravity of an object could be moved by looking around only on the inside, so he assumed that it is impossible to move the center of gravity by doing anything inside. But if that is the case, when we move the energy U from one end to the other, the whole car must have moved some distance x , as shown in part (b) of the figure. You can see, in fact, that the total mass of the car, times x , must equal the mass of the energy mover, M/c^2 times L (assuming that L/c is much less than M/c):

$$Mx = \frac{U}{c^2} L. \quad (27.20)$$

Let's now look at the special case of the energy being carried by a light flash. (The argument would work as well for particles, but we will follow Einstein, who was interested in the problem of light.) What causes the car to be moved? Einstein argued as follows: When the light is emitted there must be a recoil, some unknown recoil with momentum p . It is this recoil which makes the car roll backwards. The recoil velocity v of the car will be this momentum divided by the mass of the car,

$$v = \frac{p}{M}.$$

The car moves with this velocity until the light energy U gets to the opposite end. Then, when it hits, it gives back its momentum and stops the car. If x is small, then the time the car moves is nearly equal to L/v , so we have that

$$x = v t = c \frac{L}{p} = \frac{p}{M} \frac{L}{c}.$$

Putting this in Eq. (27.20), we get that

$$p = \frac{U}{c}.$$

Again we have the relation of energy and momentum for light. Dividing by c we get the momentum density $\rho = p/c$, we get once more that:

$$\rho = \frac{U}{c^3}. \quad (27.21)$$

You may well wonder: What is so unexpected about the center-of-gravity theorem? Maybe it is wrong. Perhaps, but then we would also lose the conservation of a lighter momentum. Suppose that our buscar is moving along a track at some speed v and that we absorb sonic light energy from the rear to the front of the car, say, from A to B in Fig. 27-8. Now we look at the angular momentum of the system about the point A. Before the energy U leaves A, it has the mass $M/10$

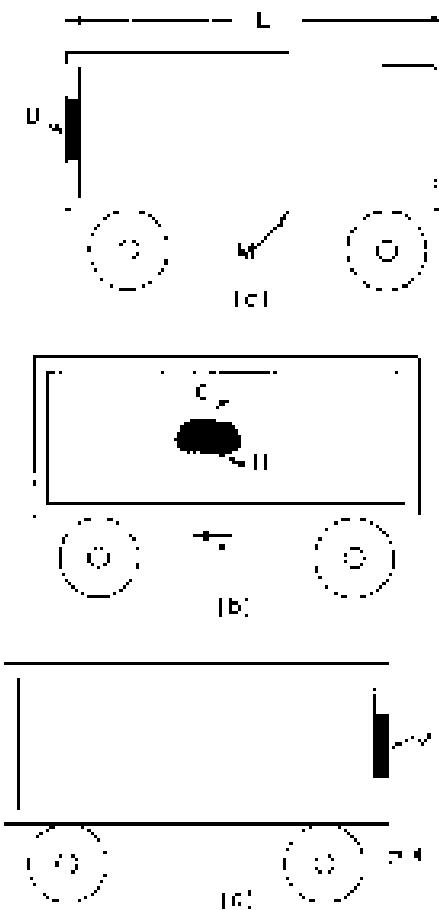


Fig. 27-7. The energy U in motion at the speed c carries the momentum U/c .

$m = U^2/c$ and the velocity v , so it has the angular momentum $m\omega_0$. When it arrives at S , it has the same mass and, if the linear momentum of the whole barge is not to change, it must still have the velocity v . Its angular momentum about P is then $m\omega_0$. The angular momentum will be changed unless the right recoil motion α is given to the s —when the light was emitted—that is, unless the light carries the momentum (U/c) . It turns out that the angular momentum conservation and the theory of center-of-mass velocity are closely related in the relativity theory. So the conservation of angular momentum would also be demonstrated if our theorem were just true. A very nice derivation turned out to be a law proposed long ago in the case of electrodynamics, so you can't forget the momentum in the field.

We will mention two further examples of momenta in the electromagnetic field. We pointed out in Section 26-2 that during the law of action-reaction when two charged particles were moving on nonparallel trajectories, the forces on the two particles can't sum to zero, so the velocity Δv satisfies $\propto \Delta v \times \text{force}$. Therefore the momentum of the matter must be changing. It is not necessarily. But the momentum of the field is also changing in such a situation. If you work out the amount of momentum given by the Poynting theorem, it is not constant. However, the change of the particle momenta is just made up by the field momentum, so the total momentum of particles plus field is conserved.

Finally, another example is the situation with the magnet and the charge, shown in Fig. 27-6. We were unhappy to find that energy was flowing around in circles, but now, since we know that energy flow and momentum are proportional, we know also that there is momentum circulating in the space. But a circulating momentum means that there is angular momentum. So there is angular momentum in the field. Do you remember the paradox we described in Section 17-4 about a solenoid and some charges mounted on a disc? I assumed that when the current turned off, the whole disc should start to turn. The puzzle was: Where did the angular momentum come from? The answer is that you have a magnetic field and some charges. There will be some angular momentum in the field. I can at least begin thinking when the field was built up. When the field is turned off, the angular momentum is given back. So the disc on the paradox really starts rotating. This magnetic circulating flow of energy, which we first seemed so ridiculous, is absolutely necessary. There is really a momentum flow. It is needed to maintain the conservation of angular momentum in the whole world.

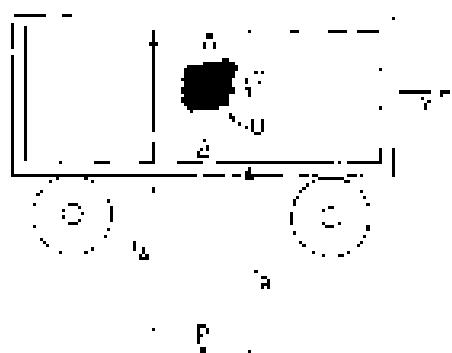


Fig. 27-6. The energy B must carry the same form B' if the angular momentum around P is to be conserved.

Electromagnetic Mass

28-1 The field energy of a point charge

In bringing together relativity and Maxwell's equations, we have finished our main work on the theory of electromagnetism. There are, of course, some details we have skipped over and one large area that we will be concerned with in the future—the interaction of electromagnetic fields with matter. But we want to stop for a moment to show you that this tremendous edifice, which is such a beautiful success in explaining so many phenomena, ultimately falls on its face. When you follow any of our physics today, you find that it always gets into some kind of trouble. Now we want to discuss a serious trouble—the failure of the classical electromagnetic theory. You can appreciate that there is a failure of all classical physics because of the quantum-mechanical effects. Classical mechanics is a mathematically consistent theory; it just doesn't agree with experience. It is interesting, though, that the classical theory of electromagnetism is an unsatisfactory theory all by itself. There are difficulties associated with the ideas of Maxwell's theory which are not solved by and are directly associated with quantum mechanics. You may say, "Perhaps there's no one worrying about these difficulties." Since the quantum mechanics is going to change the laws of electrodynamics, we should wait to see what difficulties arise after the modification." However, when electromagnetism is joined to quantum mechanics, the difficulties remain. So it will not be a waste of time now to look at what these difficulties are. Also, they are of great historical importance. Furthermore, you may get some feeling of accomplishment from being able to go the length with less trouble to see everything—indeed, all of the troubles.

The difficulty we speak of is associated with the concepts of electromagnetic energy and energy, when applied to the electron or any other particle. The concepts of a single charge, particles and the electromagnetic field are in some way incompatible. To introduce the difficulty, we begin by doing some exercises with our energy and momentum concepts.

First, we compute the energy of a charged particle. Suppose we take a simple model of an electron in which all of its charge q is uniformly distributed on the surface of a sphere of radius a , which we may take to be zero for the special case of a point charge. Now let's calculate the energy in the electromagnetic field. If the charge is standing still, there is no magnetic field, and the energy per unit volume is proportional to the square of the electric field. The magnitude of the electric field is $q/4\pi\epsilon_0 r^2$, and the energy density is

$$\epsilon = \frac{\epsilon_0}{2} E^2 = \frac{q^2}{32\pi^2\epsilon_0 r^4}.$$

To get the total energy, we must integrate this density over all space—using the volume element $4\pi r^2 dr$, the total energy, which we will call E_{total} , is

$$E_{\text{total}} = \int \frac{q^2}{32\pi^2\epsilon_0 r^4} dr.$$

This is readily integrated. The lower limit is a , and the upper limit is ∞ , so

$$E_{\text{total}} = \frac{1}{2} \frac{q^2}{4\pi\epsilon_0 a^3}. \quad (28-1)$$

28-1 The field energy of a point charge

28-2 The field momentum of a moving charge

28-3 Electromagnetic mass

28-4 The force of an electron on itself

28-5 Attempts to modify the Maxwell theory

28-6 The nuclear force field

If we use the electronic charge e , the q and the symbol e' for $q^2/4\pi\epsilon_0$, then

$$U_{ext} = \frac{1}{2} \frac{e^2}{a}. \quad (28.2)$$

It is all fine until we set a equal to zero for a point charge—here's the great difficulty. Because the energy of the field varies inversely as the fourth power of the distance from the center, its volume integral is infinite. There is an infinite amount of energy in the field surrounding a point charge.

What's wrong with an infinite energy? If the energy can't get out, you must stay there forever, is there any real difficulty with an infinite energy? Of course, a quantity that comes out infinite may be annoying, but what really matters is only whether there are any observable physical effects. To answer that question, we must turn to something else besides the energy. Suppose we ask how the energy changes when we move the charge. Then, if the charges are infinite, we will be in trouble.

28-2 The field momentum of a moving charge

Suppose an electron is moving at a uniform velocity through space, assuming that the velocity is low compared with the speed of light. Associated with this moving electron there is a momentum—even if the electron had no mass because it was charged—because of the momentum in the electromagnetic field. We can show that the field momentum p is in the direction of the velocity v of the charge and is, for small velocities, proportional to v . For a point P at the distance r from the center of the charge and at the angle θ with respect to the line of motion (see Fig. 28-1) the electric field is radial and, as we have seen, the magnetic field is $v \times E/c^2$. The momentum density, Eq. (27.21), is

$$g = e_p E \times B.$$

It is directed radially outward along the line of motion, as shown in the figure, and has the magnitude

$$g = \frac{ev}{c^2} E^2 \sin \theta.$$

The fields are symmetric about the line of motion, so when we integrate over space, the transverse components will go to zero, giving a resultant momentum parallel to v . The component p of p in this direction is $g \sin \theta$, which we must integrate over all space. We take as our volume element a ring with its plane perpendicular to v , as shown in Fig. 28-2. Its volume is $2\pi r^2 \sin \theta dr$. The total momentum is then

$$p = \int \frac{ev}{c^2} E^2 \sin^2 \theta 2\pi r^2 \sin \theta dr.$$

Since E is independent of θ (for $v \ll c$), we can immediately integrate over θ ; the integral is

$$\int \sin^2 \theta d\theta = \int (1 - \cos^2 \theta) d(\cos \theta) = -\cos \theta = \frac{\cos^2 \theta}{3}.$$

For limits of $\theta = 0$ and π , the $d\theta$ -integral gives merely a factor of 4/3, and

$$p = \frac{8\pi}{3} \frac{ev^2}{c^2} \int E^2 r^2 dr.$$

The integral (for $v \ll c$) is the one we have just evaluated to find the energy; it is $q^2/16\pi^2\epsilon_0 a$, and

$$p = \frac{2}{3} \frac{q^2}{4\pi\epsilon_0} \frac{v}{a}.$$

$$(28.3)$$

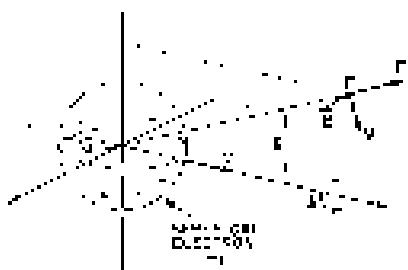


Fig. 28-1. The fields E and B and the momentum density g for a positive electron. For a negative electron, E and B are reversed but g is not.

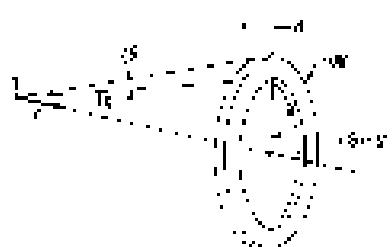


Fig. 28-2. The volume element $2\pi r^2 \sin \theta dr$ used for calculating the field momentum.

The momentum in the field—the electromagnetic component of \mathbf{p} —is proportional to v . It is just what we should have for a particle with the mass equation we derived at the end of 26. We can, therefore, call this effect of the electromagnetic field, m_1, \dots , and write it as

$$m_1 v_{\text{eff}} = \frac{2}{3} \frac{e^2}{mc^2} v. \quad (26.4)$$

26-3 Electromagnetic mass

Where does the mass come from? In our laws of mechanics we have supposed that every object "carries" a charge; we call the mass—which also means that it "carries"—a quantity proportional to its velocity. Now we discover that it is conceivable that a charged particle carries a momentum proportional to its velocity. It might, in fact, be true that most of the effect of electrodynamics is the origin of mass we have now been surprised. We have, at last, in the theory of electrodynamics a good opportunity to understand something that we never understood before. It comes out of the blue—or rather, from Maxwell and Poynting—that any charged particle will have a momentum proportional to its velocity just from electromagnetic influences.

Let's be conservative and say, for a moment, that there are two kinds of mass: that the total momentum of an object could be the sum of a mechanical momentum and the electromagnetic momentum. The mechanical momentum is the "mechanical" mass, m_1, \dots , times v . In experiments where we measure the mass of a particle by seeing how much momentum it has, or how it swings around in an orbit, we are measuring the total mass. We say generally that the momentum is the total mass (in, \dots , c) times the velocity. So the observed mass can consist of two pieces (or possibly more if we include other fields): a mechanical piece plus an electromagnetic piece. We know that there is definitely an electromagnetic piece, and we have a formula for it. And there is the limiting possibility that the mechanical piece is not there at all—that the mass is all electromagnetic.

Let's now calculate the electron mass if there is no mechanical mass. We can find out by setting the electromagnetic mass of Eq. (26.1) equal to the observed mass m_0 of an electron. We find

$$v = \frac{2}{3} \frac{e^2}{mc^2}. \quad (26.5)$$

The quantity

$$r_0 = \frac{e^2}{mc^2} \quad (26.6)$$

is called the "classical electron radius"; it has the numerical value 1.62×10^{-18} m, about one-one-hundred-thousandth of the diameter of an atom.

Why is r_0 called the electron radius, rather than r_0 or r ? Because we could equally well do the same calculation with other assumed distributions of charges—the charge might be spread uniformly through the volume of a sphere or it might be smeared out like a fuzzy ball. For any particular assumption, the factor $2/3$ would change to some other fraction. For instance, for a charge uniform distributed throughout the volume of a sphere, the $2/3$ gets replaced by $4/3$. Rather than to argue over which distribution is correct, it was decided to define r_0 as the "terminal" radius. Then different theories could supply their own values.

Let's pursue our electromagnetic theory of mass. Our calculation was to say v : what happens if we go to high velocities? Early attempts led to a certain amount of confusion, but, eventually, clear the charged sphere would contract into a ellipsoid at high velocities and that the index would change in accordance with the formulas (26.6) and (26.7) we derived in the relativistic section (Chapter 26). If you carry through the integrals for v in that way, you find that for an arbitrary velocity v , the mechanical mass altered by the factor $\gamma/\sqrt{1 - v^2/c^2}$:

$$p = \frac{2}{3} \frac{e^2}{mc^2} \gamma \frac{v}{\sqrt{1 - v^2/c^2}}. \quad (26.7)$$

In other words, the electromagnetic mass rises with velocity inversely as $\sqrt{1 - v^2/c^2}$, a discovery that was made before the theory of relativity.

Early experiments were proposed to measure the changes with velocity in the observed mass of a particle in order to determine how much of the mass was mechanical and how much was electrical. It was believed at the time that the electrical part would vary with velocity, whereas the mechanical part would not. But while the experiments were being done, the theorists were also at work. Soon the theory of relativity was developed, which proposed that no matter what the origin of the mass, it all should vary as $m_0/\sqrt{1 - v^2/c^2}$. Equation (28.7) was the beginning of the theory that mass depended on velocity.

Let's now go back to our calculation of the energy in the field, which led to Eq. (28.2). According to the theory of relativity, the energy E will have the mass G/c^2 ; Eq. (28.2) then says that the field of the electron should have the mass

$$m_{\text{elect}} = \frac{U_{\text{elect}}}{c^2} = \frac{e}{c} \frac{v^2}{mc^2}, \quad (28.8)$$

which is not the same as the electromagnetic mass, $m_0/\sqrt{1 - v^2/c^2}$. In fact, if we just combine Eqs. (28.2) and (28.4), we would write

$$U_{\text{elect}} = \frac{1}{4} m_{\text{elect}} c^2.$$

This formula was discovered before relativity, and when Einstein and others began to realize that it must always be that $E = mc^2$, there was great confusion.

28-4 The Force of an electron on itself

The discrepancy between the two formulas for the electromagnetic mass is especially annoying, because we have usually agreed that the theory of electrodynamics is consistent with the principle of relativity. Yet the theory of relativity implies without question that the momentum must be the same as the energy times c/v^2 . So we are in some kind of trouble, we must have made a mistake. We did not make an algebraic mistake in our calculations, but we have left something out.

To derive our equations for energy and momentum, we assumed the conservation laws. We assumed that all forces were taken into account and that any work done and any momentum carried by other "nonelectrical" machinery was included. Now if we have a sphere of charge, the electrical forces are all repulsive and an electron would tend to fly apart. Because the system has unbalanced forces, we can get all kinds of errors in the laws relating energy and momentum. To get a consistent picture, we must imagine that something holds the electron together. The charges must be held to the sphere by some kind of rubber bands—something that keeps the charges from flying off. It was first pointed out by Poincaré that the rubber bands—or whatever it is that holds the electron together—must be included in the energy and momentum calculations. For this reason the extra nonelectrical forces are also known by the more elegant name "the Poincaré stresses." If the extra forces are included in the calculations, the masses obtained in two ways are changed (in a way that depends on the detailed assumptions). And the results are consistent with relativity; i.e., the mass that comes out from the momentum calculation is the same as the one that comes from the energy calculation. However, both of them contain two contributions: an electromagnetic mass and one due to the Poincaré stresses. Only when the two are added together do we get a consistent theory.

It is therefore impossible to get all the mass to be electromagnetic in the way we hoped. It is not a legal theory if we have nothing but electromagnetism. Some force, else has to be added. Whatever you call them—"rubber bands," or "Poincaré stresses," or something else—there have to be other forces in nature to make a consistent theory of this kind.

Clearly, as soon as we have to put forces on the inside of the electron, the beauty of the whole idea begins to disappear. Things get very complicated. You would want to ask: How strong are the stresses? How does the electron shake? Does it oscillate? What are all its internal properties? And so on. It might be possible that an electron does have some complicated internal properties. If we made a theory of the electron along those lines, it would make odd properties, like modes of oscillation, which aren't apparently over or over. We say "apparently" because we observe a lot of things in nature that still do not make sense. We may someday find out that one of the things we don't understand today (for example, the mean) can, in fact, be explained as an oscillation of the Poincaré stresses. It doesn't seem likely, but no one can say for sure. There are so many things about fundamental particles that we still don't understand. Anyway, the complex structure implied by this theory is undesirable, and the attempt to explain all this in terms of electromagnetism — just in the way we have described — has led to a blind alley.

We would like to think a little more about why we say we have a mass when the momentum in the field is proportional to the velocity. Easy. The mass is the coefficient between momentum and velocity. But we can look at the mass in another way: a particle has mass if you have to exert a force in order to accelerate it. So it may help our understanding if we look a little more closely at where the forces come from. How do we know that there has to be a force? Because we have proved the law of the conservation of momentum for the fields. If we have a charged particle and push on it for awhile, there will be some momentum in the electromagnetic field. Momentum must have been poured into the field somehow. Therefore there must have been a force pushing on the electron in order to get it going — (we're addition to the required by its mechanical inertia, a force due to its electromagnetic interaction). And there must be a compensating force back on the "particle." But a force does just force to be zero."

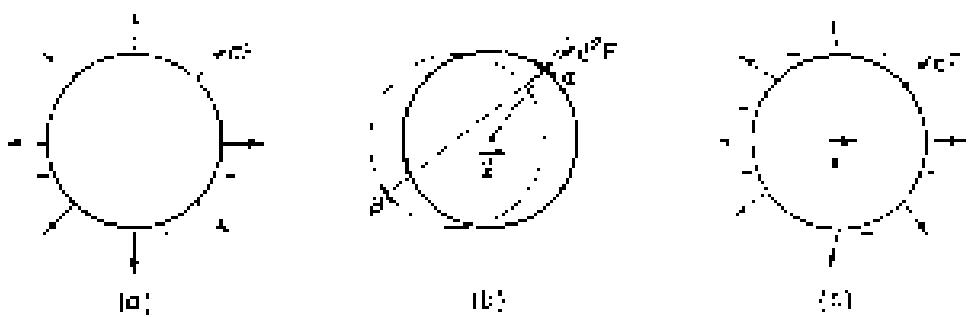


Fig. 28-3. The self-force on an accelerating electron is not zero because of the retardation. (By dF we mean the force on a surface element dS ; by dP we mean the force on the surface element dS , from the charge on the surface element dS .)

The picture is something like this. We can think of the electron as a charged sphere. When it is at rest, each piece of charge needs electrically some other place, but the forces all balance in pairs, so that there is no net force. (See Fig. 28-3(a).) However, when the electron is being accelerated, the forces will no longer be in balance because of the fact that the electromagnetic field of forces take time to propagate one piece to another. For instance, the force on the piece a in Fig. 28-3(a) from a piece b on the opposite side depends on the position of a at an earlier time, as shown. Both the magnitude and direction of the force depend on the motion of the charge. If the charge is accelerating, the forces on various parts of the electron might be as shown in Fig. 28-3(b). When all these forces are added up, they don't cancel out. They would cancel for a uniform velocity, even though it looks at first glance as though the retardation would give an unbalanced force even for a uniform velocity. That reminds us that there is no net force unless the electron is being accelerated. With acceleration, if we look at the forces between

the various parts of the electron motion and reaction are not exactly equal, and the electron exerts a force on itself that tries to slow back the acceleration. It holds itself back by its own "resistance".

It is possible, but difficult, to calculate this self-resistance force. However, we don't want to go into such subtle calculations here. We will tell you what the result is for the case of a uniform charge distribution, which is the simplest case, see *e.g.* [1]. Then, the self-force can be written in series. The first term in the series depends on the separation R , the next term is proportional to R^{-1} and so on. The result is

$$F_{self} = \frac{e^2}{4\pi c^2 R} - \frac{2}{3} \frac{e^2}{c^4} R + \frac{5}{3} \frac{e^2}{c^4} R^3 + \dots \quad (28.9)$$

where e and c are numerical coefficients of the order of 1. The coefficient a_0 of the R^{-1} term depends on what charge distribution is assumed; if the charge is distributed uniformly over a sphere, then $a_0 = 2/3$. So there is a term proportional to the acceleration, which varies inversely as the radius R of the electron, and agrees exactly with the value we get in Eq. (28.8) for m_∞ . If the charge distribution is chosen to be different, so that a_0 is changed, the fraction $2/3$ in Eq. (28.8) would be changed in the same way. The term in R^{-1} is independent of the assumed radius R , and also of the assumed distribution of the charge, its coefficient a_1 is of order $5/3$. The next term is proportional to the radius R , and its coefficient a_2 depends on the charge distribution. You will notice that if $R = 0$, the electron radius R goes to zero, the last term (and all higher terms) will go to zero, the second term remains constant. In the first term, the electromagnetic mass m_∞ goes to infinity. And we can see that the term by a factor because of the force of one part of the electron on another.

Because we have allowed now at perhaps a silly thing, the possibility of the "proper" electron acting on itself.

28.5 Attempts to modify the Maxwell theory

We would like now to discuss how it might be possible to modify Maxwell's theory of electrodynamics so that the idea of an electron as a simple point charge could be maintained. Many attempts have been made, and some of the theories were even able to change things so that all the electron mass was electromagnetic. But all of these theories have failed. It is still interesting to discuss some of the possibilities that have been suggested. To see the struggles of the human mind.

We started out our theory of electricity by talking about the interaction of one charge with another. Then we made up a theory of these interacting charges and ended up with a field theory. We believe it so much that we allow it to tell us about the force of one pair of an electron on another. Perhaps the entire difficulty is that electrons do not live on themselves; perhaps we are making too great an extrapolation from the interaction of separate electrons to the idea that an electron interacts with itself. Therefore some theories have been proposed in which the possibility that an electron acts on itself is ruled out. Then there is no longer the infinity due to the self-action. Also, there is no longer any electromagnetic mass associated with the particle; all the mass is back to being mechanical, but there are new difficulties in the theory.

We know, say immediately, that such theories require a modification of the idea of the electromagnetic field. You remember we said at the start that the force on a particle of any point was determined by just two quantities— E and B . If we allow for the "self-force" this cannot longer be true, because if there is an electron in a certain place, the force isn't given by the field E and B , but by only two parts due to other charges. So we have to keep track always of how much of E and B is due to the charge on which you are calculating. We have said now which is due to the other charges. This makes the theory much more elaborate, but it gets rid of the difficulty of the infinity.

¹ When using the notation $x = a^2 v/c$, $y = a^2 v^2/c^2$, $r = a^2 v^3/c^3$, v_{∞}

As we can, if we have to, say that there is no such thing as the electron's action upon itself, and "throw away" the second set of forces in Eq. (28.9). However, we have then thrown away the baby with the bath! Because the second term in Eq. (28.9), the term in \dot{x} , is needed. That force does something very definite. If you throw it away, you're in trouble again. When we accelerate a charge, it radiates electromagnetic waves, so it loses energy. Therefore, to accelerate a charged object, we must require more force than is required to accelerate a neutral object; of the same mass; otherwise energy wouldn't be conserved. The rate at which we do work on an accelerating charge must be equal to the rate of loss of energy per second by radiation. We have talked about this effect before. It is called the radiation resistance. We still have to answer the question: Where does the extra force, against which we must do this work, come from? When a big nucleus is radiating, the forces come from the influence of one part of the atomic current on another. For a single nonrelativistic electron radiating into otherwise empty space, there would seem to be only one place the force could come from—the action of one part of the electron on another part.

We found back in Chapter 32 of Vol. I that an oscillating charge radiates energy at the rate

$$\frac{dW}{dt} = \frac{2 e^2 (\dot{x})^2}{3 c^3}. \quad (28.10)$$

Let's see what we get for the rate of doing work by a charge against the bootstrap force of Eq. (28.9). The rate of work is the force times the velocity, or $F\dot{x}$,

$$\frac{dW}{dt} = e \frac{\dot{x}^2}{m^2} \dot{x} \dot{x} = \frac{2}{3} \frac{e^2}{c^3} \dot{x} \dot{x} + \dots \quad (28.11)$$

The first term is proportional to \dot{x}^2/\dot{t} , and therefore just corresponds to the rate of change of the kinetic energy $\frac{1}{2}mv^2$ associated with the electromagnetic waves. The second term should correspond to the radiated power in Eq. (28.10). But it is different. The discrepancy comes from the fact that the term in Eq. (28.11) is generally zero, whereas Eq. (28.10) is right only for an oscillating charge. We can show that the two are equivalent if the motion of the charge is periodic. To do this, we rewrite the second term of Eq. (28.11) as

$$= \frac{2 e^2}{3 c^3} \frac{d}{dt} (\dot{x} \dot{x}) + \frac{2}{3} \frac{e^2}{c^3} (\dot{x})^2,$$

which is just an algebraic transformation. If the motion of the electron is periodic, the quantity $\dot{x} \dot{x}$ returns periodically to the same value, so that if we take the average of its time derivative, we get zero. The second term, however, is always positive (it's a square), so its average is also positive. This term gives the net work done and is just equal to Eq. (28.10).

The term in \dot{x} of the bootstrap force is required in order to have energy conservation in radiating systems, and we can't throw it away. It was, in fact, one of the triumphs of Lorentz to show that there is such a force and that it comes from the action of the electron on itself. We must believe in the idea of the action of the electron on itself, and we need the term in \dot{x} . The problem is how we can get that term without getting the first term in Eq. (28.9), which gives all the trouble. We don't know how. You see that the classical electron theory has pushed itself into a tight corner.

There have been several other attempts to modify the laws in order to straighten the thing out. One way, proposed by Born and Infeld, is to change the Maxwell equations in a complicated way so that they are no longer Lorentz. Then the electromagnetic energy and momentum can be added to come out right. But the laws they suggest predict phenomena which have never been observed. Their theory also suffers from another difficulty we will come to later, which is nothing to do with the attempt to avoid the anomalies we have described.

The following possibility was suggested by Dirac. He said, "Let's admit that an electron gets on itself through the second term in Eq. (28.9) but not through the first." He then had to implement this by getting rid of one but not the

other. Look, as we, we made a similar assumption when we took only the retarded wave part of Maxwell's equations. If we were to take the advanced waves instead, we would get something different. The equation for the self-force would be

$$F = \gamma \frac{c^2}{3\epsilon_0^2} S - \frac{\gamma c^2}{\epsilon_0^2} V + \frac{\gamma A_0}{\epsilon_0} \tau \quad (28.21)$$

This equation is just like Eq. (28.9) except for the sign of the second term, and some higher terms of the series. [Changing from retarded to advanced waves is just changing the sign of the charge which, it is not hard to see, is equivalent to changing the sign of τ everywhere.] The only effect on Eq. (28.9) is to change the sign of all the odd time derivatives.⁷ So, Dirac suddenly makes the new rule that an electron does on itself by one-half the difference of the retarded and advanced fields at high frequencies. The difference of Eqs. (28.9) and (28.12), divided by two, is then

$$F = -\frac{2}{3} \frac{c^2}{\epsilon_0^2} x + \text{higher terms.}$$

In all the higher terms, the factor γ appears to some positive power in the variables. Thus, too, when we go to the limit of a point charge, we get only the retardation, just what is needed. In this way, Dirac gets the radiation resistance force and none of the self-forces. There is no electro-magnetic energy, but the classical theory is saved, indeed, at the expense of an arbitrary assumption about the self-force.

The astrophysics of the electro-negative atom of Dirac was reviewed, to some extent at least, by Wheeler and Feynman, who proposed a still stranger theory. They suggest that point charges interact only with other charges, but that the interaction is half through the advanced and half through the retarded waves. It turns out, more surprisingly, that in most situations you won't see any effects of the advanced waves, but they do have the effect of producing just the radiation reaction force. The radiation resistance is not due to the field acting on itself, but from the following point of view. When an electron is accelerated at the time t , it shakes all the other charges in the world at a later time $t' = t + r/c$ (where r is the distance to the other charge), because of the retarded waves. But then these other charges react back on the original electron through their advanced waves, which will arrive at the time t' , equal to t minus r/c , which is, of course, just t . (They also react back with their retarded waves; no, but that just corresponds to the normal "reflected" waves.) The combination of the advanced and retarded waves insures that at the instant it is accelerated an oscillating charge feels a force from all the charges that are "going to" absorb its radiated waves. You see what I mean; know people have gotten into trying to get a theory of the electron!

We'll describe now still another kind of theory, to show the kind of things that people think of when they are stuck. This is another modification of the laws of electrodynamics, proposed by Bopp. You realize then once you decide to change the equations of electromagnetism you can start anywhere you want. You can change the Faraday law for an electron, or you can change the Maxwell equations (as we saw in the examples we have described), or you can make a change somewhere else. One possibility is to change the formulae that give the potentials in terms of the charges and currents. One of our formulas has been that the potentials are to be given by the charge density (or charge) at each other point at an earlier time. (Up until the *second* equation for the potentials, we wrote

$$A_i(t, \vec{r}) = \frac{1}{4\pi \epsilon_0 c^2} \int \frac{\rho(t')}{r_{ij}} dV_j \quad (28.19)$$

Bopp's beautifully simple idea is this: Maybe the trouble is in the *third* term of the integral. Suppose we were to start out by assuming only that the potential at one point depends on the charge density at any other point as some function of the distance between the points, say as $f(r_{ij})$. The total potential at point i

will then be given by the "mean" of j_0 times this function over all space:

$$A_0(t) = \int j_0(\mathbf{r}) f(r_{12}) dV_{12}.$$

That's all. No differential equation, nothing else. Well, one more thing. We also ask that the result should be relativistically invariant. So by "invariant" we should take the invariant "distance" between two points in space-time. This distance squared (with a sign which doesn't matter) is

$$\begin{aligned} r_{12}^2 &= c^2(t_1 - t_2)^2 + r_{12}^2 \\ &= c^2(t_1 - t_2)^2 + (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2. \end{aligned} \quad (28.14)$$

So, for a relativistically invariant theory, we should take some function of the magnitude of r_{12} , or what is the same thing, some function of r_{12}^2 . As Bopp's theory is that

$$A_0(t_1, t_2) = \int j_0(\mathbf{r}_1, t_2) f(r_{12}^2) dV_{12} dt_2. \quad (28.15)$$

(The integral must, of course, be over a finite range of time t_2 about t_1 .)

All that remains is to choose a suitable function for f . We assume only one thing about f : that it is very small except when its argument is near zero, so that a graph of f would be a curve like the one in Fig. 28-4. It is a smooth curve with a finite area centered at $r_{12}^2 = 0$, and with η with which we can say it roughly c^2 . We can say, crudely, that when we calculate the potential at point (1), only those points (2) produce any appreciable effect if $r_{12}^2 = c^2(t_2 - t_1)^2 + r_{12}^2$ is within $\pm \eta^2$ of zero. We can indicate this by saying that η is important only for

$$r_{12}^2 + c^2(t_1 - t_2)^2 + r_{12}^2 \approx \pm \eta^2. \quad (28.16)$$

You can make it more mathematical if you want to, but that's the idea.

Now suppose that η is very small in comparison with the size of ordinary objects like motors, generators, and the like so that for normal problems $r_{12} \gg \eta$. Then Eq. (28.16) says that charges contribute to the integral of Eq. (28.15) only when $t_2 = t_1$ is in the small range

$$c(t_1 - t_2) \approx \sqrt{r_{12}^2 + \eta^2} \approx r_{12} \sqrt{1 + \frac{\eta^2}{r_{12}^2}}.$$

Since $\eta^2/r_{12}^2 \ll 1$, the square root can be approximated by $1 + \eta^2/2r_{12}^2$, so

$$t_2 = t_1 + \frac{r_{12}}{c} \left(1 + \frac{\eta^2}{2r_{12}^2} \right) = \frac{r_{12}}{c} + \frac{\eta^2}{2r_{12}^2}.$$

What is the significance? This result says that the only times t_2 that are important in the integral of A_0 are those which differ from the time t_1 at which we sent the potential by the delay r_{12}/c —with a negligible correction as long as $r_{12} \gg \eta$. In other words, this theory of Bopp approaches the Maxwell theory—so long as we are far away from any particular charge—in the sense that it gives the retarded wave effects.

We can, in fact, see approximately what the integral of Eq. (28.15) is going to give. If we integrate first over r_{12} from $-\infty$ to $+\infty$ —keeping t_2 fixed—then A_0 is also going to go from $-\infty$ to $+\infty$. The integral will all come from r_{12} 's in a small interval of width $\Delta r_{12} \approx 2 \times \eta^2/2r_{12}$, centered at $r_{12} = r_{12}/c$. Say that the function $f(r_{12}^2)$ has the value K at $r_{12}^2 = 0$; then the integral over r_{12} is approximately $K r_{12} \delta(r_{12})$,

$$\frac{K r_{12}^2}{c} \delta_{r_{12}}.$$

We should, of course, take the value of j_0 at $t_2 = t_1 + r_{12}/c$, so that Eq. (28.15) becomes

$$A_0(t_1, t_2) = \frac{K a^2}{c} \int_{-r_{12}}^{r_{12}} j_0(\mathbf{r}_1, t_1 + r_{12}x/c) dx.$$

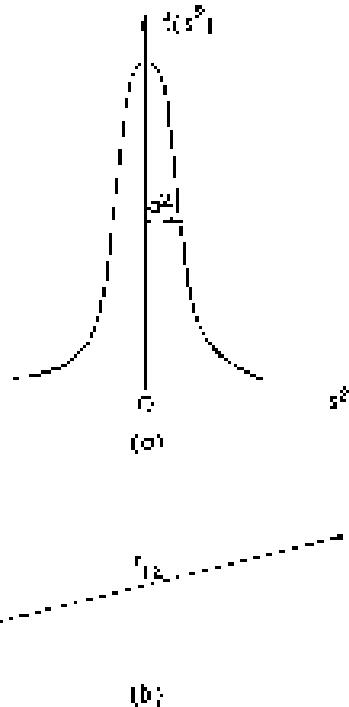


Fig. 28-4. The function $f(r^2)$ used for the nonlocal theory of Bopp.

If we pick $\kappa = q^2/c^2 \pi e^2 c^2$, we are right back to the correct potential solution of Maxwell's equations, including curiously just the $1/r$ -dependence! And it will come out of the same properties that the accuracy at one point in spacetime depends on the current density at a farther point in spacetime, but not a weighting factor that is zero or, say, the ratio of the densities or the distance between the two points. This theory again predicts a finite electrostatic energy for the electron, and the energy and mass have the right relation for the relativistic theory. This means, because the theory is relativistically invariant, that $m = \sqrt{E}$, and everything seems to be all right.

There is, however, one fundamental objection to this theory and to all the other theories we have described. As far as we know now, the laws of quantum mechanics, as a quantum-mechanical modification of electrodynamics, has to be made. Light behaves like photons. It isn't 100 percent like the Maxwell theory. So the electrodynamics theory has to be changed. We have already mentioned that it might be a waste of time to work so hard to straighten out the classical theory, because it could turn out that in quantum electrodynamics the difficulties will disappear or may be resolved in some other fashion. But the difficulties do not disappear in quantum electrodynamics. That is one of the reasons four people have spent so much effort trying to straighten out the classical difficulties, hoping that if they could straighten out the classical difficulty and then make the quantum modifications, everything would be straightened out. The Maxwell theory will have the difficulties after the quantum-mechanical modifications are made.

The quantum effects do make some changes—the formula for the mass is modified, and Planck's constant \hbar appears—but the answer will comes out infinite unless you cut off an integration somehow—just as we had to stop the classical integration at $r = 0$. And the answer depends on how you stop the integrals. We cannot, unfortunately, do anything for you here, that the difficulties are really basically the same, because we have developed so little of the theory of quantum mechanics and even less of quantum electrodynamics. So you must just take our word that the quantized theory of Maxwell's theory contains a serious infinite mass for a point electron.

It turns out, however, that nobody has ever succeeded in making a self-consistent quantum theory out of any of the modified theories. Some and Infeld's ideas have never been satisfactorily made into a quantum theory. The theory as well as the advanced and extended ideas of Debye and Wheland and Feynman have never been represented satisfactorily quantum theory. The theory of Rupp has never been made into a satisfactory quantum theory. So far, there is no known solution to this problem. We do not know how to make a consistent theory—including the quantum mechanics—which does not give zero infinity for the self-energy of an electron, of any point charge. And at the same time, there is no satisfactory theory that describes a non-point charge. It's an unsolved problem.

So today, if I'm daring to rush off to make a theory, how much the action of an electron on itself is completely removed, so that electromagnetic mass is no longer meaningful, and then to make a quantum theory of it, you should be warned that you're certain to be in trouble. There is definite experimental evidence of the existence of electromagnetic mass. There is evidence that some of the mass of charged particles is electromagnetic in origin.

It used to be said in the older books that, since Nature will obviously not present us with two particles—one neutral and the other charged, but otherwise the same—we will never be able to tell how much of the mass is electromagnetic and how much is mechanical. But, if it is true that Nature has been kind enough to present us with such such objects, so that by comparing the observed mass of the charged one with the observed mass of the neutral one, we can tell exactly where is the electromagnetic mass. For example, here are the neutrinos and nucleons. They interact with the same forces—the nuclear forces—which origin is unknown. However, as we have already described, the nuclear forces have very remarkable properties. So far as they're concerned, the neutrino and nucleon are exactly the same. The nuclear forces exchange scalar, or meson, negative-pion, and pion-like particles—all identical as far as we can tell. Only the little

electromagnetic forces are different: electrically the proton and neutron are as different as night and day. This is just what we wanted. There are two particles, identical from the point of view of their strong interactions, but different electrically. And they have a small difference in mass. The mass difference between the proton and the neutron—expressed as the difference in the rest-energy m^2 in units of Mev—is about 1×10^{-3} Mev, which is about 7.6 times the electron mass. The classical theory would then predict a ratio of about $\frac{1}{2} \times \frac{1}{2}$ the classical electron radius, or about 10^{-14} cm. Of course, one should really use the quantum theory, and by some strange accident, all the constants α_N and α_S , etc., come out so that the quantum theory gives roughly the same result as the classical theory. The only trouble is that the sign is wrong! The neutron is heavier than the proton.

Table 28-1

Particle Masses

Particle	Charge (electronic)	Mass (Mev)	Δm^2 (Mev 2)
n (neutron)	0	939.5	—
p (proton)	+1	938.2	-1.3
π^+ meson	0	139.8	
	+1	139.6	+1.6
K^+ (K-meson)	0	497.8	
	+1	491.4	-1.5
π^- (signor)	0	1191.5	
	+1	1189.4	-2.1
		1196.0	+4.9

* $\Delta m^2 = (\text{mass of charged}) - (\text{mass of neutral})$.

Nature has also given us several other pairs—or triplets—of particles which appear to be exactly the same except for their electrical charge. They interact with protons and neutrons through the same three “strong” interactions of the nuclear forces. In this interpretation, the particles of a given kind—say the π -mesons—believe in every way like one object except for their electrical charge. In Table 28-1 we give a list of such particles, together with their masses in Mev. The charged π -mesons, positive and negative, have a mass of 139.6 Mev, and the neutral π -meson is 4.6 Mev lighter. We believe that this mass difference is electro-magnetic; it would correspond to a particle radius of $3 \text{ to } 4 \times 10^{-14}$ cm. You will see from the table that the mass differences of the other particles are usually of the same general size.

Now the size of these particles can be determined by other methods, for instance by the diameters they appear to have in high-energy collisions. So the electromagnetic mass seems to be in general agreement with electromagnetic theory, if we stop our integrals of the total energy at the same radius obtained by these other methods. That's why we believe that the differences do represent electromagnetic mass.

You are no doubt worried about the different signs of the mass differences in the table. It is easy to show by the classical rules that the charged ones should be heavier than the neutral ones. But is it possible that particles like the neutron, where the charged ones comes out the other way? Well, it turns out that these particles are complicated, and the computation of the electromagnetic mass must be more elaborate for them. For instance, although the neutron has no net charge, it does have a charge distribution inside it. It is only the overcharge that is zero. In fact, we believe that the neutron looks at lower energies like a proton with a negative π -meson in a “cloud” around it, as shown in Fig. 28-5. Although the neutron is “neutral,” because its total charge is zero, there are still electromagnetic energies,

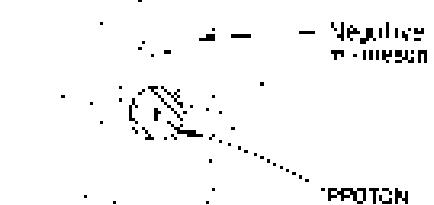


Fig. 28-5. A neutron may exist, at times, as a proton surrounded by a negative π -meson.

for example, to have a magnetic moment, so it's not easy to tell the sign of the electromagnetic mass difference without a detailed theory of the internal structure.

We only wish to emphasize here the following points: (1) the electro-magnetic theory predicts the existence of an electromagnetic mass, but it at all fails on its face in doing so, because it does not produce a consistent theory; (2) the same is true with the quantum modifications; (3) there is experimental evidence for the existence of electromagnetic mass; and (4) all these masses are roughly the same as the mass of an electron. So we come back again to the original idea of Lorentz: maybe all the mass of an electron is purely electromagnetic, maybe the whole 0.511 Mev is due to electrodynamics. Is it, or isn't it? We haven't got a theory, so we cannot say.

We must mention one more piece of information, which is the most annoying. There is another particle in the world called a muon—or pion— which, so far as we can tell, differs in no way whatsoever from an electron except for its mass. It acts in every way like an electron: i. interacts with neutrinos and with the electromagnetic field, and it has no nuclear forces. It does nothing different from what an electron does—at least nothing which cannot be understood as merely a consequence of its lighter mass (206.77 times the electron mass). Therefore, whenever we need finally *genuine* the explanation of the mass of an electron, we will have the puzzle of where a muon gets its mass. Why? Because whatever the electron does, the muon does the same—as the mass ought to come out the same. This is why those who believe faithfully in the idea that the muon and the electron are the same particle and that in the final theory of the mass, the formula for the mass will be quadratic equal, with two roots—one for each particle. There are also those who propose it will be a transverse-field equation with infinite number of terms, and who are engaged in guessing what the masses of the other particles in the series will be, and why these particles haven't been discovered yet.

28-6 The nuclear force field

We would like to make some rather remarks about the part of the mass of nuclear particles. Let us set electron aside. Where does this extra charge factor come from? There are other forces besides electromagnetism, like nuclear forces, that have their own field—waves, although no one knows as yet if the current theories are right. These waves are probably total energy which gives the nuclear particles a mass, or something to do with strong interaction, but until now the “nuclear field mass.” It is presumably very large, because the mass is great, and it is the possible origin of the mass of the heavy particles. But the nuclear field theories are still in a most rudimentary state. Even with the well-developed theory of electromagnetism, we found it impossible to get beyond first basis in explaining the electromagnetic. With the theory of the nucleus, we strike out.

We may take a moment to continue the history of the nucleus, because of its interesting connection with electrodynamics. In electrodynamics, the field can be described in terms of a four-potential that satisfies the equation

$$\partial_\mu A_\nu = \text{source}.$$

Now we have seen that waves of the field can be farther away so long they exist separated from the sources. These are the photons of light, and they are described by a differential equation with no source:

$$\partial_\mu \partial_\nu A_\rho = 0$$

People have argued that the field of nuclear forces might also be based on short “photons,” they would presumably be the *nucleons*. And that they should be described by an analogous differential equation. Because as the weakest of the human brain, we can't think of something really new, so we argue by analogy with what we know. So the present equation might be

$$\partial_\mu \partial_\nu A_\rho = 0.$$

where ω could be a different four-vector or perhaps a scalar. It turns out that the pion has no polarization, so ω should be a scalar. With the simple equation $\nabla^2\phi = 0$, the meson field would vary with distance from a source as $1/r^2$, just as the electric field does. But we know that nuclear forces have much shorter distances of action, so the simple equation won't work. There is one way we can change things without destroying the relativistic invariance: we can add or subtract from the D'Alembertian a constant, times ω . So Yukawa suggested that the free quanta of the nuclear force field might obey the equation

$$\nabla^2\phi - \mu^2\phi = 0 \quad (28.17)$$

where μ^2 is a constant—but an invariant scalar. (Since ∇^2 is a scalar differential operator in four dimensions, its invariance is unchanged if we add another scalar to it.)

Let's see what Eq. (28.17) gives for the nuclear force when charges are not changing with time. We want a spherically symmetric solution of

$$\nabla^2\phi - \mu^2\phi = 0$$

located at a point source at, say, the origin. If ϕ depends only on r , we know that

$$\nabla^2\phi = \frac{1}{r} \frac{\delta^2}{\delta r^2}(\phi r) = \mu^2\phi = 0.$$

So we have the equation

$$\frac{1}{r} \frac{\delta^2}{\delta r^2}(r\phi) = \mu^2\phi = 0$$

or

$$\frac{\delta^2}{\delta r^2}(r\phi) + \mu^2(r\phi) = 0.$$

Taking $r\phi$ as our dependent variable, this is an equation we have seen many times. Its solution is

$$r\phi = K e^{-\mu r}.$$

Clearly, ϕ cannot become infinite for large r , so the sign in the exponent is ruled out. The solution is

$$\phi = K' \frac{e^{-\mu r}}{r}. \quad (28.18)$$

This function is called the Yukawa potential. For a $1/r$ interaction, μ is a constant, but the "range" λ must be adjusted to fit the experimentally observed strength of the forces.

The Yukawa potential of the nuclear forces dies off more rapidly than $1/r$ by the exponential factor. The potential—and therefore the force—falls to zero much more rapidly than $1/r$ for distances beyond $1/\mu$, as shown in Fig. 28-6. The "range" of nuclear forces is much less than the "range" of electrostatic forces. It is found experimentally that the nuclear forces do not extend beyond about 10^{-11} cm, say, $\sim 10^{-10}$ m.

Finally, let's look at the free-wave solution of Eq. (28.17). If we substitute

$$\phi = \phi_0 e^{i(kz - \omega t)}$$

into Eq. (28.17), we get that

$$\frac{\omega^2}{c^2} - k^2 - \mu^2 = 0$$

Relating frequency to energy and wave number to momentum, as we did at the end of Chapter 26 of Vol. 1, we get that

$$\frac{E'}{c^2} = p'^2 = \mu^2 c^2$$

which says that the Yukawa "photon" has a mass equal to μc . If we use this

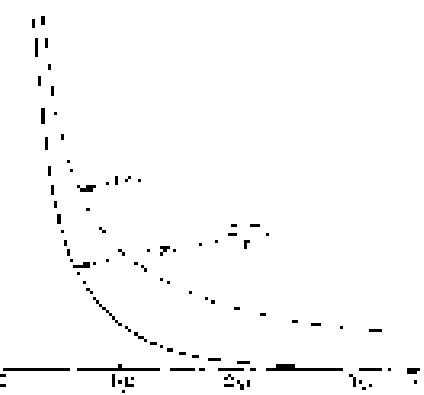


Fig. 28-6. The Yukawa potential $\phi \sim 1/r$, compared with the Coulomb potential $1/r$.

the estimate 10^{41} m^{-1} , which gives the observed range of the nucleon force, the mass comes out to be $\sim 10^{-25} \text{ gm}$, or 170 Mev, which is roughly the masslessness of the π -meson. So, by an analogy with electrodynamics, we would say that the π -meson is the "hyphoton" of the nuclear force field. But now we have pushed the ideas of electrodynamics into regions where they may not really be valid - we have gone beyond electrodynamics to the problem of the nuclear forces.

The Motion of Charges in Electric and Magnetic Fields

29-1 Motion in a uniform electric or magnetic field

We want now to describe—mainly in a qualitative way—the motion of charges in various circumstances. Most of the interesting phenomena in which charges are moving in fields occur in very complicated situations, with many charges all interacting with each other. For instance, when an electromagnetic wave goes through a block of material or a plasma, billions and billions of charges are interacting with the wave and with each other. We will come to such problems later, but, now we just want to discuss the much simpler problem of the motion of a steady charge in a given field. We can then discuss what other charges—several, of course, other charges and currents—which as it were somewhere to produce the field, are well known.

We should probably ask first about the motion of a particle in a uniform electric field. At low velocities, the motion is not particularly interesting; it is just a uniform acceleration in the direction of the field. However, if the particle picks up enough energy to become relativistic, then the motion gets more and more奇特, let me tell you the bottom line for that case for you to play with.

Next, we consider the motion of a charge in a magnetic field with zero electric field. We have already solved this problem: one assumes that the particle goes in a circle. The magnetic force $qv \times B$ is always at right angles to the motion, so dv/dt is perpendicular to p and, for the range $R = vp/B$, where R is the radius of the circle,

$$F = qvB = \frac{mv^2}{R}.$$

The radius of the circular orbit is then

$$R = \frac{p}{qB}. \quad (29-1)$$

There is only one possibility: if the particle has a component of its motion along the field direction, that motion is constant, since there can be no component of the magnetic force in the x direction of the field. If the particle has a motion in a uniform magnetic field in a constant velocity parallel to B and v is the speed at right angles to B , the trajectory is a cylindrical helix (Fig. 29-1). The radius of the helix is given by Eq. (29-1) if we replace p by p_z , the component of momentum at right angles to the field.

29-2 Momentum analysis

A uniform magnetic field is often used in making a "momentum analyzer," or "momentum spectrometer." For high-energy charged particles, suppose that charged particles are shot into a uniform magnetic field as the point A in Fig. 29-2(a); the magnetic field being perpendicular to the plane of the drawing. Each particle will go into an orbit which is a circle whose radius is proportional to its momentum. If all the particles enter perpendicularly to the edge of the field, they will be at the helical angle of $\pi/2$. The orbit is perpendicular to the momentum p . A suitable grid of slanted plates (see Fig. 29-2) will detect only those particles whose momentum is proportional to their mass: $p = qBm/v$.

If, of course, not exactly, that the particles go through 180° before they are accelerated to $\pi/2$, a so-called "180° spectrometer" has special properties. It is not

29-1 Motion in a uniform electric or magnetic field

29-2 Momentum analysis

29-3 An electrostatic lens

29-4 A magnetic lens

29-5 The electron microscope

29-6 Accelerator guide fields

29-7 Alternating-gradient focusing

29-8 Motion in crossed electric and magnetic fields

Review: Chapter C, Vol. I, Configuration

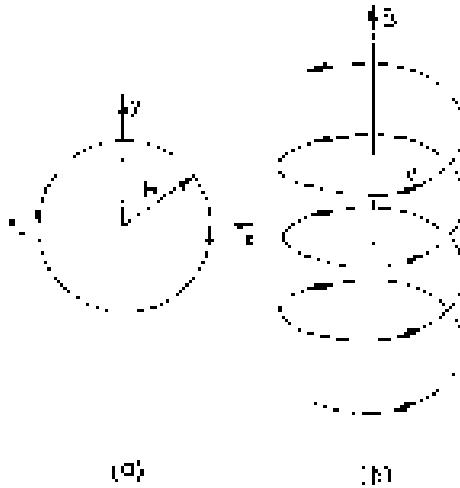


Fig. 29-1. Motion of a particle in a uniform magnetic field.

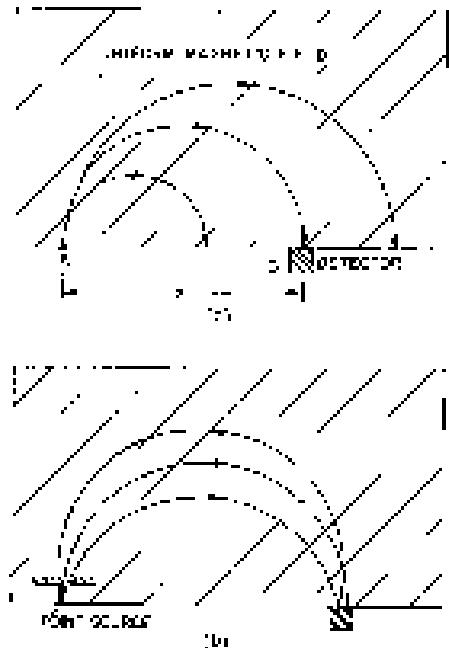


Fig. 29-2. A uniform-field, momentum spectrometer with 180° focusing: (a) different momenta, (b) different angles. (The magnetic field is directed perpendicular to the plane of the figure.)

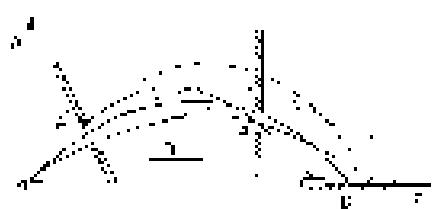


Fig. 29-3. An axis-feed spectrometer.

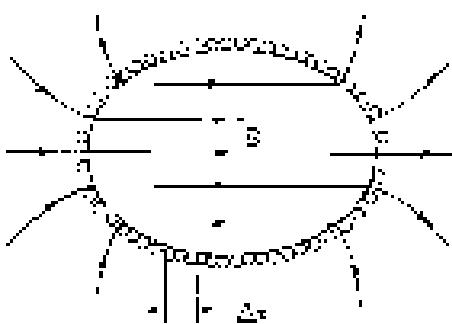


Fig. 29-4. An ellipsoidal coil with coils wound in each axial interval Δx produces a uniform magnetic field inside.

necessary than if the particles enter at right angles to the field edge. Figure 29-3(b) shows the trajectories of three particles, all with the same momentum but entering the field at different angles. You see that they make different trajectories, but all leave the field very close to the point α . We say that there is a "focus." Such a focusing property has the advantage that larger angles can be accepted at a although some "focusing" is imposed, as shown in the figure. A larger angular acceptance usually means that more particles are counted in a given time, decreasing the time required for a given measurement.

By varying the magnetic field, or moving the counter along its x , or by using many counters to cover a range of x , the "momentum spectrum" ($f(p)$, we mean that the number of particles with momenta between p and $p + dp$) is obtained. Such measurements have been made, for example, to determine the distribution of energies in the α -decay of various nuclei.

There are many other forms of momentum spectrometers, but we will describe just one more, which has an especially large solid angle of acceptance. It is based on the radial orbits in a uniform field, like the one shown in Fig. 29-1. Let's think of a cylindrical coordinate system— (p, θ, z) —set up with the z -axis along the direction of the field. If a particle is emitted from the origin at some angle θ with respect to the z -axis, it will move along a spiral whose equation is

$$y = a \sin k z, \quad z = b z,$$

where a , b , and k are parameters you can calculate with the help of p , θ , and the "magnetic field" B . If we plot the distance y from the z -axis as a function of z as a good coordinate, for the several starting angles, we will get curves like the ones shown in Fig. 29-3. (Remember that this is just a kind of projection of a helical trajectory y . When the angle between z -axis and the starting direction is larger, the peak value of y is larger but the longitudinal velocity v_z less, so the trajectories for different angles tend to "join" to a kind of "cloud" near the point β on the right.) If we put a narrow aperture of A , particles with a range of initial angles can still get through and pass on to the side, where they can be counted by the long detector D .

Particles are not forced to start at the origin with a big initial momentum but at the same angles, follow the paths shown by the broken lines and just go through the aperture at β . So the spectrometer selects a small interval of momenta. The advantage over the first spectrometer described is that the aperture A —and the aperture A' —can be much larger so that particles which "leave the source" in a rather large solid angle are measured. A large fraction of the particles from the source are used—an important advantage for weak sources or for very precise measurements.

One pays a price for this advantage, however, because a large volume of uniform magnetic field is required, and this is usually only practical for low-energy particles. One way of making a uniform field, you remember, is to wind a coil on a sphere with a surface current density proportional to the sine of the angle. You can also show that the same thing is true for an ellipsoid of rotation. So such spectrometers are often made by winding an elliptical coil on a wooden (or aluminum) ellipsoid. All that is required is that the current in each interval of axial distance Δz be the same, as shown in Fig. 29-4.

29-3 An electrostatic lens

Particle focusing has many applications. For instance, the electrons that leave the cathode in a TV picture tube are brought to a focus at the screen—to make a bright spot. In this case, one wants to take electrons all of the same energy but with different impact angles and bring them together in a small spot. The problem is like focusing light with a lens, and the question is what the corresponding focusing particles are called a lens.

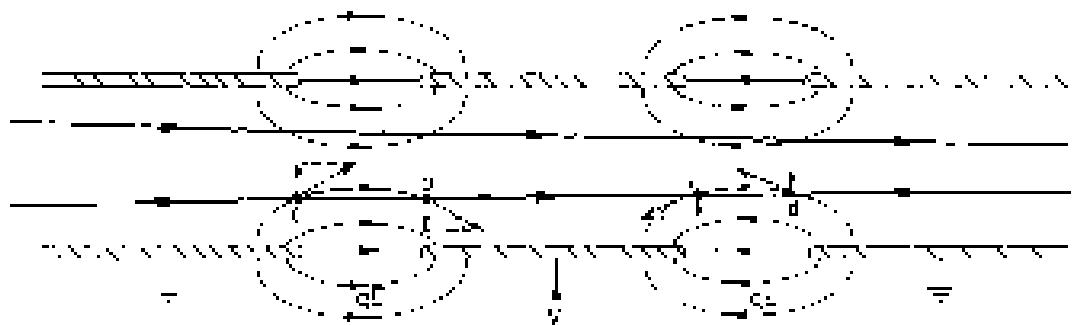


Fig. 29-5. An electrostatic lens. The field lines shown are "lines of force," since they "curve" the law of qE .

One example of an electron lens is sketched in Fig. 29-5. It is an "electrostatic" lens whose operation depends on the electric field between two adjacent electrodes. Its operation can be understood by considering what happens to a particle, being bent away from the axis. When two electrodes are used, the region in between a particle will have a component of force which points them toward the axis. You might think that they would get an equal and opposite impulse in the region B , but that's not so. By the time the electrons reach B , they have gained energy and so pass far now in the region A . The forces are the same, but the time is shorter, so the impulse is less. In going through the regions A and B , there is a net axial impulse, and the electrons are bent away from the axis. In passing the high-voltage region, the particles get slowed or sick toward the axis. The force is outward, in a region very靠近 the axis, so the particle stays longer in the last region, so there is again a net impulse. For distances not too far from the axis, the total impulse through the lens is proportional to the distance from the axis. Can you see why? And this is the axial or accessory for lens-type focusing.

You can use the same arguments to show that there is focusing if the potential of the middle electrode is either positive or negative with respect to the other two. Electrostatic lenses of this type are commonly used in cathode-ray tubes and in some electron microscopes.

29-4 A magnetic lens

Another kind of lens—often found in electron microscopes—is the magnetic lens sketched schematically in Fig. 29-6. A cylindrically symmetric electromagnet has very sharp circular pole tips which produce a strong, non-uniform field in a small region. Electrons which travel radially through this region are focused. You can understand this mechanism by looking at the longitudinal view of the pole tip region shown in Fig. 29-7. Consider two electrons, 1 and 2, that leave the source S of Fig. 29-6 with respect to the axis. As electron 1 moves horizontally toward the field, it is deflected away from the axis by the horizontal component of the field. But since it will have a lateral velocity, as that which it passes through the air, or vacuum, field, it will get an impulse toward the axis. As this motion is taken out by the magnetic force as it leaves the field, so the net effect is an impulse toward the axis, plus a "sidetoss" about the axis. At the same time particle 2 also passes through the air, it is deflected toward the axis. By the figure, the divergent electrons are brought into parallel paths. The action is just like a lens with an image at the focal point. Another similar lens upstream can be used to focus the electrons back to a single point, giving an image of the source S .

29-5 The electron microscope

You know that electron microscopes can "see" objects too small to be seen by optical microscopes. We discussed in Chapter 20 of Vol. I the basic limitations of any optical system due to diffraction of the lens opening. If a lens opening with



Fig. 29-6. A magnetic lens.

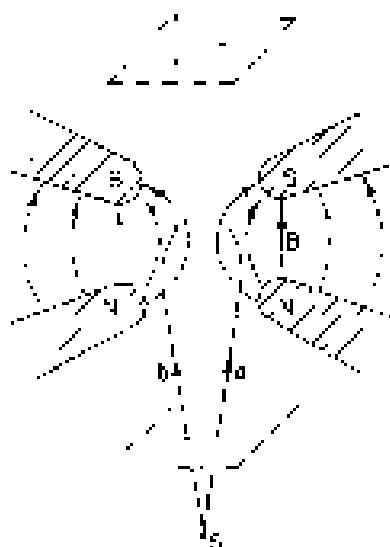


Fig. 29-7. Electron motion in the magnetic lens.

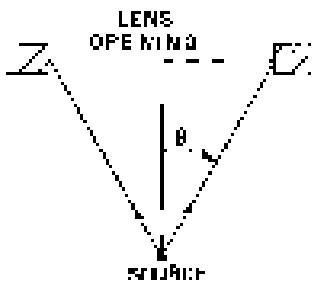


Fig. 24-0. The resolution of a microscope is limited by the angle subtended from the source.

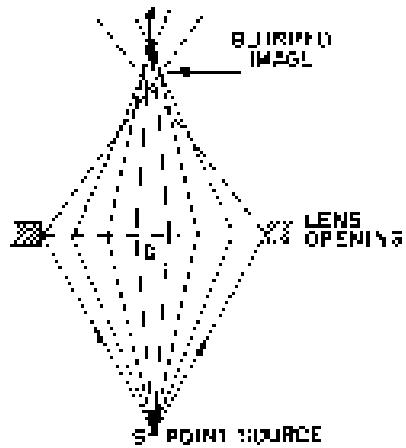


Fig. 2P-5. Spherical aberration of a lens.

turns the angle α_A from zero (see fig. 2a-3). Two initial bearing signs at the source can now be seen in sequence if they are closer than about

$$\phi = \frac{\lambda}{\sin \theta}$$

where λ is the wavelength of the light. With the best optical microscopes, θ approaches the theoretical limit of 90° , so δ is about equal to λ , or approximately 500 to 800 nm.

The same limitation would also apply to our electron microscope, but there the wavelength is the Planck-Jewett distance, about 0.03 angstrom. If one could use a lens opening of even $\times 10^4$, it would be possible to see objects only ± 0.001 angstrom apart. Since the distance of the DNA is typically 1 or 2 angstroms apart, we could get photographs of such DNA. But why would be easy; we would need a photograph of the DNA structure. What a tremendous thing that would be! Most of present-day research in materials' biology is an attempt to figure out the shapes of various protein molecules. How could one be satisfied?

Uniformly, the best case single power that has been employed in an electron microscope is more like 13 magnifications. The reason is that the objective lens has its diaphragm at a large opening. A 13-magn. lens "hyperfocal" distance, which means that rays at large angles from the axis have a different point of inflection than the rays nearer the axis, as shown in Fig. 20-19. By special techniques, optical microscope lenses can be made with a negligible aberration at infinity, but no one has yet been able to make an electron lens which avoids spherical aberration.

In fact, one can show that any electrostatic or magnetic lens of the type we have described must have an irreducible amount of spherical aberration. This aberration—together with diffraction limits the resolving power of electron microscopes to their apertural value.

The limitation we have mentioned does not apply to electric and magnetic fields which are not axially symmetric or which are not constant in time. Perhaps some day someone will think of a new kind of electron lens that will overcome the inherent aberration of the simple electron lens. Then we will be able to photograph atoms directly. Perhaps one day chemical compounds will be analyzed by looking at the positions of the atoms rather than by looking at the color of some photographic film.

29-4 Accelerator with feedback

Magnetic fields are also used to produce special particle trajectories in high-energy particle accelerators. Machines like the cyclotron and synchrotron bring particles to high energies by passing the particles repeatedly through a strong electric field. The particles are held on their orbits by a magnetic field.

We have seen that a particle in a uniform magnetic field will go in a circular orbit. This, however, is true only for a perfectly uniform field. Imagine a field B which is nearly uniform over a large area, but which is slightly stronger in one region than in another. If we put a particle of a magnetomagnetic field, it will go in a nearly circular orbit over the entire $B = \mu_0 H$. The radius of curvature will, however, be slightly smaller in the region where the field is stronger. The orbit is not a closed circle but will "walk" through the field, as shown in Fig. 36-10. We can, if we wish, consider that the "light" "on" in the field produces an extra angular kick which sends the particle off on a new track. If the particles travel millions of revolutions in an accelerator, some kind of "light trapping" is needed which will tend to keep the trajectory close to some design orbit.

Another difficulty with a uniform field is that the particles do not remain in a plane. If they start out with the slightest angle—or are given a slight angle by any small error in the field—they will go in a helical path that will eventually take them into the magnet pole or the ceiling or floor of the vacuum tank. Some arrangement must be made to inhibit such vertical drifts; the field must provide "guiding centers," as well as radial focusing.

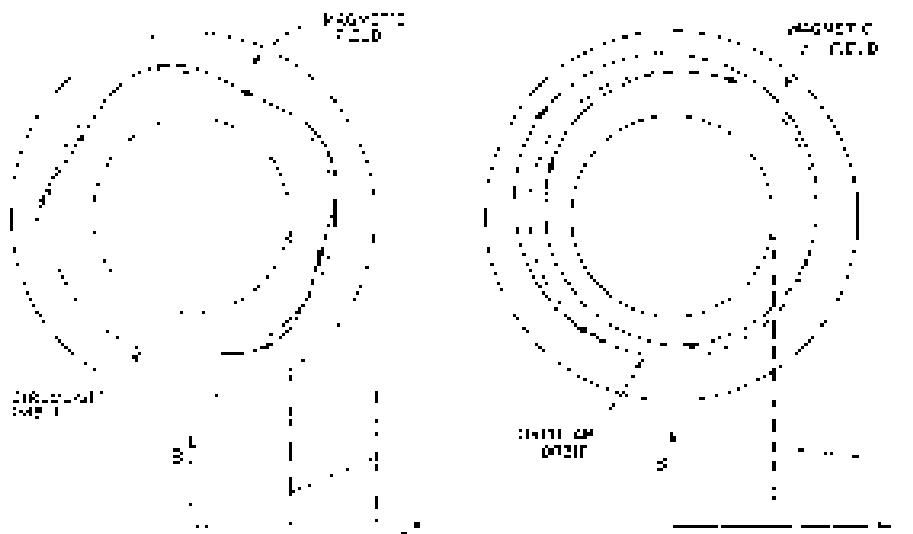


Fig. 29-11. Radial motion of a particle in a magnetic field with a large positive slope.

Fig. 29-12. Radial motion of a particle in a magnetic field with a small negative slope.

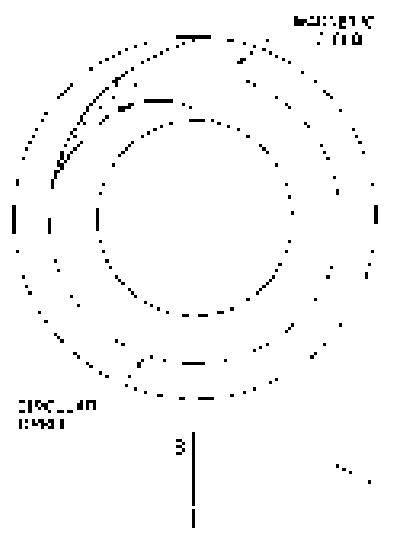


Fig. 29-12. Radial motion of a particle in a magnetic field with a large negative slope.

One would, at first, guess that radial focusing could be provided by making a magnetic field which increases with increasing distance from the center of the design path. Then if a particle goes out to a large radius, it will be in a stronger field where it will bend back toward the central radius. If it goes to too small a radius, the bending will be less, and it will be returned toward a larger radius. If the particle is once started off, say angle α with respect to the design orbit, it will tend to return to the design orbit, as shown in Fig. 29-11. The radial focusing would keep the particle on the circular path.

Actually there is still some radial focusing even with the opposite field slope. This can happen if the radius of curvature of the trajectory does not increase as rapidly as the increase in the distance of the particle from the center of the field. The particle orbits will be as drawn in Fig. 29-12. If, however, the field is too large, however, the orbits will not return to the design radius but will spiral inward to a point as shown in Fig. 29-13.

We quickly see that we must add the right equations of the "relative gradient" at $y = 0$ and $x = 0$:

$$\alpha = \frac{\partial B_x}{\partial r}, \quad (29.2)$$

A guide field gives no focusing if this relative gradient is greater than -1.

A radial field gradient will also produce vertical forces on the particles. Suppose we have a field that is stronger nearer to the center of the orbit and weaker at the outside. A section cross section of the magnet at right angles to the orbit might be as shown in Fig. 29-14. (For present the orbits would be coming out of the page.) If the field is to be stronger to the left and weaker to the right, the lines of the magnetic field must be curved as shown. We can see that this must be so by using the fact that the circulation of B is zero in free space. If we take coordinates as shown in the figure, then

$$(\nabla \times B)_x = \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} = 0,$$

$$\text{or} \quad \frac{\partial B_z}{\partial y} = \frac{\partial B_y}{\partial z}. \quad (29.3)$$

Since we assume that $\partial B_z/\partial x$ is negative, there must be some negative $\partial B_y/\partial z$.

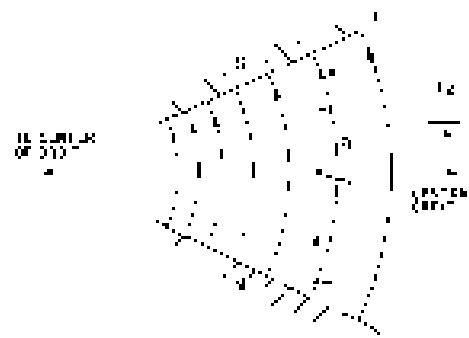


Fig. 29-14. A section guide field as seen in a cross section perpendicular to the orbits.

If the "horizontal" plane of the orbit is a plane of symmetry where $B_0 = 0$, then the radial component B_r will be negative above the plane and positive below. This must be cured as shown.

Such a field will have no real focusing properties. Imagine a particle that is travelling more or less parallel to the central orbit path, say at $\theta = 90^\circ$. The horizontal component of B will exert a downward force on it. If the proton is beyond the central orbit, the force is reverse. So there is an effective "spring" that holds the vertical orbit. From our arguments just will be vertical focusing provided that the vertical field decreases with increasing radius. In the field B which is positive, there will be "vertical defocusing." So for real focusing up the field must be less than zero. We know, however, that the radial focusing a had to be greater than -1. The parameter a is taught to give the condition that

$$-1 < a < 0$$

The particle must be kept in stable orbits in any attractive values very near zero so we must keep the focusing and defocusing small. The value $a = -1$ is typically used.

29-7 Alternating-gradient Focusing

Since the focusing of w_0 is then "weak" focusing, it is clear that much more effort is required than up until now given by a large positive gradient (w_0) . By this, the vertical forces would be strongly defocusing. Similarly, large negative slopes in w_0 would give stronger vertical forces and would cause radial defocusing. It was realized a year 10 years ago, however, that a force that alternates between strong focusing and strong defocusing can still have total focusing force.

To explain how alternating-gradient focusing works, we will first describe the operation of a quadrupole lens, which is based on the same principle. Imagine that a uniform negative magnetic field is added to the field of Fig. 29-15, with the strength adjusted to make zero field at the orbit. The resulting field, at small distances from the central point, would be like the field shown in Fig. 29-15. Such a four-pole magnet is called a "quadrupole lens." A positive particle that enters (from the rest) to the right or left of the center is pushed back toward the center. If the particle enters above or below, it is pushed away from the center. This is a horizontal focusing lens. If the horizontal gradient is reversed—it can be done by reversing all the polarities—the signs of all the forces are reversed and we have a vertical focusing lens, as in Fig. 29-16. For each lens, the field strength—and therefore the focusing forces—increases linearly with the distance of the lens from the axis.

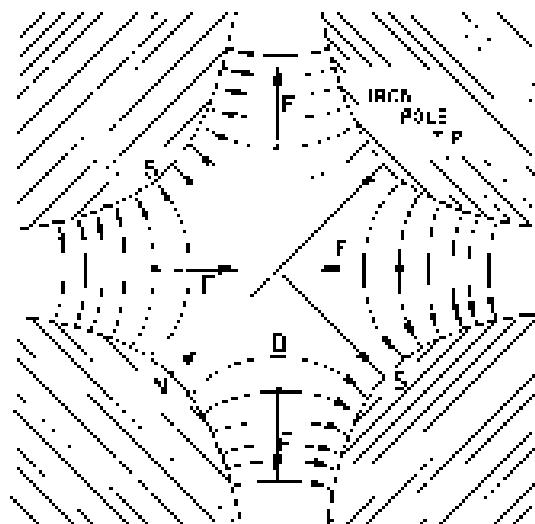


Fig. 29-15. A horizontal focusing quadrupole lens.

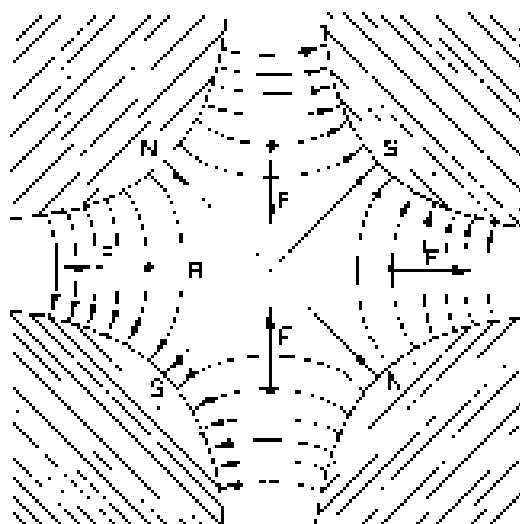


Fig. 29-16. A vertical focusing quadrupole lens.

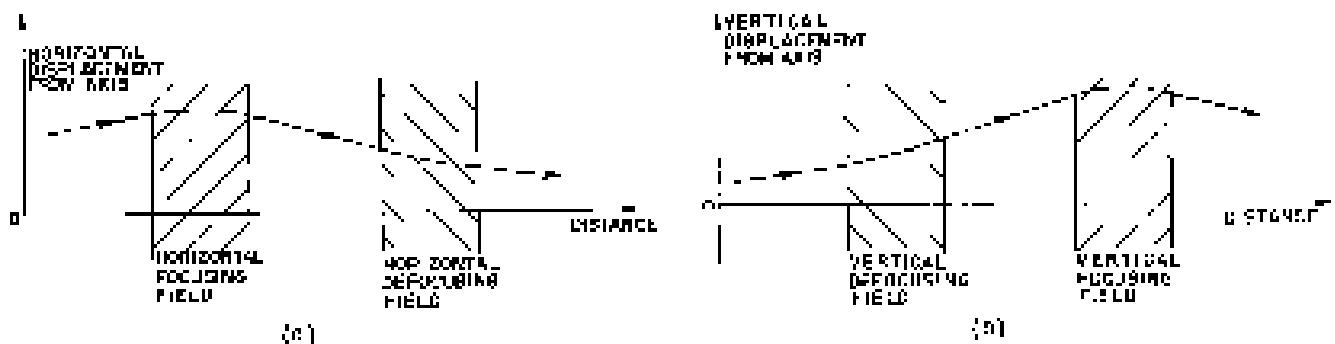


Fig. 29-17. Horizontal and vertical focusing with a pair of quadrupole lenses.

Now imagine that two such lenses are placed in series. If a particle enters with some horizontal displacement from the axis as shown in Fig. 29-17(a), it will be deflected toward the axis in the first lens. When it arrives at the second lens it is closer to the axis, so the force exerted is less and the outward deflection is less. There is a net horizontal toward the axis, the energy effect is horizontally focusing. On the other hand, if we look at a particle which enters off the axis in the vertical gradient, the path will be as shown in Fig. 29-17(b). The particle is first deflected along from the axis, but then it arrives at the second lens with a larger displacement. Since a stronger force, and this is best toward the axis. Again the net effect is focusing. Thus a pair of quadrupole lenses acts independently for horizontal and vertical motion, very much like the septuads. Quadrupole lenses are used to focus and control beams of particles in all the same way that optical lenses are used for light beams.

We should point out that an alternating-gradient system does not always produce focusing. If the gradients are too large (in relation to the particle momentum or to the spacing between the lenses), the net effect can be a defocusing one. You can see how that could happen if you imagine that the spacing between the two lenses of Fig. 29-17 were increased, say, by a factor of three or four.

Let's return now to the synchrotron guide magnet. We can consider that it consists of an alternating sequence of "positive" and "negative" lenses with a superimposed uniform field. The uniform field serves to bend the particles, on the average, in a "horizontal" circle (with no effect on the vertical motion), and the alternating lenses act on any particles that might tend to go astray—pushing them always toward the central orbit (on the average).

There is a nice mechanical analog which demonstrates that a force which alternates between a "focusing" force and a "defocusing" force can have a net "focusing" effect. Imagine a mechanical "pendulum" which consists of a rod with a weight on the end, suspended from a pivot which is caused to be moved rapidly up and down by a motor driven crank. Such a pendulum has two equilibrium positions. Besides the normal, downward-hanging position, the pendulum is in an equilibrium "hanging upward"—with its "bob" above the pivot. Such a pendulum is shown in Fig. 29-18.

By the following argument you can see that the vertical motion motion is equivalent to an alternating, corrective force. When the pivot is accelerated downward, law "gug" tends to pull inward, as indicated in Fig. 29-19. When the pivot is accelerated upward, the effect is reversed. This force restoring the "bob" back to the axis is defocusing, but the average effect is a force toward the axis. So the pendulum will swing clockwise for small initial positions which is just opposite the normal one.

There is, of course, a much easier way of keeping a pendulum upside down, and that is by balancing it on your finger! But try to balance the suspended stick on the other finger! Or one stick with your eyes closed! Balancing involves making a correction for what is going wrong. And this is not possible, in general, if there are several things going wrong at once. In a synchrotron there are billions of particles going around together, each one of which may start out with a different "error." The kind of focusing we have been describing works on them all.

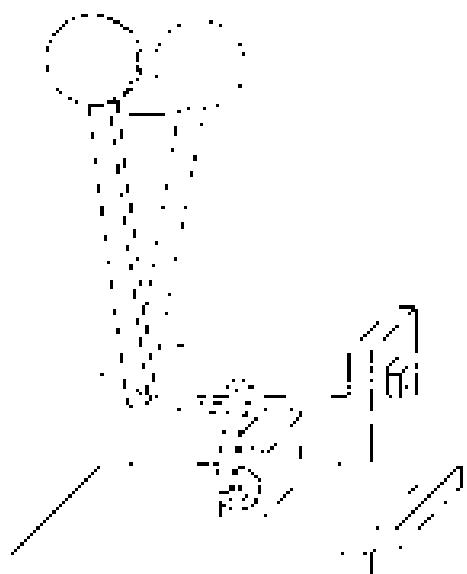


Fig. 29-18. A pendulum with an oscillating pivot can have a stable position with the bob above the pivot.

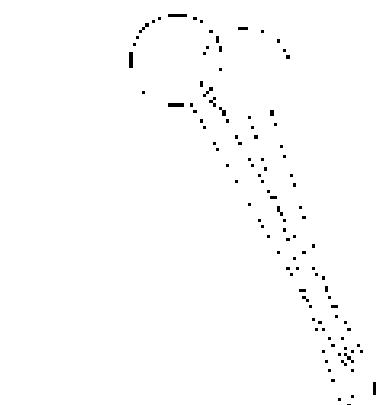


Fig. 29-19. A downward acceleration of the pivot causes the pendulum to move toward the vertical.

29-8 Motion in crossed electric and magnetic fields

So far we have talked about variables in electric fields only or in magnetic fields only. There are some interesting effects when there are both fields & fields at the same time. Suppose we have a uniform magnetic field B and a uniform field E at right angles. Suppose also that v is perpendicular to B with velocity v as shown in Fig. 29-10. (The figure is a photograph from a color slide.) We can understand this motion qualitatively. When the particle (assumed positive) moves in the direction of E , it picks up speed, and so it is bent less by the magnetic field. When it is going against the E field, it loses speed and is continuously bent more by the magnetic field. The net effect is that it has an average "drift" in the direction of $E \times B$.

We can, in fact, show that the motion is a uniform circular motion superimposed on a uniform rectilinear motion in the rapid $v_{\parallel} = E/B$. The trajectory in Fig. 29-10 is a cycloid. Imagine an observer who is moving to the right at a constant speed. In his frame our magnetic field gets transformed to a new magnetic field plus an electric field in the v_{\parallel} direction. If he has just the right speed, his local electric field will be zero, and he will see the electron going in a circle. So the motion we see is a circular motion, plus a translation in the drift speed.

29-5 The motion of electrons in crossed electric and magnetic fields: the basis of the synchrotron tubes, i.e., oscillators used for generating microwave energy.

There are many other interesting examples of particle motions in electric and magnetic fields—such as the orbits of the electrons and protons trapped in the Van Allen belts—but we do not sufficiently have the time to deal with them here.

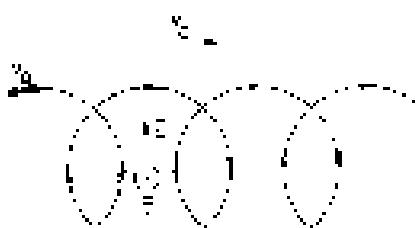


Fig. 29-10. Path of a particle in crossed electric and magnetic fields.

The Internal Geometry of Crystals

30-1 The internal geometry of crystals

We have finished the study of the basic laws of electricity and magnetism, and we are now going to study the electromagnetic properties of matter. We begin by describing solids—that is, crystals. When the atoms of matter are not moving around very much, they get stuck together and arrange themselves in a configuration with as low an energy as possible. If the atoms in a certain place have found a pattern which seems to be of low energy, then the atoms somewhere else will probably make the same arrangement. For these reasons, we have in a solid material a repetitive pattern of atoms.

In other words, the conditions in a crystal are this way: The environment of a particular atom in a crystal has a certain arrangement, and if you look at the same kind of an atom at another place further along, you will find one whose surroundings are exactly the same. If you go on from further along by the same distance you will find the conditions exactly the same once more. The pattern is repeated over and over again—and, of course, in three dimensions.

Imagine the problem of designing a wallpaper—or a cloth, or some geometric design for a plane area—in which you are supposed to have a design element which repeats and repeats until repeats, so that you can make the area as large as you want. This is the two-dimensional analog of a problem which a crystal solves in three dimensions. For example, Fig. 30-1(a) shows a common kind of wallpaper design. There is a single element repeated in a pattern that can go on forever. The geometric characteristics of this wallpaper design, considering only its repetition properties and not worrying about the geometry of the flower itself as is artist's merit, are contained in Fig. 30-1(b). If you start at any point, you can find the corresponding point by moving the distance a along the direction of arrow 1. You can also get to a corresponding point if you move the distance b in the direction of the other arrow. There are, of course, many other direct axes. You can go, for example, from point a to point b and reach a corresponding position, but such a step can be considered as a combination of a step along direction 1, followed by a step along direction 2. One of the basic properties of the pattern can be described by the two shortest steps to nearby equal positions. By "equal" positions we mean that if you were to stand in any one of them and look around you, you would see exactly the same thing as if you were to stand in another one. That's the fundamental property of a crystal. The only difference is that a crystal is a three-dimensional arrangement instead of a two-dimensional arrangement; and naturally, instead of flowers, each element of the lattice is some kind of an arrangement of atoms, perhaps six hydrogen atoms and two carbon atoms, in some kind of pattern. The pattern of atoms in a crystal can be found out experimentally by x-ray diffraction. We have mentioned this method briefly before, and won't say any more now except that the precise arrangement of the atoms in space has been worked out for most simple crystals and also for some fairly complex ones.

The internal pattern of a crystal shows up in several ways. First, the bonding strength of the atoms in certain directions is usually stronger than in other directions. This means that there are certain planes through the crystal where it is more easily broken than others. They are called the cleavage planes. If you break a crystal with a knife blade it will often split apart along such a plane. Second, the internal structure often appears at the surface because of the way the crystal was formed. Imagine a crystal being deposited out of a solution. There are the atoms floating around in the solution and finally settling down when they find a position

30-1 The internal geometry of crystals

30-2 Chemical bonds in crystals

30-3 The growth of crystals

30-4 Crystal lattices

30-5 Symmetries in two dimensions

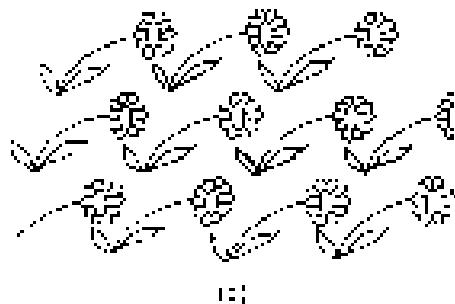
30-6 Symmetries in three dimensions

30-7 The strength of metals

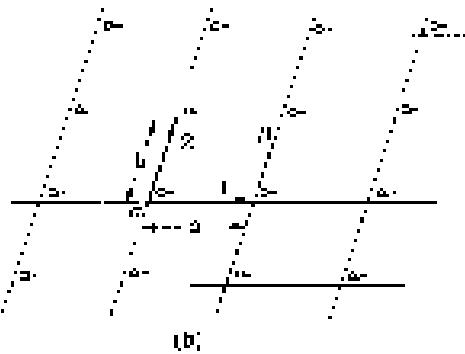
30-8 Dislocations and crystal growth

30-9 The Bragg-Nye crystal model

Reference: C. E. Lof, *Introduction to Solid State Physics*, John Wiley and Sons, Inc., New York, 2nd ed., 1956.



(a)



(b)

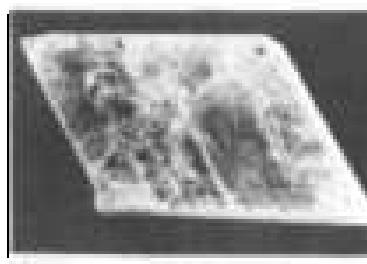
Fig. 30-1. A repeating pattern in two dimensions.



(a)



(b)



(c)

Fig. 20-2. Three crystals of sodium chloride, (a) cube, (b) pyramidal shape, (c) fine grains.

of lowest energy. (It's as if the wallpaper got made by flowers drifting around until one decided to get stuck, and then the next, and the next, until the pattern gradually grew.) You can appreciate that there will be additional forces in what it will grow to if the net speed is not constant; thereby grows up at first slow and of peripherated shape. Because of such effects, the resulting shapes of many crystals show some of the character of the material in addition to the shape of the lattice.

For example, Fig. 20-2(a) shows the shape of a typical pyramidal crystal whose internal pattern is cubic! (If you look closely at such a crystal, you will notice that the facets do not make a very good rectangle because the sides are not all of equal length—they are "wavy," and very unequal. But in any case it is a very pyramidal shape—the angle between the faces is exactly 70° .) Clearly the size of any particular nucleus controls the shape, but the angle is an intrinsic function of the internal geometry. So every crystal of a given kind has a different shape, even though the angles between the spindles of the lattice stay the same.

The internal geometry of a crystal is often also very evident in its external shape. Figure 20-2(b) shows the shape of a typical pyramid of salt. Again the crystal is not a perfect cube, but the faces are exactly straight, angles to one another.

A more complicated crystallization, which has the shapes shown in Fig. 20-2(c), is a highly anisotropic crystal, as is easily seen from the fact that it is very tough (it's hard to pull apart in one direction [the one along the figure], but it's very easy to pull it by pulling apart in the other direction [vertically]). It has commonly been used to obtain very long, thin sheets—Magnesite quartz is two asbestos-related minerals, one being silvery. A third example of a mineral with such a regularity, which has the interesting property that it is easily pulled apart in two directions but not in the third, appears to be mica, a very strong, tough mineral.

20-2 Chemical bonds in crystals

The mechanical properties of crystals clearly depend on the kind of chemical bonding between the atoms. The strikingly different strength of bonds along different directions depends on the kinds of interatomic binding in the different directions. You have already learned in chemistry, no doubt, about the different kinds of chemical bonds. First, there are ionic bonds, as we have already discussed for sodium chloride. Roughly speaking, the sodium atoms have lost an electron and become positive ions; the chlorine atoms have gained an electron and become negative ions. The positive and negative ions are attracted in a three-dimensional checkerboard and are held together by electrical forces.

The covalent bond—in which electrons are shared between two atoms—is more common and is usually very strong. In a diamond, for example, the carbon atoms have covalent bonds in all four directions to the nearest neighbors, so the crystal is very hard indeed. There is also covalent bonding between silicon and oxygen in a quartz crystal, but there the bond is really only partially covalent. Because there is now complete sharing of the electrons, the atoms are partly charged, and the crystal is somewhat ionic. Nature is not as simple as we try to make it; there are really all possible gradations between covalent and ionic bonding.

A last crystal has still another kind of bonding. In it there are large molecules in which the atoms are held strongly together by covalent bonds so that the molecule is a tough structure. But since the strong bonds are completely localized, there are only relatively weak attractions between the separate, individual molecules. In such molecular crystals the molecules keep their individual identity, so to speak, and the internal arrangement might be as shown in Fig. 20-3. Since the molecules are not held strongly to each other, the crystals are easy to break. They are quite different from something like diamond, which is really one giant molecule that cannot be broken anywhere without disrupting strong covalent bonds. Paraffin is another example of a molecular crystal.

An extreme example of a molecular crystal occurs in a substance like solid neon. There is very little attraction between the atoms—each atom is a completely

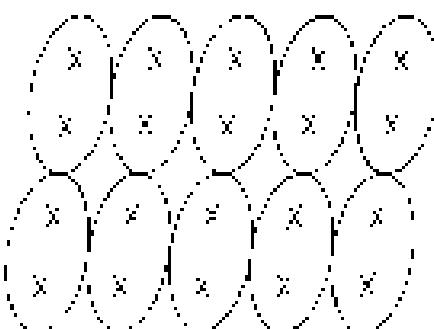


Fig. 20-3. The lattice of a molecular crystal.

separated from one another by a distance of about 10⁻⁸ cm. At very low temperatures, the thermal motion is very small, so the slight interaction forces cause the atoms to settle down into a regular array like a pile of closely packed spheres.

The metals form a completely different class of substances. The bonding is of a very different kind. Not only the bonding is not dependent upon distance but it is a property of the whole atom. The valence electrons are not attached to one atom or to a pair of atoms but to a whole group and the oxygen. Each atom contributes one electron to a common pool of electrons, and this sum is positive ion resulting in the net negative electrons. The valence shells hold together like a bunch of glued spheres.

In the metals, since there are no special bonds, any particular direction there is no strong dependence on the bonding. They are still cohesive, however, because the total energy is lowest when the atoms are all arranged in some definite way—although the energy of the preferred arrangement is not usually much lower than other possible ones. By a just approximation, the atoms of many metals act like small spheres packed in as tightly as possible.

30-3 The growth of crystals

Try to imagine the natural formation of crystals in the earth. In the soil's surface there is a big mixture of all kinds of atoms. They are being continually jolted about by volcanic action, by wind, and by water—continually being moved about and mixed. Yet, by some trick, silicon atoms gradually begin to form each other, and to find oxygen atoms, to make silica. One atom at a time is added to the others to build up a crystal—the mixture goes unmixed. And as new ones nearby, sodium and chlorine atoms are finding each other and building up a crystal of salt.

How does it happen that once a crystal is started, it grows only in a particular kind of street to stick on? It happens because the whole system is working toward the lowest possible energy. A growing crystal will seek a new atom that is going to make the energy as low as possible. But how does it know that a silicon or an oxygen atom at some particular spot is going to result in the lowest possible energy? It does it by trial and error. In the liquid, all of the atoms are in perpetual motion. Each atom bounces against its neighbors about 10¹⁰ times every second. If it hits against the right spot of a growing crystal, it can do somewhat sort of jumping off again if the energy is low. By continually testing over periods of millions of years at a rate of 10¹⁰ tests per second, the atoms gradually build up at the places where they find their lowest energy. Eventually they grow into big crystals.

30-4 Crystal lattices

The arrangement of the atoms in a crystal—the crystal lattice—can take on many geometric forms. We would like to describe first the simpler lattices, which are characteristic of most of the metals and of the solid form of the nonmetals; they are the cubic lattices which can occur in two forms: the body-centered cubic, shown in Fig. 30-4(a), and the face-centered cubic shown in Fig. 30-4(b). The drawings show, of course, only one cubic lattice; the pattern continues in all directions. When we make the drawing clearer, only the "centers" of the atoms are shown. In a actual crystal, the atoms are mere little spheres in contact with each other. The dark and light spheres in the drawings may, in general, stand for different kinds of atoms as may be the same kind. For instance, iron has a body-centered cubic form at low temperatures, but a face-centered cubic form at high temperatures. The physical properties are quite different in the two crystalline forms.

How do such forms come about? Imagine that you have the problem of packing spheres on man together as tightly as possible. This Na_2 would be formed by making a layer of a "hexagonal close-packed group," as shown in Fig. 30-5(a). Then you could build up a second layer like the first, but displaced horizontally,

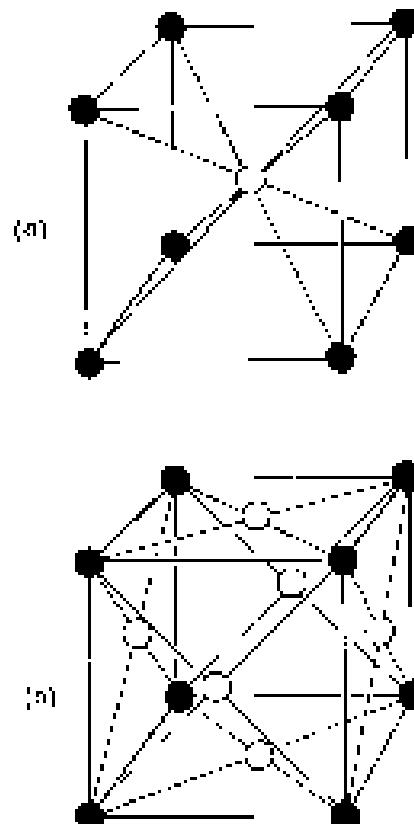


Fig. 30-4. The unit cell of cubic crystals. (a) body-centered, (b) face-centered.

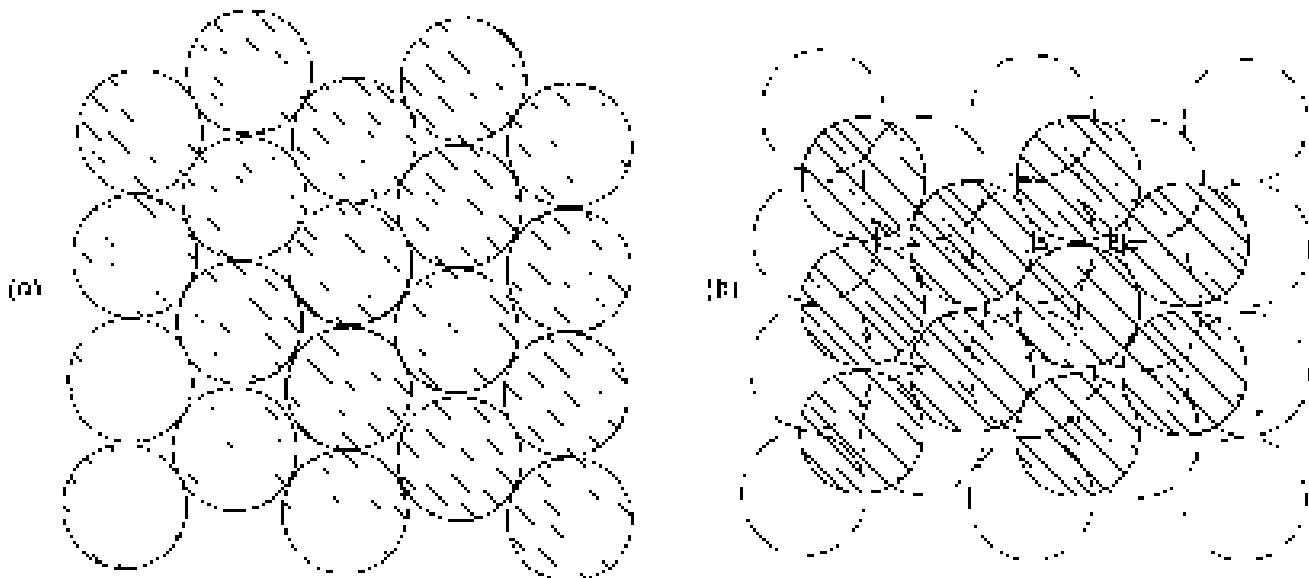


Fig. 20-5. Building up a hexagonal close-packed lattice.

as shown in Fig. 20-5(b). Now, you can put on the third layer. But, notice there are two distinct ways to placing the spheres, etc. If you start the third layer by placing an atom at A in Fig. 20-5(b), each atom in the third layer is directly above an atom of the bottom layer. On the other hand, if you start the third layer by putting an atom at the position B, the atoms of the third layer will be replaced at points exactly in the middle of a triangle formed by three atoms of the bottom layer. Any other starting place is equivalent to A or B, so there are only two ways of placing the third layer.

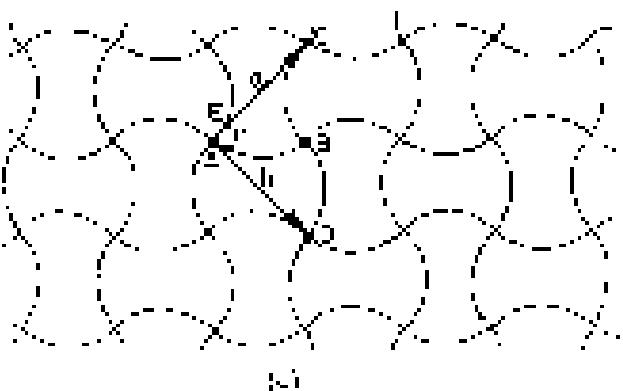
If the third layer has an atom at point A, the crystal lattice is a face-centered cube—but seen at an angle. I assure you that starting with hexagons you can end up with cubes. But notice that a cube looked at from a corner has a hexagonal pattern. For instance, Fig. 20-6 could represent a plane drawing of a cube seen in perspective!

The third layer is added to Fig. 20-5(b) by starting with an atom at A; this is no surface, however, and the lattice has instead only a hexagonal symmetry. It is clear that both presentations we have described are equally close-packed.

Some metals—for example, copper and zinc—choose the first alternative, the face-centered cube. Others—for example, beryllium and magnesium—choose the other alternatives; they form hexagonal crystals. Clearly, which a particular lattice structure choice depends only on the packing of the spheres, but this also be determined in part by other factors. In particular, it depends on the slight remaining angular connectedness of the interatomic forces (or, in the case of the metals, on the energy of the electron pair). You will no doubt learn all about such things in your chemistry courses.

20-5 Symmetries in two dimensions

We would now like to discuss some of the properties of crystals from a somewhat different point of view. The main feature of a crystal is that, if you start at one atom, and move to a neighboring atom one lattice unit away, you are again in the same kind of an environment. That's the fundamental property. But if you were an atom, there would be another kind of change that could take you again to the same environment—that is, another possible “symmetry.” Figure 20-7(a) shows another possible “wallpaper-type” design (though one you have probably never seen). Suppose we compare the environments for points A and B. You might, at first, think that they are the same—but not quite. Points C and D are equivalent to A, but the environment of B is like that of A only if the surroundings are reversed, as in a mirror reflection.



(a)

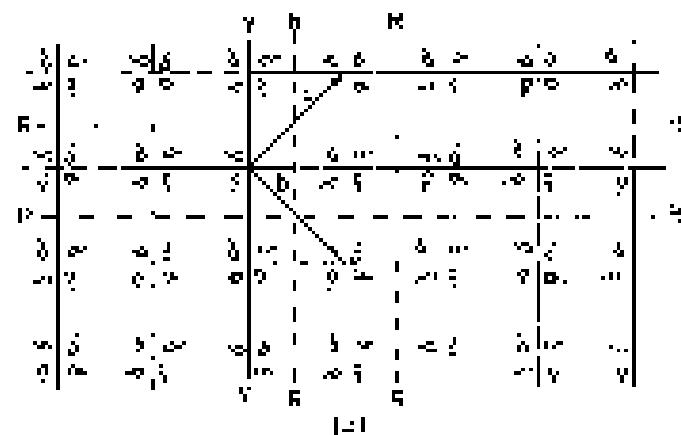


Fig. 30-7. A pattern of high symmetry.

There are other kinds of "equivalent" points in the pattern. For instance, the points E and F have the "same" environment except that one is horizontal and the other vertical. The pattern is more special. A rotation of 30° is another multiple of $\pi/6$, about a vertex such as G for the same pattern of overlapping circles. If you look at such a situation, you see it occurs on the mid-side, but inside it is much complicated than a simple rule.

Now that we have described rather special examples, let's try to figure out all the possible environments of equivalent points. First, we consider what happens in a plane. A plane has four corners, but the two overall "prevailing" corners that you can't get out of the blocks to the two others of the "equivalent" points. The two corners (read 2) are the minimum number of the lattice of $\mathbb{Z} \times \mathbb{Z}$. In Fig. 30-7(a), the two corners and the $(0,0)$ in Fig. 30-7(b) are the prevailing corners of the two patterns. We consider a corner point, say x , opposite y by $\pi/6$, or z by $-\pi/6$. Since x and y are equivalent under a 60° clockwise rotation, x is 90° from y and z is 90° from x . So x is 180° from z .

We say that these two vertices which have a "four-fold" symmetry. And we consider the other two edges which have a "three-fold" symmetry. And we consider the other two edges which have a "two-fold" symmetry. All these three types of symmetry. A rotation of a symmetry of $2\pi/3$ in Fig. 30-7(a) by rotating 120° about the center of any three edges that return back to itself.

What other kinds of symmetry should we consider? For example, a two-fold or a single-fold rotational symmetry. It is very reasonable that we consider. The only symmetry that we consider that they are no 30° symmetries. But, let's show that the $\mathbb{Z} \times \mathbb{Z}$ grid does not have 30° symmetry. Suppose we try to imagine a bit how will a 30° rotation work. Let's consider, for example, the $(0,0)$ in Fig. 30-7(b). We are to suppose that point $(0,0)$ is a 30° angle rotated and that point $(0,0)$ is the position after rotation. And the distance between $(0,0)$ and that point $(0,0)$ is the same as the distance between $(0,0)$ and $(1,0)$. That is nearly impossible due to the same reason. Because $(0,0)$ is a center point of $(1,0)$. They must be a right-angle. It implies that $(0,0)$ is not $(0,0)$. The check. Just choose $(0,0)$ as one of our primitive vectors, so the angle between the two primitive vectors must be 90° . But 30° symmetry is not possible.

What about three-fold symmetry? If we rotate the two primitive vectors a and b have equal lengths and make an angle of $2\pi/3 = 120^\circ$, as in Fig. 30-8(a), then there should be some kind of problem. In the point of $(0,0)$ in Fig. 30-8(a), the vector a from $(0,0)$ is then less than b , so b is going to be a vector. There can be no three-fold symmetry. The only possibilities that we expect at least three-fold symmetry are $30^\circ, 60^\circ > 120^\circ$. 30° or 60° are also clearly possible. On way of solving this result is that the pattern can be unchanged by a rotation of one full turn (one complete circuit) of a line, one third, $1/3$ full turn, a seventh of a turn, and so on. To be precise, if we rotate a pattern n times in a 30°-separated circle, $(0,0) - 2\pi/3n$, we approach to "twofold" symmetry. You may

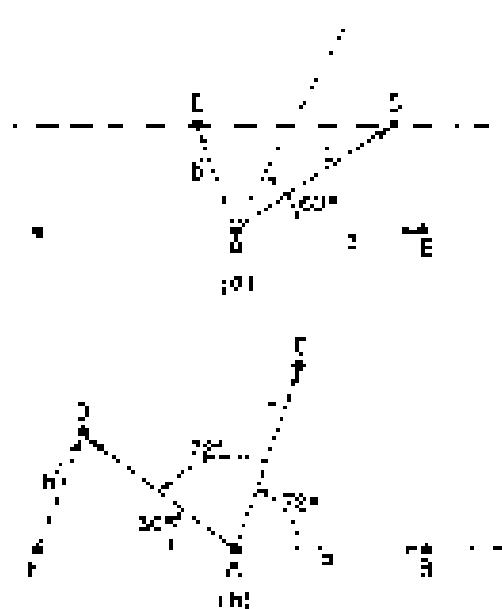


Fig. 30-8. (a) Relative symmetry of the grid of the grid. (b) the grid has no three-fold rotational symmetry is not possible.

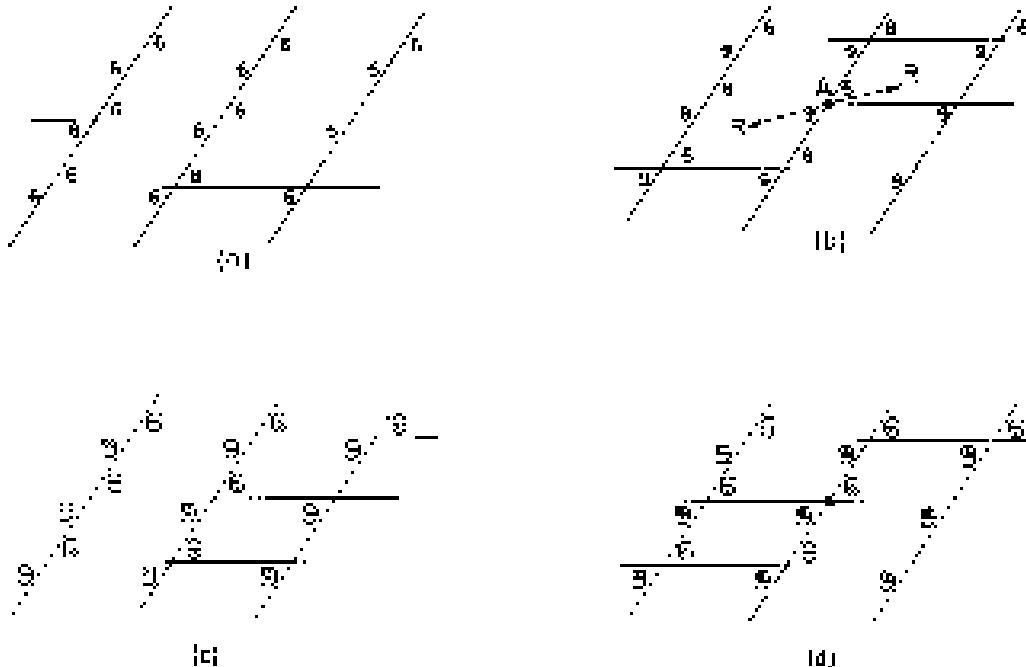


Fig. 32-9. Symmetry under inversion. Pattern (b) is unchanged if $R \rightarrow -R$, but pattern (a) is changed. In three dimensions, pattern (c) is symmetric under inversion but (d) is not.

that a pattern with a equal to 1 or to 6 has a "higher symmetry" than one with a equal to 1 or to 2.

Returning to Fig. 32-7(a), we see that the pattern has a fourfold rotational symmetry. We have drawn in Fig. 32-7(b) another design which has the same symmetry properties as part (a). The little rounded-like figures are asymmetric objects which serve to define the symmetry of the design instead of each square. Notice that the extremes are reversed in alternate squares, so that the unit cell is larger than one of the small squares. If there were no extremes, the pattern would still have fourfold symmetry, but the unit cell would be smaller. The patterns of Fig. 32-7 also have other symmetry properties. For instance, a reflection about any of the horizontal axes $R = 0$ produces the same pattern.

The pattern in Fig. 32-7 has still another kind of symmetry. If the pattern, i.e., $\mathbf{v}(R)$ is shifted by \mathbf{v}_0 in the x - y plane and then rotated by 180° around x_0 , we get back the original pattern. Such a \mathbf{v} is called a "twofold" one.

These are all the possible symmetries in two dimensions. There is one more spatial symmetry operation, which is equivalent to a combination of a 180° rotation, in which \mathbf{x} is a wave vector, up to three dimensions. It is known, for instance, that if we shift a wave point at \mathbf{k} under displacement \mathbf{R} from some origin, i.e., instead of the point \mathbf{g} in Fig. 32-9(g) it is moved to the point at $-\mathbf{R}$,

the wave at position (a) of Fig. 32-9 produces a new pattern, but is nevertheless related to the original one by reflection. For a wave-mechanical pattern (as yet, not yet seen from this point), we know that all the periodicity conditions must be equivalent to a rotation of 180° about the origin point. Suppose, however, we move the wave in Fig. 32-9(g) with three-dimensional \mathbf{k} components, then \mathbf{g} and $-\mathbf{g}$ are always at "nearest" points enough to say. At the same time, in three dimensions all the averages will be reflected at the points \mathbf{x} and $-\mathbf{x}$. This is because if we measure the heads and tails of the curves by dots and crosses, respectively, we can make a three-dimensional pattern as in Fig. 32-9(g), which is not symmetric under an inversion, but we can make a pattern like (b) as shown in (d), which does have such a symmetry. Notice that it is only possible to make a three-dimensional pattern by very careful selection of parameters.

If we characterize the spatial \mathbf{g} as a vector and average by the usual symmetry operations we can have up to 10^4 different patterns, but in fact there are 17 distinct patterns possible. We have drawn one pattern in the lowest possible

symmetry in Fig. 30-1, and one of high symmetry in Fig. 30-2. We will leave you with the game of trying to figure out all of the 17 possible patterns.

It is peculiar how few of the 17 possible patterns are used in making wallpaper and fabrics. One always sees the same three or four basic patterns. Is this because of a lack of imagination of designers, or because many of the possible patterns are not pleasing to the eye?

30-6 Symmetries in three dimensions

So far we have talked only about patterns in two dimensions. What we are really interested in, however, are patterns of atoms in three dimensions. Thus, it is clear that a three-dimensional crystal will have three primitive vectors. If we then ask about the possible symmetry operations in three dimensions, we find that there are 230 different possible symmetries! For some purposes, these 230 types can be grouped into seven classes, which is shown in Fig. 30-10. The lattice with the least symmetry is called the *random*. Its unit cell is a *para-tetrahedron*. The primitive vectors are of different lengths, and ratios of the angles between them are equal. There is no possibility of any rotations or reflection symmetry. There are, however, still two possible symmetries: the unit cell is, *a*, *b*, *c*, not changed by a "twist" or through the vertex. (By an *inversion*, another rhombus, we mean that a total displacement *R* is replaced by $-R$. In other words, that (x,y,z) goes into $(-x,-y,-z)$.) So the random class has only two possible symmetries unless there is some special twist in among the primitive vectors. For example, if all the vectors are equal and are separated by equal angles, one has the *regular tetrahedron* shown in the fig. 30-10. This figure can have no twist or reflection, but may be unchanged by a rotation about the long body diagonal.

If one of the primitive vectors, say *a*, is at right angles to the other two, we get a *hexagonal unit cell*. A new symmetry is possible: a rotation by 180° about *a*. The hexagonal cell is a hexagon in which the vectors *a* and *b* are equal and the angle between them is 60°; or a rotation of 60° or 120° or 180° about the vector *a* repeat the same lattices (five extra internal symmetries).

If all three primitive vectors are at right angles, but of different lengths, we get the *orthorhombic cell*. The figure is symmetric for rotations of 180° about the three axes. Higher-order symmetries are possible with the *orthorhombic cell*, which has all right angles and two equal primitive vectors. Finally, there is the *cubic cell*, which is the most symmetric of all.

The sum of all this adds up to various symmetries of the internal symmetries of the crystals showing up sometimes in subtle ways—in the more esoteric physical properties of the crystal. For instance, a crystal will, in general, have a linear dielectric behavior. If we describe the tensor in terms of the ellipsoid of polarizability, we should expect that some of the crystal symmetries should show up also in the ellipsoid. For example, a cubic crystal is symmetric with respect to a rotation of 90° about any one of three orthogonal directions. Clearly, the only object with this property is a sphere. A cubic crystal must be an isotropic dielectric.

On the other hand, a tetragonal crystal has a fourfold rotational symmetry (the *oblate*—that is, two of its principal axes are equal, and the third must be parallel to the axis of the crystal). Similarly, since the orthorhombic crystal has twofold rotational symmetry about three orthogonal axes, its axes must coincide with the axes of the polar ellipsoid supposed. In a like manner, any of the axes of a monoclinic crystal must be parallel to any of the principal axes of the ellipsoid. Though we can't say anything about the other axes, since none in a crystal has no rotational symmetry, the ellipsoid can have any orientation at all.

As you can see, we can make a big game of figuring out the possible symmetries and relating them to the possible physical reasons. We have considered only the *isotropized* tensor, but things get more complicated for others. For instance, for the tensor of elasticity. There is a branch of mathematics called "group theory". Let that well, such topics, but usually you can figure out what you want with common sense.

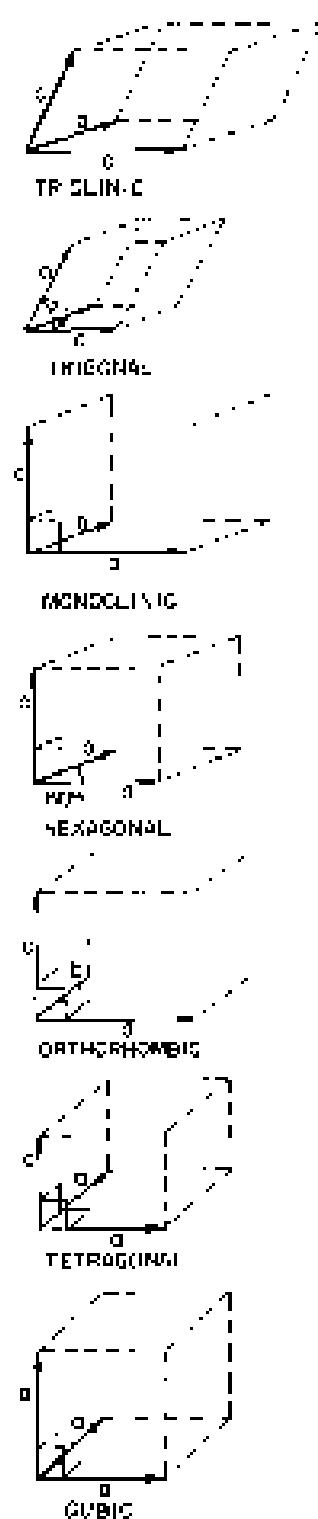


Fig. 30-10. The seven classes of crystal lattices.

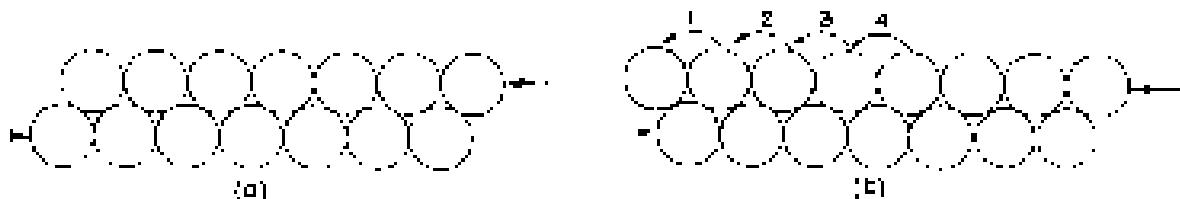


Fig. 30-11. Slippage of crystal planes.

30-7 The strength of metals

We have said that metals usually have a simple cubic crystal structure; we want now to discuss their mechanical properties—which depend on this structure. Metals are, generally speaking, very “soft,” because it is easy to slide one layer of the crystal over the next. You may think: “That’s ridiculous; metals are strong.” Not so; a single atom of a metal can be distorted very easily.

Suppose we look at two layers of a crystal subjected to a shear force, as shown in the diagram of Fig. 30-11(a). You might at first think the whole layer would deform until the force was big enough to push the whole layer “over the hump,” so that it shifted one row to the left. Although slipping does occur along a plane, it doesn’t happen that way. (If it did, you would calculate that the metal is much stronger than it really is.) What happens is more like one atom going at a time; first the atom on the left makes a jump, then the next, and so on, as indicated in Fig. 30-11(b). In effect it is the vacant space between two atoms that quickly travels to the right, with the net result that the whole second layer has moved over one atomic spacing. The slipping goes the way because it takes much less energy to lift one atom at a time over the bump than to lift a whole row. Once the force is enough to start the process, it goes the rest of the way very fast.

It turns out that, in a real crystal, slipping will occur repeatedly at one plane, then will stop there and start at some other plane. The details of why it starts and stops are quite mysterious. It is, in fact, quite strange that successive regions of slip are often fairly evenly spaced. Figure 30-12 shows a photograph of a tiny, thin copper crystal that has been stretched. You can see the various planes where slipping has occurred.

The sudden slipping of individual crystal planes is quite apparent if you take a piece of tin wire that has large crystals in it and stretch it while holding it next to your ear. You can hear a rushing “click” as the planes snap to their new positions, one after the other.

The problem of having a “loose” atom in one row is somewhat more difficult than it might appear from Fig. 30-11. When there are more layers, the situation must be something like that shown in Fig. 30-13. Such an imperfection in a crystal is called a *dislocation*. It is presumed that such dislocations are often present when the crystal was formed or are generated at some notch or crack in the surface. Once they are produced, they can move relatively freely through the crystal. The greater difficulties result from the motions of many such dislocations.

Dislocations can move freely, that is, they require little extra energy—so long as the rest of the crystal has a perfect lattice. But they may get “stuck” if they run into some other kind of imperfection in the crystal. If it takes a lot of energy to pass the imperfection, they will be stopped. This is precisely the mechanism that gives strength to imperfect metal crystals. Pure iron crystals are quite soft, but a small concentration of impurity atoms may cause enough impurities to effectively immobilize the dislocations. As you know, steel, which is primarily iron, is very hard. To make steel, a small amount of carbon is dissolved in the iron melt; if the melt is cooled rapidly, the carbon precipitates in little grains, making many microscopic dislocations in the lattice. The dislocations can no longer move above, and the metal is hard.

Pure copper is very soft, but can be “work hardened.” This is done by hammering it, or bending it back and forth. In this case, many new dislocations of various kinds are made which interfere with one another, cutting down their

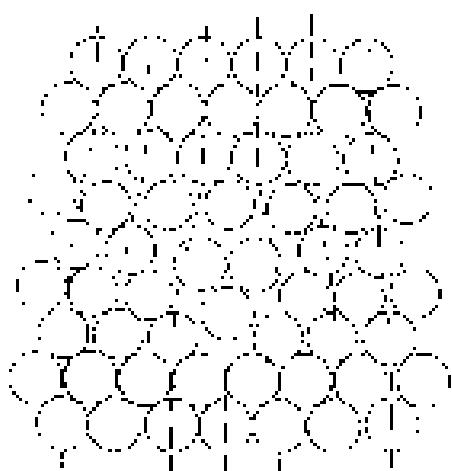


Fig. 30-12. A photograph of a copper crystal after stretching. (Courtesy of S. S. Brower, Senior Scientist, United States Steel Research Center, Monroeville, Pa.)

instability. Perhaps you've seen the trick of heating a bar of "hard soft" copper until it begins to melt; it cannot melt because it is brittle. In this process, if it becomes weak-hatched and cannot easily be pulled apart, a weak-hatched metal like copper can be made soft again by annealing at a high temperature. The twisted nature of the atoms' bonds and the deleterious short-range angle-gauge length. We know, as we described in the section about dislocation. There are many other kinds, one of which is the screw dislocation shown in Fig. 30-14. Such dislocations often play an important part in crystal growth.

30-8 Dislocations and crystal growth

One of the great problems for a long time was how crystals can possibly grow. We very frequently know it is that each atom ought to be pulled out, there is no whether it may rather be left in the crystal or not. But that means that each atom must find a place of low energy. However, an atom put on a new surface is only bound by one or two bonds from below, and doesn't have the same energy it would have if we're alone in a corner, where it would have atoms on three sides. Suppose we bring up a growing crystal as a stack of blocks, as shown in Fig. 30-15. If we try a new block at, say, position 4, it will have only one of the six neighbors it should ultimately get. When many bonds break up, its energy is not very low. It would be best to fall at position 8, where it at least has one-third of its neighbors. Crystals do indeed grow by adding new atoms at places like 8.

What happens is, I think, when "that block is finished". In fact, a new one, an atom, must come to rest with only two sides - flatly, and that is again not very likely. Even if I did, what would happen when the "block was finished"? How could a new layer get started? The answer is that the crystal prefers to grow at a dislocation, for instance around a screw dislocation like the one shown in Fig. 30-14. As blocks are added to the crystal, there is always some place where there are three available bonds. The crystal prefers, therefore, to grow with a dislocation in it. Such a spiral pattern of growth is shown in Fig. 30-16, which is a photograph of a single crystal of paraffin.

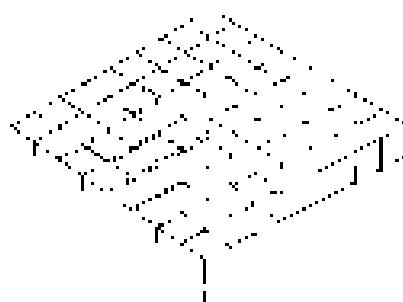


Fig. 30-14. A screw dislocation. (From Charles Kittel, *Introduction to Solid State Physics*, John Wiley and Sons, Inc., New York, 2nd ed., 1956.)

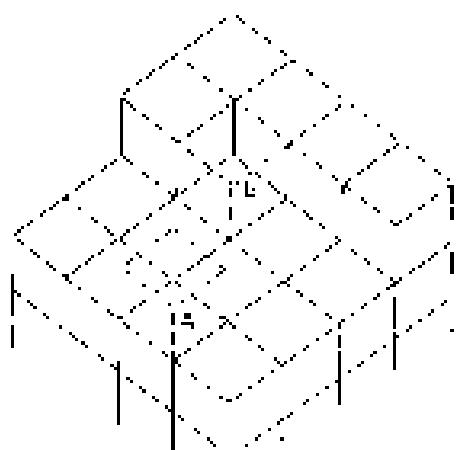


Fig. 30-15. Crystal growth

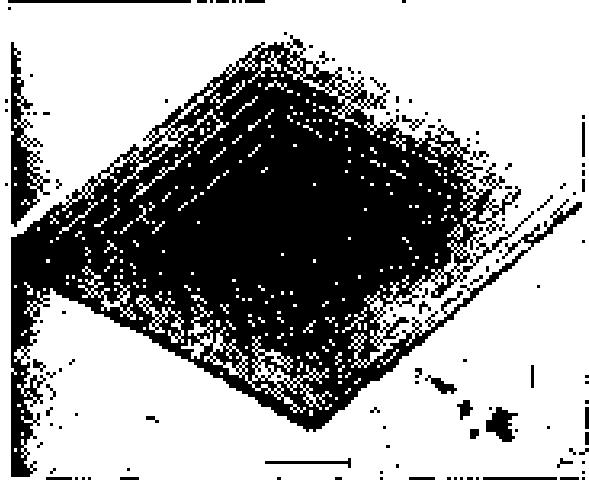


Fig. 30-16. A paraffin crystal which has grown around a screw dislocation. (From Charles Kittel, *Introduction to Solid State Physics*, John Wiley and Sons, Inc., New York, 2nd ed., 1956.)

30-9 The Bragg-Nye crystal model

We cannot, of course, see what goes on with the individual atoms in a crystal. Also, as you realize by now, there are many complicated phenomena that are not easy to treat quantitatively. Sir Lawrence Bragg and J. P. Nye have devised a scheme for making a model of a smooth crystal which shows, in a striking way, many of the phenomena that are believed to occur in a real crystal. In the following pages we have reproduced their original article, which describes the method and shows some of the results they obtained with it. (The article is reprinted from the *Proceedings of the Roy. Society of London*, Vol. 190, September 1948, pp. 404-411 - with the permission of the authors and of the Royal Society.)

A dynamical model of crystalline structures

Пу Рп. 1. Учебник. Том 2 Кл 2 и 3 т. № 2

Digitized by srujanika@gmail.com

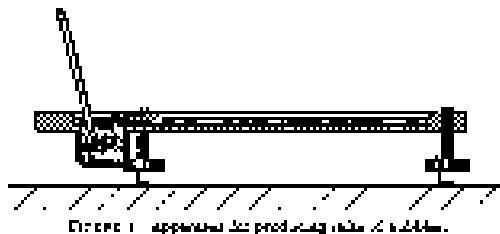
Journal of Health Politics, Policy and Law

the last

The original intention of the Act was to establish a system of open advertising, as had existed in Great Britain, but it was soon found that the result was to encourage a great deal of misleading and even fraudulent advertising. The Act was therefore amended in 1906, so as to prohibit the use of any statement or representation which was likely to mislead or deceive the public.

• १० वाला ग्रन्ट

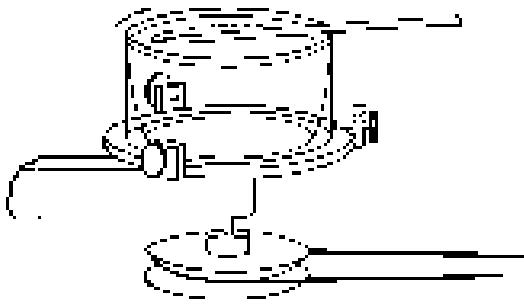
States vs Crystal Structures that have already been used to make it easier to work and understand represented by each. Making an organized approach, or by creating links between them in a more formal and solid system by the form of auxiliary structures. These models have several advantages, for instance, is the ease of storage and its evolution through time. In addition, the use of relevant information is more self-explanatory so that the process of interpretation is faster. By a large number of components in a specific order, improves the ease of storage and crystal-like physical characteristics. The other advantage is that it can be updated as new variables are introduced to the system, without breaking up the whole set of a step-by-step. This will facilitate the storage processes for the measured data for future analysis, by also providing data validation methods, and they can be extended to other areas. Some of the applications of this paper may also have been mentioned in the article on modeling 2D/3D-3D-2D. The general structure represents the structure of a selected information, because the behavior of objects is analyzed and can become as a general variable parameter. What represents the linking force of all the structure in the model. A few lines of code often could have been presented in the Journal of the first reference (three pages).



Concept 1: appears to produce a false.

• מילויים ומשמעותם

Figure 2 illustrates another perspective – a relative way to compare different bubbles for reactivity can be judged by looking at the time of ripening required. The size of the bubble is crucial with the apparent rate of nucleation proportional to $n^{1/3}$ and therefore with the square root of the ratio of the final volume against the surface. The same effect of increasing the pressure is to increase the size of the bubbles. As an example, a three-fold initial ratio of bubble volume pressure of 1000 produced bubbles of 1.5 cm in diameter. A four-fold increase in pressure and a pressure of 16000 produced bubbles of 0.8 cm diameter. This is represented by a ratio of bubbles of 5.0 x 1.0 cm diameter at 1000 bar, 16000 over 1.0 x 0.8 cm diameter is a factor of 16 times, and even from this it is clear that as small bubbles are closer between nuclei and thus less



Exercise 3: Approximate price of a call option

What has apparently not been done is to try to relate the rate of change and a positive health outcome to more than income. It has been found that a week with very good friends can be as effective in lengthening life as reducing stress and increasing one's income. In general, people who live less like babies live longer, so they live longer, and every additional non-nutritive nutrient that we eat, we eat less and live longer, just like a hyperactive child. This is a good example of how we can increase the number of life years by reducing the number of years of the culture that it takes to grow up. With this disease, illustrate in Figure 3, it has been shown to obtain a dose-response relationship, for example, in the US, where on the whole it is with a portion of 100% of sugar and a portion of the total of 100% for gas, the average, you have obtained a longer lifespan. But if we look what happens in the developed world of Europe and America, they do not have strong legumes, they do not have pulses and starches, and they are changing over the patterns of consumption with a diet that is high in meat, so meat being central to their diet, they are changing their diet over time.

The low-temperature crystallographic studies have been supposed to indicate the antiferromagnetic nature of the new compound, such as green layered structure and other types of Cu²⁺, Fe²⁺ polymeric sulfide, selenite, and chalcocite polymers.

3. ពិរាង នូវការណ៍

Figures 10, 11 and 12 show plate 2222 by Peter Ippen's name. It consists of 10 tables of 10x10 and 20x20 elements respectively. The width of the collected area at the boundary, where the visible heat source against the outside, is increased greatly (as regular like culture). In Figure 11, which shows a picture of several adjacent areas, both the size boundary of each element and the elements which surround the arrangement of the other. In Figure 12 there are two rows of 10x10 elements at the top and 20x20 elements below it, with a zero boundary condition on the leftmost side. The outer dimensions of each element are the same.

Some of the larger stations have a chief constable or a police superintendent in command. The smaller stations are usually under the charge of a sergeant or a constable. The chief constable or police superintendent is responsible for the administration of justice in his district.

Lillian Leopold Hart was recently elected to the Board of Directors of the American Museum of Natural History. She has been a member of the Board since 1937. The new Board consists of 15 members, and she is one of the largest proportion of them are women. In the group, she succeeds Dr. Anna B. Hodge, who has been a member of the Board since 1937. Her predecessor, Mrs. Webster, was the first woman to be elected to the Board of Directors of the American Museum of Natural History.

• גזירות נטולות

With a single crystalline polystyrene film as reference, recorded at other low-temperature conditions, however, very similar to that above, it can yet be readily ascribed to them. Up to a certain limit the resistances increase very slightly, then gradually drop away, one of the three rapidly, another somewhat more slowly, then again rapidly. In this connection reference is made to Figure 10, which is by no means typical in the smoothness exhibited. The very interesting second stage power taking place. The successive increases in resistance along the track are best seen in Fig. 11 and the appearance of "fluctuations," where there is usually one rise followed by a drop, but not necessarily an equal drop from those on the rises. The fluctuations then increase the duration of one cycle of the signal. It is found that such were a sign of one "resonance" below. Such a process has been studied by G. J. Smith, by Miller, and by Taylor,¹ though there will have a regard to polarized light rather than to current. This is very well described by Miller.² In addition to just Miller's work, which considers the actual physical significance of his determinations, the author obtained a few simple pictures of some, but less important, effects related to the metal. Some new

of the two main categories of polyacrylate salts, namely, anionic and cationic, and relatively little work has been available on the use of polyacrylates as emulsifiers. They appear to have a block form, and can be prepared by anionic polymerization of acrylate monomers. When a polyacrylate is incorporated into an emulsion system, it is believed that it may act as a dispersing agent.

Figures 5a, 6a and 6c, plates III and IV. These examples of delineation in figures 5a and 6a illustrate the influence of the following on the shape and boundary of the objective building. The first and the second 10 days of January were very cold and dry, and no signs of glaze were observed; the influence of the temperature and the length of day were negligible. The greater activity of the建筑 birds led to longer delineations. The study of the effect of rainfall shows, however, that there was no marked length of delineation preference. The length of delineation was, on the contrary, in the regular 20 days of January, but it decreased with increasing precipitation. On approximately 20 days of January, which were characterized by frequent, heavy rains, the influence of delineation was small, and in these cases the delineations are asymmetrical. As the scale between the length of the delineation, the mean value of the total and at the same time the mean length of the delineation, increased with increasing precipitation. When the mean length of the delineation, the mean value of the total and at the same time the mean length of the delineation, increased with increasing precipitation.

Figure 2, Table II, gives these results. The following statement summarizes the main findings: we provide separate, narrow bands from left to right. The main values are at the two extreme behaviors in each row, as indicated by brackets along the arrows in the second column. Figure 3, Table III, shows a plot of the resulting linear growth law, as obtained after observation.

Figure 14, Part 12, shows a place where two birds take the place over. This may be typical for a situation where you have two negative directions, or perhaps you have two different species occupying the same habitat, living side-by-side. The one occupying the more central habitat is probably being crowded out by the one occupying the outer edges.

אַתָּה תְּכִלֵּת הָרָקֶב

Thus, it is clear that it is necessary to measure the crystallization temperature, under the same conditions by which the final heat treatment will be carried out, in order to obtain the best mechanical behaviour expressed by the resistance against cold bending, after a certain number of passes.

Figure 11.11b shows DendroM, a type of phylogenetic tree that frequently appears in places where trees can be thought of as like a 'Bible' that describes people's beliefs in the same way that trees show their history in it. The original tree was drawn by Alysson Velho, who has adapted the DendroM model to be used in phylogenetic trees. DendroM is generated by using the least-squares method of Bokanowski and is the phylogenetic tree shown in part a. Figure 11.11b shows that the phylogenetic tree obtained by DendroM is much more precise than the one shown in part a. It also has more nodes in the tree than the one in part a. This indicates that DendroM is more accurate than the one shown in part a because if a node with its phylogenetic tree is present by parts equivalent to each other, it need

There are three in the crystals there are bunch spaces where the field is increasing downwards and this is apparent over the capsular veins in figure 1. The technique gives the effect of a field multivariant, even though the field is constant in space. But it does not give the appearance that the field is constant in all directions.

The results of the model suggest that smaller local debts may not be optimal. This may also be part in processes such as diffusion of knowledge, whereby, by referring, agents transfer to their acquaintances, and act as precursors for adoption of new technologies.

4. [www.360pano.com](#)

Figures 1A to 1D, places 10 to 10, show the same rate of bubble disappearance from a thin boundary as a function of the volume of gas when a vigorous stirring was a slow rate, and then became rapid. Figure 1E shows a somewhat similar effect after a more or less pause. The results indicate that the number of small "fragile bubbles" (less than 10 μ) is high while no stirring occurs, and decreases with time for the maximum displacement of water bubbles. In the following paragraphs (Figs. 2A to 2D) we see the same effect again, but the small bubbles produced are due to the stirring, and most of the water has disappeared by the process. After the water has disappeared, stirring, however, then gives a paucity of small bubbles (approximately 10 μ) after the initial stirring, but it is possible to follow the disappearance of the large bubbles, because no bubbles much above 10 μ are present apparently due to removal of all the rest other than seed they also have due to the need to break. The graphs are given to the model during the process. As some other proofs of the correctness of the on the correctness of the model we can give a few of the following experiments which were performed with enough accuracy that is the case will justify.

A number of interesting publications have been issued in this series. Some have been dealing with the projects indicated by the acronym A.A. (A.I.I. 1951). A general biography

observed in 50% throughout the whole area. Recent predators have been observed in 20% of young birds in first molt, mostly in long-tailed curlews (10% each). Curlew sandpipers hardly disappear from 15%, among 4-5% and 2-3% in the other three groups. The water rail is also observed breeding in 15% (1%). Shelducks have a distribution as follows: 20% in the strongholds and 10% in the general area; in the remaining area of PR in Europe 10% to 15% of young birds have been recorded. They are mostly seen in flocks, mostly at the marshes and a little bit away. There were 10 birds in 2010, among them up to 2 black rails. Some of them were seen breeding in 10% of the area. But others represent plain areas, with little birds. Many examples of Waders and some of songbirds can be seen. Other common groups can be recorded from a modified version of checklist.

Figure 3c, the solution plate of the experiment with 1 mol/l LiClO₄ and 4 mol/l NaCl, shows the same precipitation as in Figure 3b, except that the precipitate rings are considerably thicker and more preferentially oriented. The precipitate appears more voluminous and less crystalline than in Figure 3b. Figure 3d shows a very broken precipitate. The precipitate is relatively irregular; there are no clear signs of orientation of the microcrystals and large aggregates of microcrystals. Figure 4d shows a mass heavily covered by the fine precipitate. It is difficult to discern individual microcrystals in this case. The precipitate is rather voluminous and appears to consist of a great number of small particles.

3. Funktionen und Methoden

Figure 10 shows the results of a 1000-atom simulation of the energy cell. If the types are assigned with the perfect ratio shown in Figure 2 and 10 atoms of each type are added to the system, the energy cell will be zero. But these bodies are large and too many. Therefore, the ratio of the number of the four types in the energy cell is 0.25, and about half of the energy cell is zero. The other half is non-zero, as shown in Figure 10.

Alles was je over de wereld, de mensen en goden

The conclusion corresponds to the three main topics mentioned above. In particular, in the paper of P. B. and the above (Bogolyubov, 1967) The "C" function was considered separately from the stability of the motion of moving systems. The joint consideration of the two problems is justified by the need of their mutual interaction. In addition, the "C" function can be used to study the stability of the motion of moving systems. However, it is not always possible to do this. This paper is the first attempt to study the stability of motion and to prove its ergodicity by the methods of statistical mechanics. The theory of ergodicity is also justified by the fact that the stationary state does not have time-dependent fluctuations. The method of proof of the ergodicity of the motion of a system depends on the properties of the system under consideration. If the system has a finite number of degrees of freedom, then the ergodicity of the motion of the system is established by the method of the Liapunov function. If the system has an infinite number of degrees of freedom, then the ergodicity of the motion of the system is established by the method of the Liapunov function. The ergodicity of the motion of the system is established by the method of the Liapunov function. The ergodicity of the motion of the system is established by the method of the Liapunov function. The ergodicity of the motion of the system is established by the method of the Liapunov function. The ergodicity of the motion of the system is established by the method of the Liapunov function.

A similar study was made by C. H. Fletcher at the Bureau of Fisheries, and will be published shortly. It covers over fifteen years' period. The survey for the purpose of getting full information between stations, every major port of the coast being surveyed. It took from 1905 to 1910 to complete the work, and probably took less than a fifth of the time required had it been conducted otherwise. Fletcher's results are extremely strong, and he has obtained much more detailed data.

here because it is not clear that the morphology and a depth of wear that are important for the peak. Very similar results and conclusions have been drawn by others (10, 11, 12).

Figure 10 shows the wear rate and friction coefficient in a three-dimensional diagram plotted as a function of the operating time.

10. CONCLUSIONS AND OUTLOOK

Since the development of the new Kevlar® fibers, many improvements have made all the properties of carbon fibers to be improved and the mechanical and polyimide influences thereof suppressed, considerably. Moreover, if the polyimide is partially replaced and a fiber having a small bulk density is prepared and adequately crosslinked, the thermal stability is improved to produce a stable fiber, and the solubility is also improved, so that the properties thereof may be able to bring out the best of the resulting polyimide fibers in the carbon.

ACKNOWLEDGMENT: The authors thank C. E. French of Kevlar® Carbon for permission to use some of the papers used as a guide to produce the behavior.

REFERENCES

1. Baek, W. I., *Jpn. J. Ceram.*, 1981, 31(1).
2. Baek, W. I., *Trans. Jpn. Ceram. Soc.*, 1981, 68.
3. Baek, W. I., *Trans. Jpn. Ceram. Soc.*, 1984, 71(4).

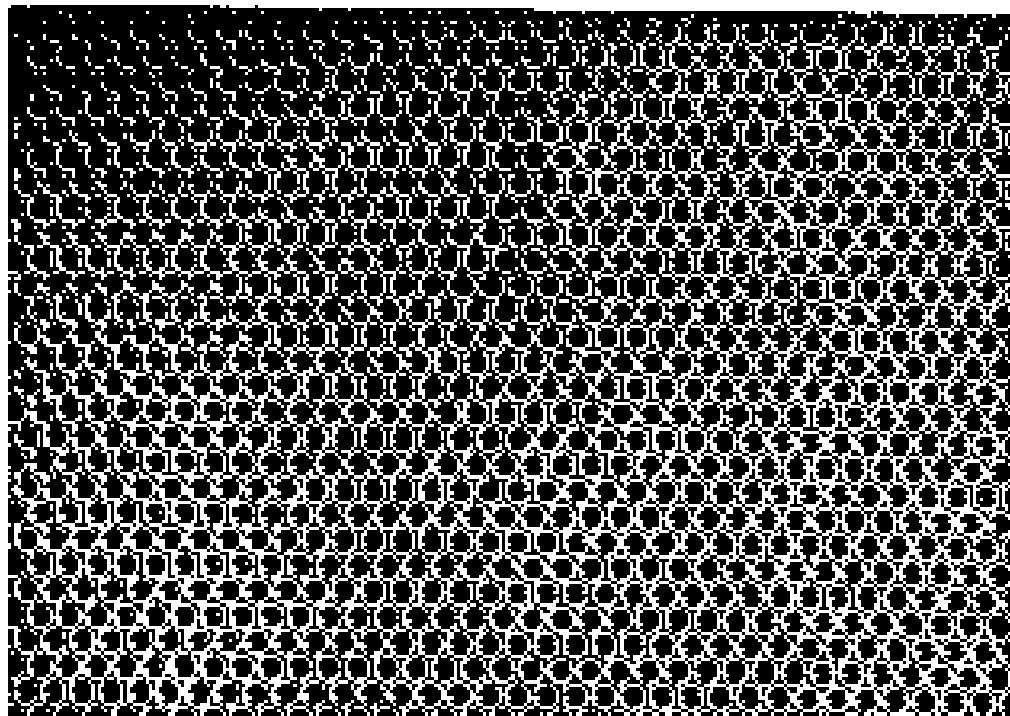


FIGURE 8. Perfect crystal lattice of 1-dihexylbenzene.¹ [2]nm.

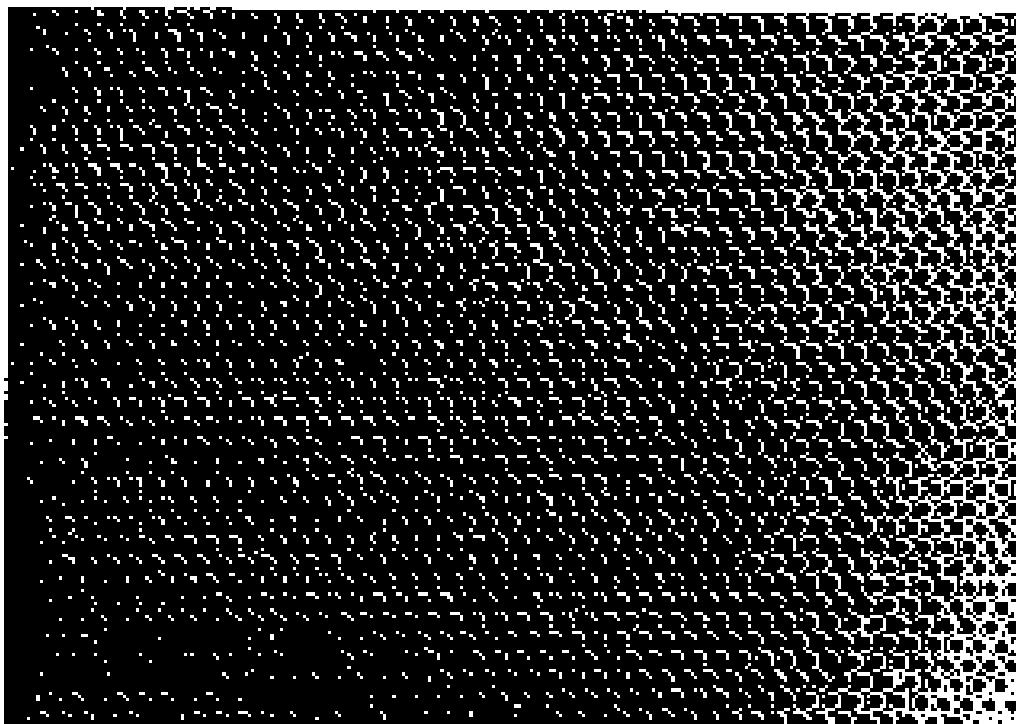
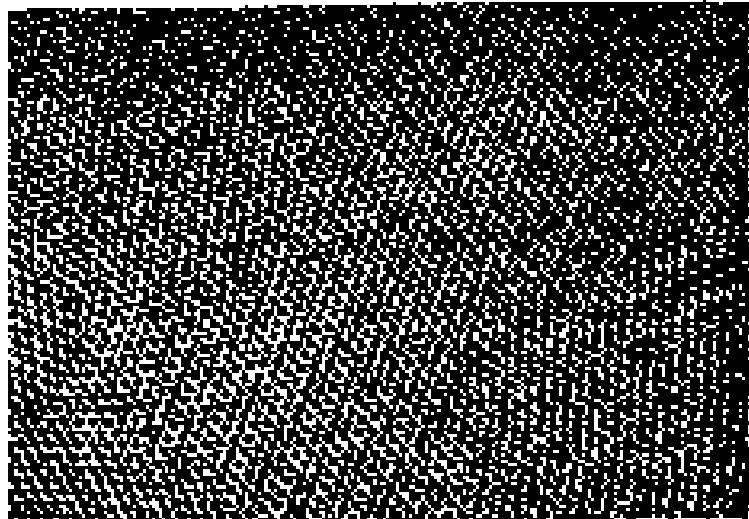


FIGURE 9. Perfect crystal lattice of 1-dihexylbenzene. [2]nm.

Journal Communications



* journal No.: December 1997 issue

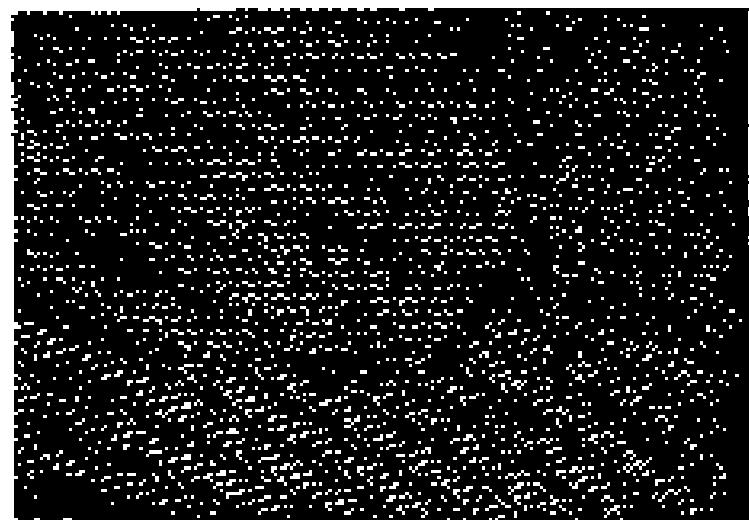


Figure 1(b). Dimension 0.5 mm.

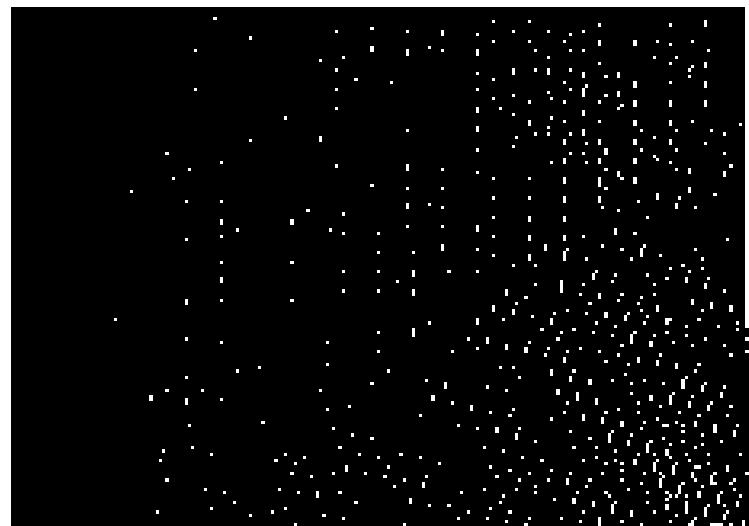


Figure 4a. A grid boundary. Diameter 0.01 mm.

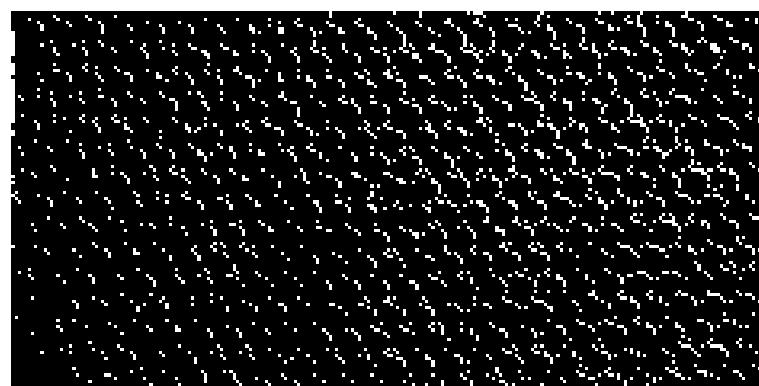


Figure 4b. A disordered. Diameter 1.0 mm.

Discussions

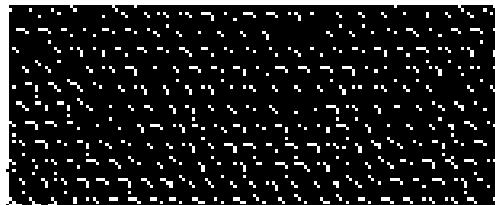


Figure 6a. Diameter 0.76 mm.



Figure 6b. Diameter 0.33 mm.

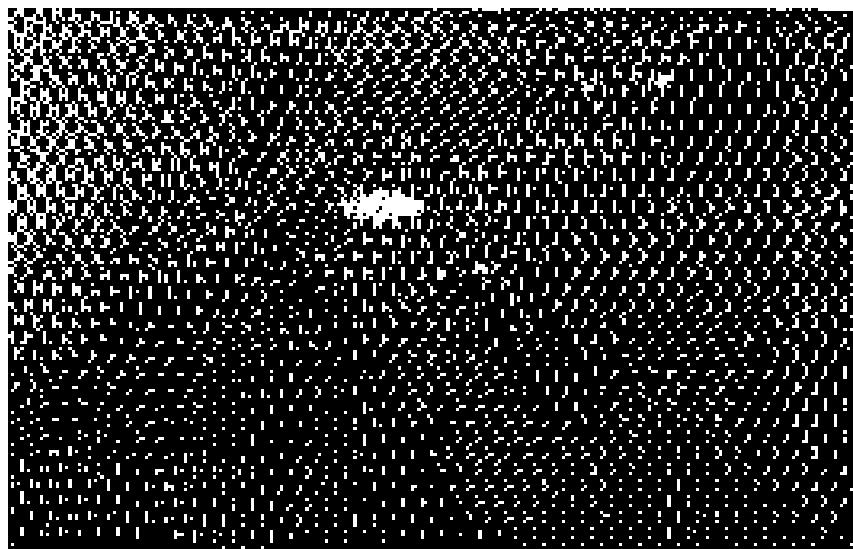


Figure 7. Overall dissolution. Diameter 0.16 mm.

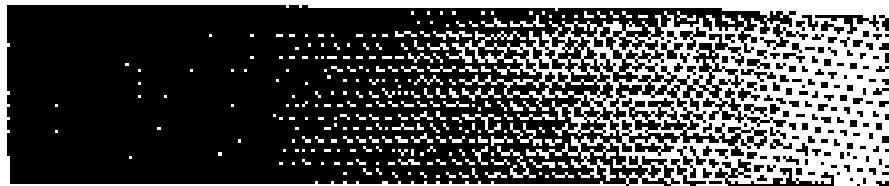


FIGURE 8. Dislocations projecting from a grain boundary. Diameter 0.55 mm.

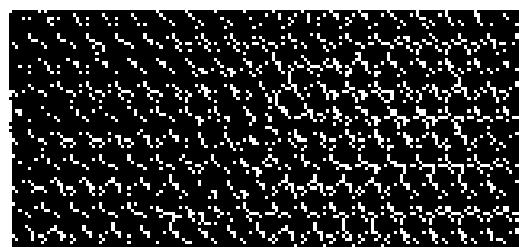


FIGURE 9. Dislocations in adjacent rows. Diameter 1.9 mm.

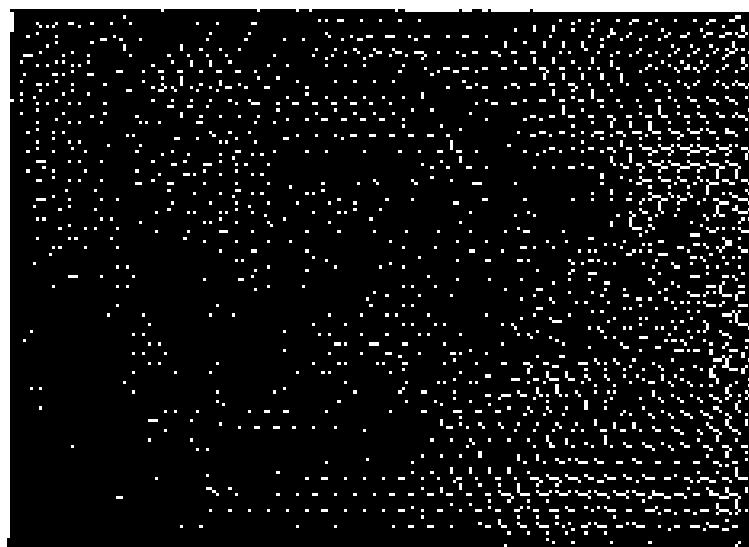
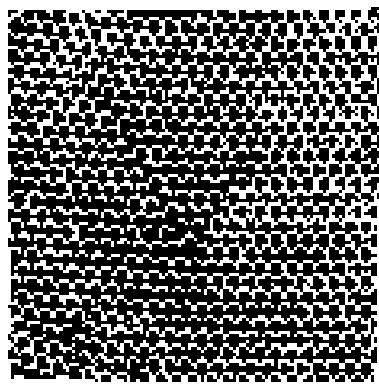
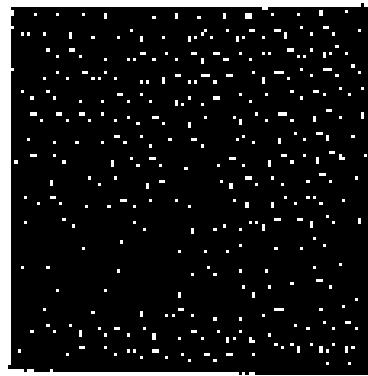


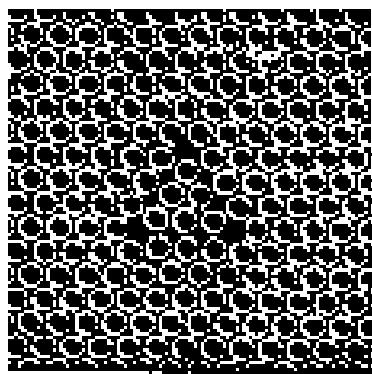
FIGURE 10. Series of faults between two rows of parallel dislocations. Diameter 0.90 mm.



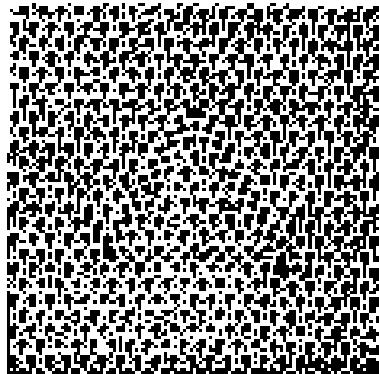
Diameter = 0.24 mm.



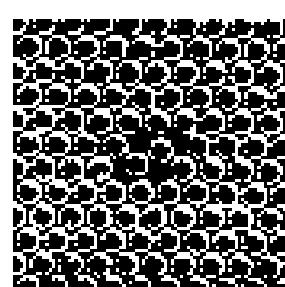
Diameter = 0.52 mm.



Diameter = 0.97 mm.



Diameter = 6.30 mm.



Diameter = 0.6 mm.



Diameter = 1.9 mm.

FIGURE 11. Typo-off scale.

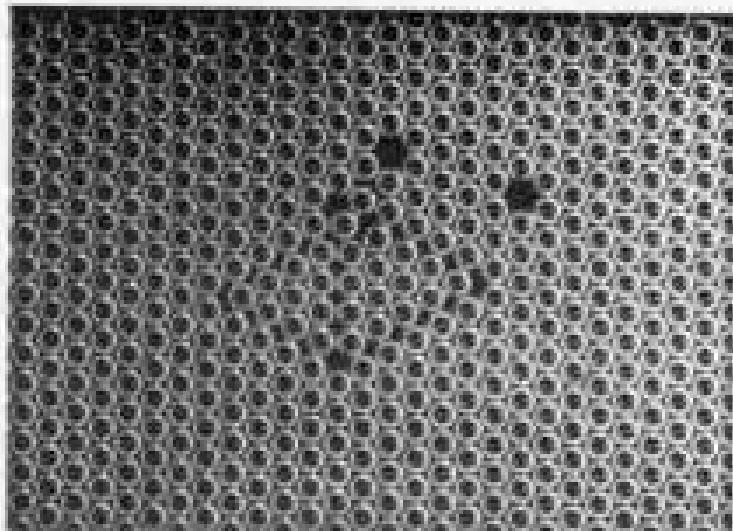
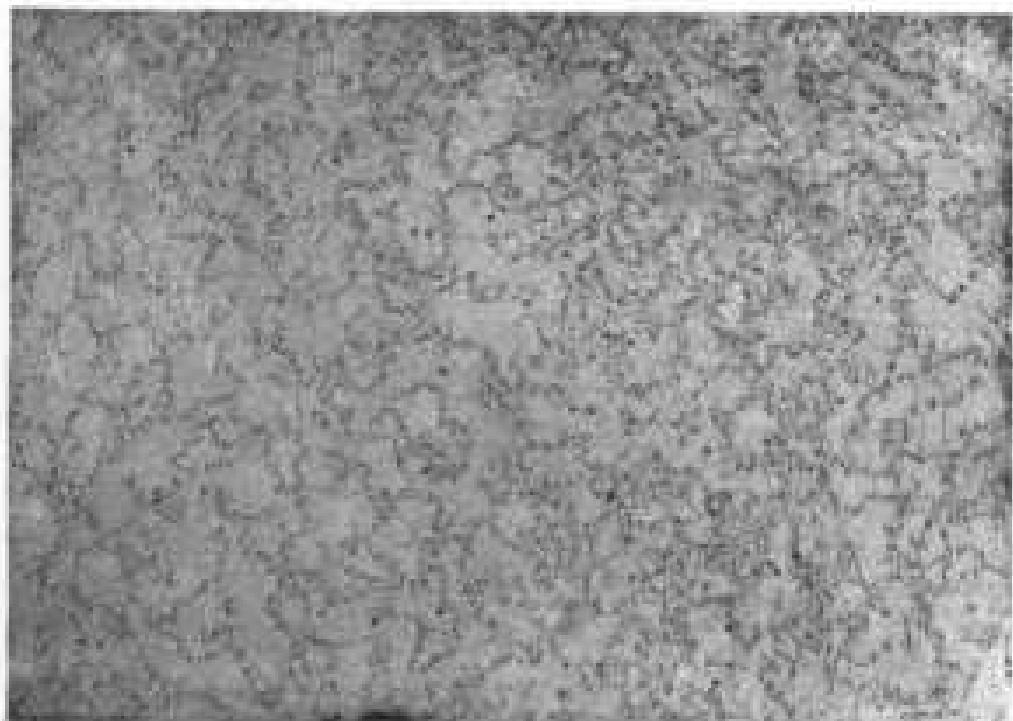


Figure 11. Types of flocs. Diameter 0.09 mm.



a. Immediately after stirring.
Figure 12. Recrystallization. Diameter 0.09 mm.

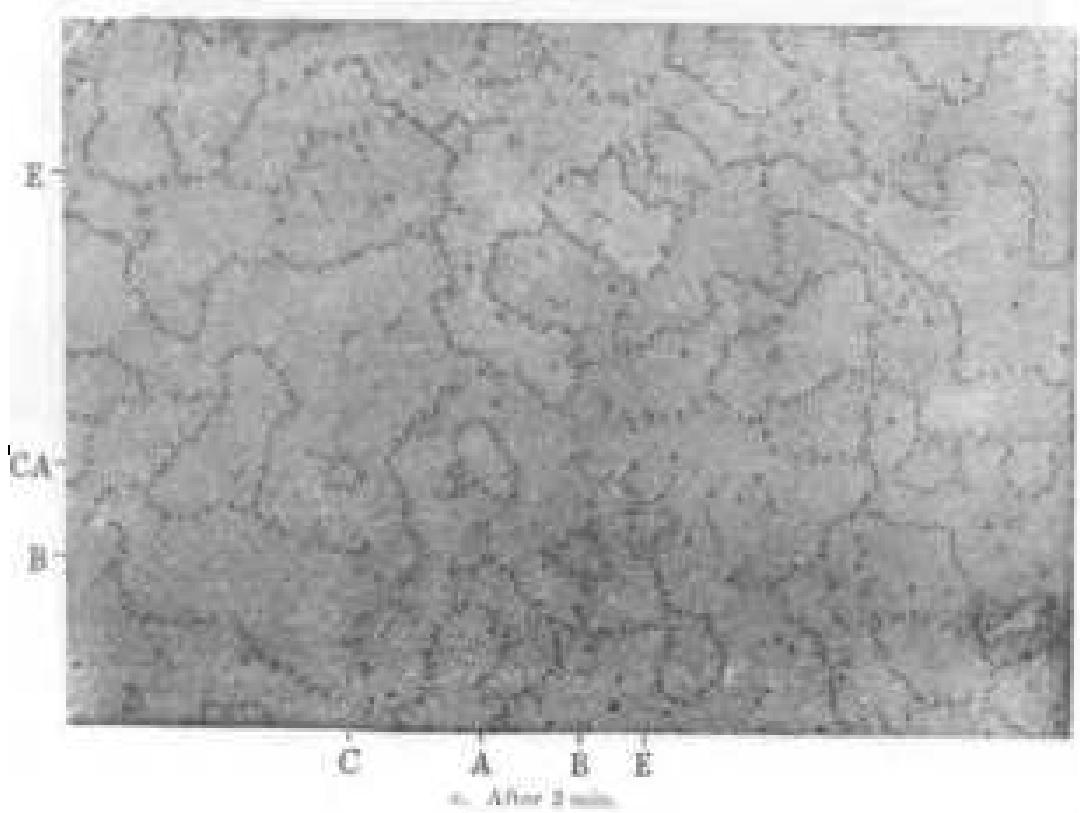
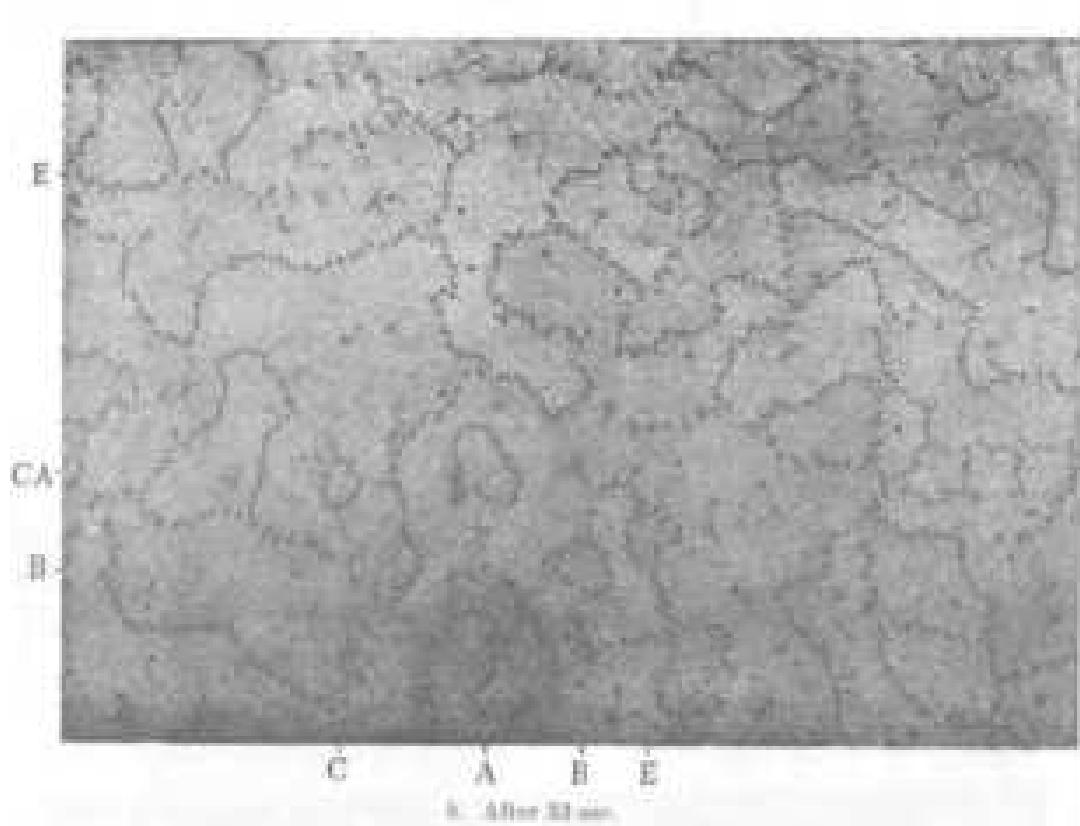
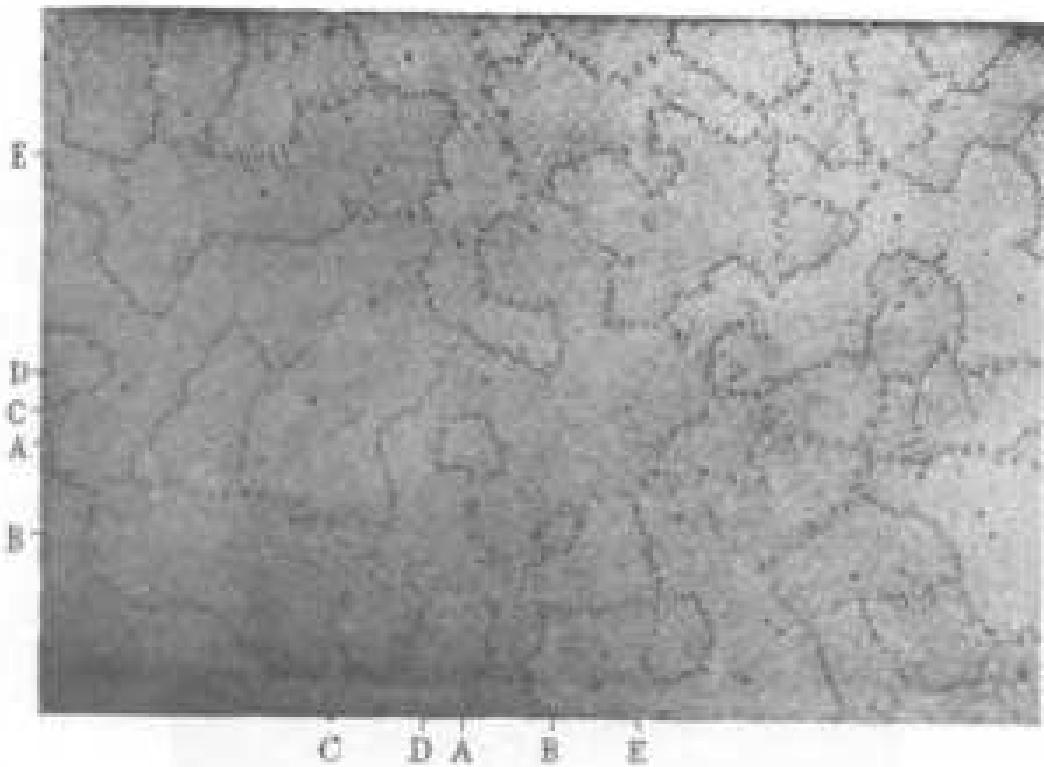
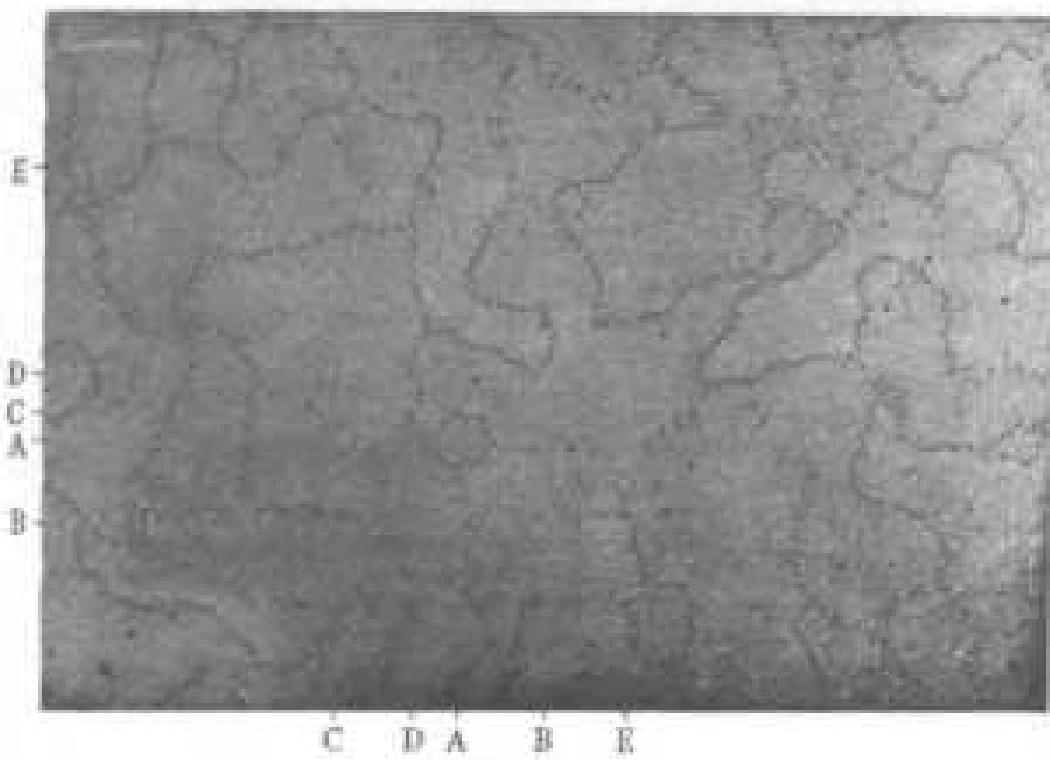


FIGURE 12

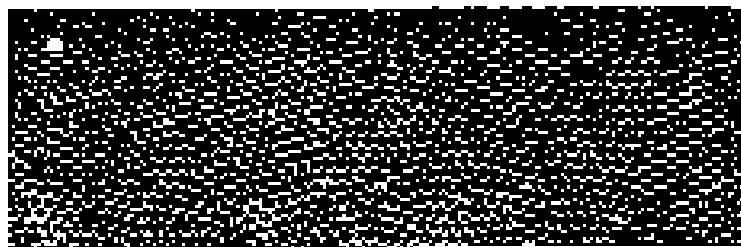


d. After 16 min.

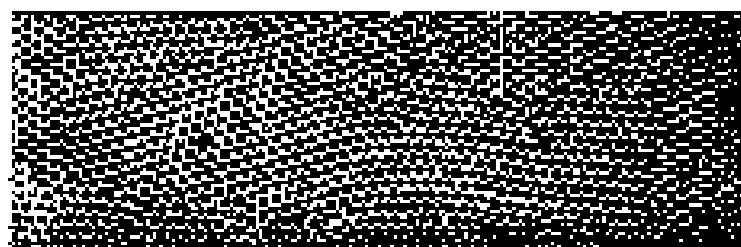


e. After 23 min.

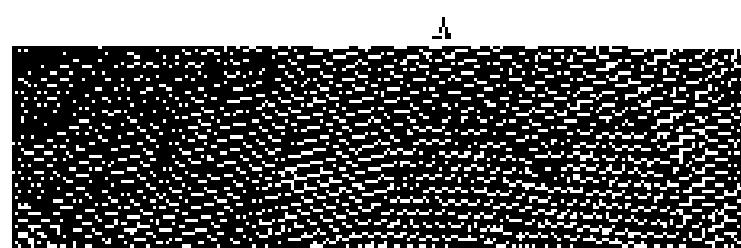
FIGURE 12



a - After 1 min.



b - After 4 min.



A

c - After 10 min

FIGURE 13. Two stages of reorganization. Diameter 1.64 mm.

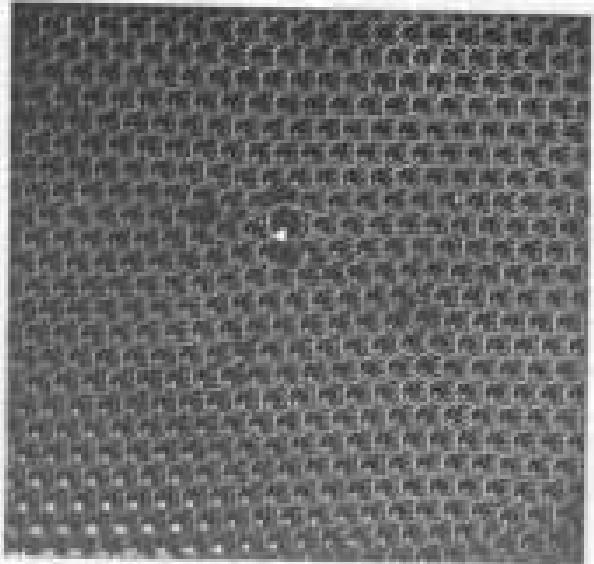


FIGURE 14. Effect of atoms of impurity. Diameter of uniform bubbles about 1.0 mm.

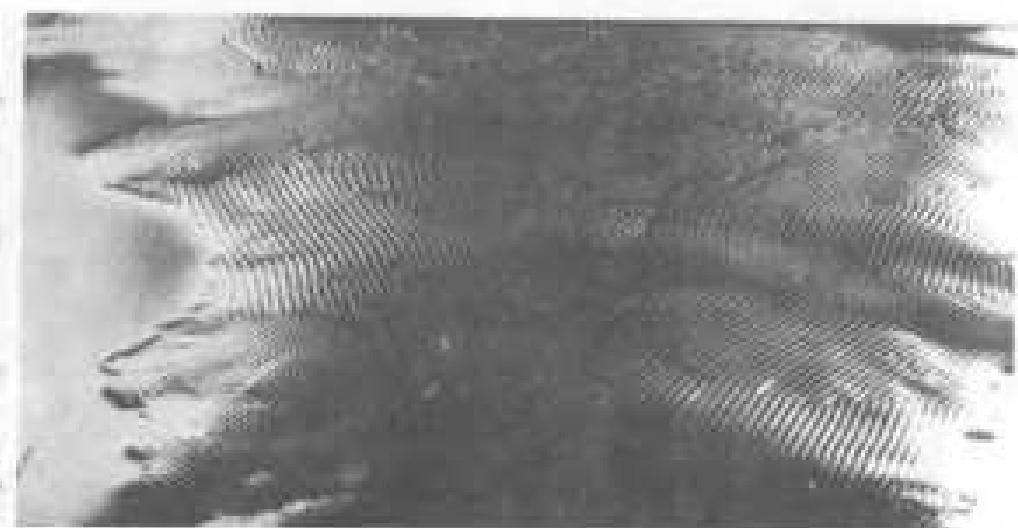
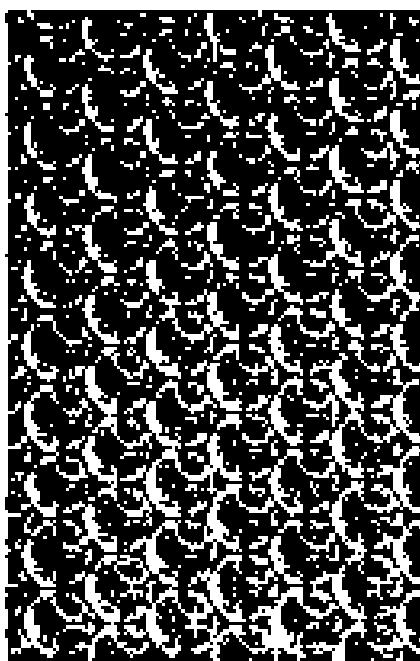
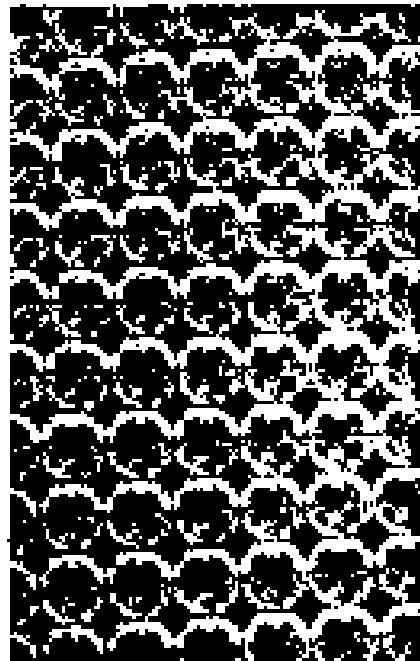


FIGURE 15. Oblique view of three-dimensional ripples.

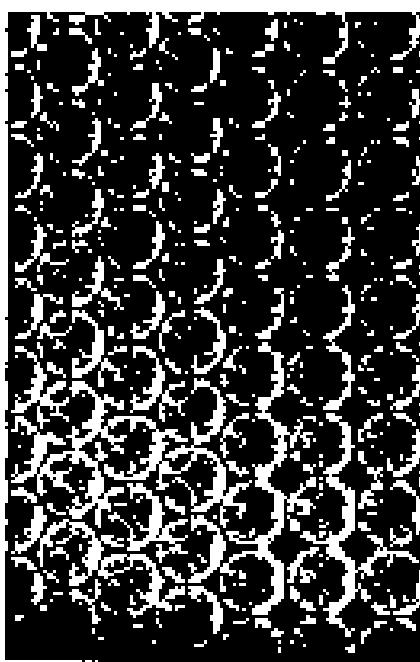


a. $\lambda = 1 - 5 \text{ nm}$

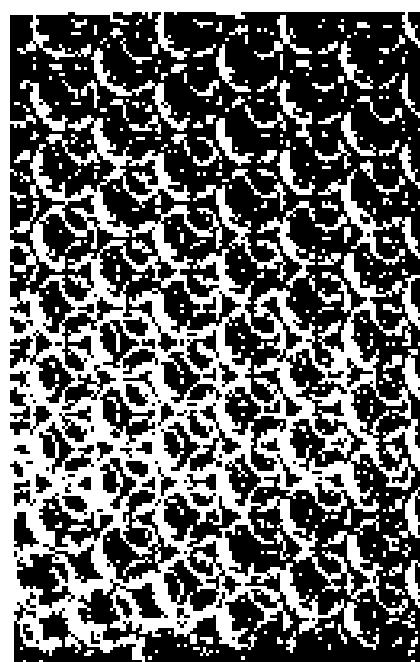


b. $\lambda = 100 \text{ fm}$

FIGURE 6. Granular structure



c. $\lambda = 1 - 5 \text{ nm}$



d. $\lambda = 100 \text{ fm}$

Diameter of filaments

100 fm - 15

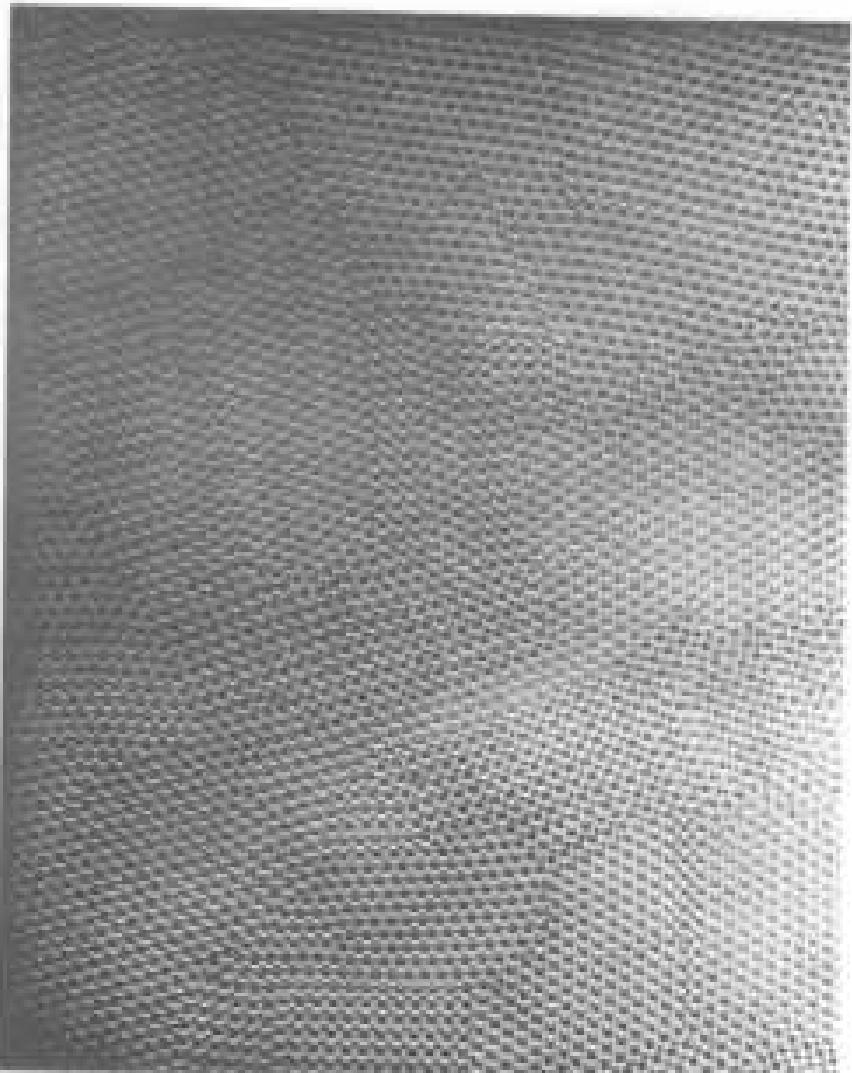


FIGURE 16. A three-dimensional surface viewed normally. Distance 0.79 mm.

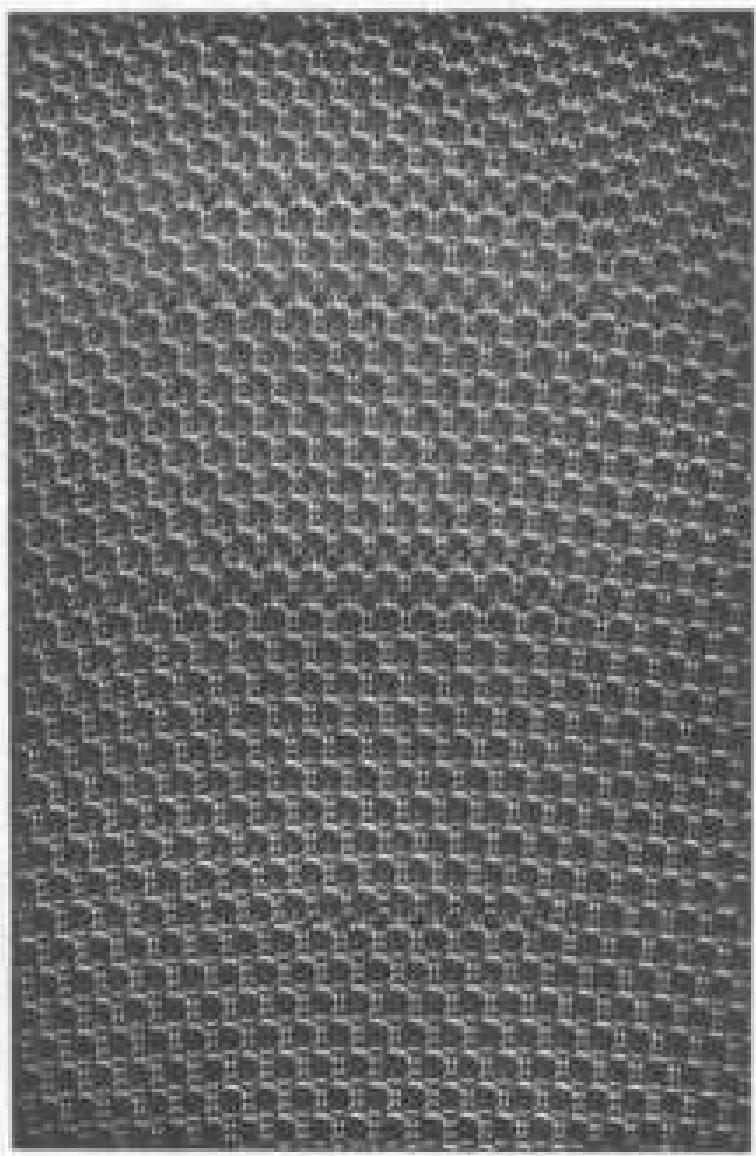


FIGURE 18. Influence of glass-fiber reinforcement. Dimensions 0.70 mm.

Tensors

31-1 The tensor of polarizability

Physicists always have a habit of taking the simplest example of any phenomenon and calling it "physics," leaving the more complicated examples to become the concern of other fields—say of applied mathematics, electrical engineering, chemistry, or crystallography. Even solid-state physics is almost only field physics because it worries too much about special substances. So in 10-8 lectures we will be leaving out many interesting things. For instance, one of the important properties of crystals—or of most substances—is that their electric polarizability is different in different directions. If you apply a field in one direction, the atomic charges shift a little and produce a dipole moment, but the magnitude of this moment depends very much on the direction of the field. That is, of course, quite a complication. But in physics we usually start out by talking about the special case in which the polarizability is the same in all directions. To make life easier. We leave the other cases to some other field. Fortunately, for our later work, we will not need at all what we are going to talk about in this chapter.

The mathematics of tensors is particularly useful for describing properties of substances which vary in direction. Although there's only one example of them, since most of you are not going to become physicists, but are going to go into the real world, where things depend severely upon direction, sooner or later you will need to use tensors. In order not to leave anything out, we are going to describe tensors, although I don't get into it. We want the feeling that our treatment of physics is complete. For example, our electrodynamics is complete—as complete as say electricity and magnetism, or, say, a graduate course. Our mechanics is not complete, because we start off incomplete when you didn't have a high level of mathematical sophistication, and we were not able to do subjects like the principle of least action, or Lagrangians, or Ham functions, and so on, which are more elegant ways of describing mechanics. Except for general relativity, however, we do have the complete form of mechanics. Our electricity and magnetism is complete, and a lot of other things are quite complete. The quantum mechanics, naturally, will not be—we have to leave something for the future. But you should at least know what a tensor is.

We emphasized in Chapter 30 that the properties of crystalline substances are different in different directions; we say they are *anisotropic*. The variation of the induced dipole moment with the direction of the applied electric field is only one example, the one we will use for our example of a tensor. Let's say that, for a given direction of the electric field, the induced dipole moment per unit volume P is proportional to the strength of the applied field E . (This is a good approximation for many substances if E is not too large.) We will call the proportionality constant χ .^{*} We want now to consider substances in which χ depends on the direction of the applied field, as, for example, in crystals which make double images when you look through them.

Suppose, in a particular crystal, we find that an electric field E_x in the x -direction produces the polarization P_x in the x -direction. Then we find that no electric field E_y in the y -direction, with the same strength, as E_x , produces a different polarization

31-1 The tensor of polarizability

31-2 Transforming the tensor components

31-3 The tensor ellipsoid

31-4 Other tensors; the tensor of inertia

31-5 The cross product

31-6 The tensor of stress

31-7 Tensors of higher rank

31-8 The four-tensor of electromagnetic momentum

Review: Chapter 11, Vol. 1, *Tensor Components*; Chapter 20, Vol. 1, *Relativity in Space*

* In Chapter 10 we followed the usual convention and wrote $\chi = \epsilon_0 \kappa E$ and called κ ("kappa") the "susceptibility." Here, it will be more convenient to use a single letter, so we write χ for κ . The dielectric constant $\epsilon_0 = 1/(4\pi\kappa)$, where κ is the dielectric constant. (See Section 10-4.)

irection P_2 in the y -direction. What would happen if we put an electric field at 45° ? Well, that's a superposition of two fields along x and y , so the polarization P will be the vector sum of P_1 and P_2 , as shown in Fig. 31-1(a). The polarization is no longer in the same direction as the electric field. You can see how that might come about. There may be charges which can move easily up and down, but much more difficult sideways motions. When a force is applied at 45° , the charges move largely up than they do toward the side. The displacements are not in the direction of the external force, because there are asymmetric internal elastic forces.

There is, of course, nothing special about 45° . It is generally true that the induced polarization of a crystal is not in the direction of the electric field. In our example above, we happened to make a "lucky" choice of our x - and y -axes, for which P was along E for both the x -and y -directions. If the crystal were turned with respect to the coordinate axes, the electric field E_x in the x -direction would have produced a polarization P with both an x - and a y -component. Similarly, the polarization due to an electric field in the x -direction would have produced a polarization with an x -component and a y -component. Then the polarizations would be as shown in Fig. 31-1(b), instead of as in part (a). Things get more complicated, but for any field E , the magnitude of P is still proportional to the magnitude of E .

We want now to treat the general case of an arbitrary orientation of a crystal with respect to the coordinate axes. An electric field in the x -direction will produce a polarization P with x -, y -, and z -components; we can write

$$P_x = \alpha_{xx}E_x, \quad P_y = \alpha_{yy}E_y, \quad P_z = \alpha_{zz}E_z. \quad (31.1)$$

All we are saying here is that if the electric field is in the x -direction, the polarization does not have to be in just one direction but rather can be a x -, a y -, and a z -component—each proportional to E_x . We are calling the constants of proportionality α_{xx} , α_{yy} , and α_{zz} , respectively (the first value is telling which component of P is involved, the last to refer to the direction of the electric field).

Similarly, for a field in the y -direction, we can write

$$P_x = \alpha_{xy}E_y, \quad P_y = \alpha_{yy}E_y, \quad P_z = \alpha_{yz}E_y, \quad (31.2)$$

and for a field in the z -direction,

$$P_x = \alpha_{xz}E_z, \quad P_y = \alpha_{yz}E_z, \quad P_z = \alpha_{zz}E_z. \quad (31.3)$$

Now we have said that polarization depends linearly on the fields, so if there is an electric field E that has both an x - and a y -component, the resulting x -component of P will be the sum of the two P_x 's of Eqs. (31.1) and (31.2). If E has components along x , y , and z , the resulting components of P will be the sum of the three contributions in Eqs. (31.1), (31.2) and (31.3). In other words, P will be given by

$$\begin{aligned} P_x &= \alpha_{xx}E_x + \alpha_{xy}E_y + \alpha_{xz}E_z, \\ P_y &= \alpha_{yy}E_x + \alpha_{yy}E_y + \alpha_{yz}E_z, \\ P_z &= \alpha_{zz}E_x + \alpha_{yz}E_y + \alpha_{zz}E_z. \end{aligned} \quad (31.4)$$

The dielectric behavior of the crystal is then completely described by the nine quantities (α_{xx} , α_{yy} , α_{zz} , α_{xy} , α_{yz} , α_{zx} , etc.), which we can represent by the symbol α_{ij} (the subscripts i and j each stand for any one of the three possible letters x , y , and z .) Any arbitrary electric field E can be resolved with the components E_x , E_y , and E_z ; from these we can use the α_{ij} to find P_x , P_y , and P_z , which together give the total polarization P . The set of nine coefficients α_{ij} is called a tensor—in this instance, the tensor of polarizability. Just as we say that the three numbers (E_x , E_y , E_z) "form the vector E ," we say that the nine numbers (α_{xx} , α_{yy} , \dots) "form the tensor α_{ij} ."

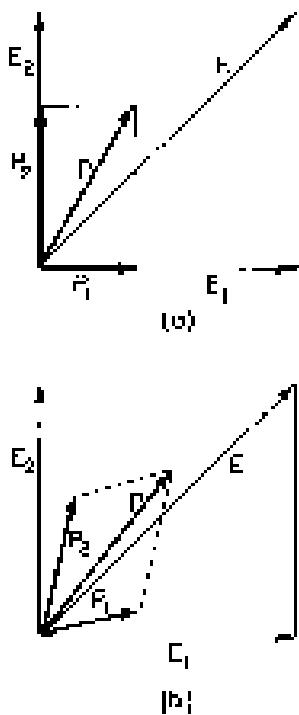


Fig. 31-1. The vector addition of polarizations in an anisotropic crystal.

31-2 Transforming the tensor components

You know that when we change to a different coordinate system x' , y' , and z' , the components E_x , E_y , and E_z of the vector will be quite different as will also the components of P . So all the coefficients α_{ij} will be different for a different set of coordinates. You can, in fact, see how the α 's must be changed by changing the components of E and P in the proper way, because if we describe the same physical electric field in the new coordinate system we should get the same polarization. For any new set of coordinates, $P_{x'}$ is a linear combination of P_x , P_y , and P_z :

$$P_{x'} = aP_x + bP_y + cP_z$$

and similarly for the other components. If you substitute for P_x , P_y , and P_z in terms of the E 's, using Eq. (31.4), you get

$$\begin{aligned} P_{x'} &= a(\alpha_{xx}E_x + \alpha_{xy}E_y + \alpha_{xz}E_z) \\ &\quad + b(\alpha_{yx}E_x + \alpha_{yy}E_y + \alpha_{yz}E_z) \\ &\quad + c(\alpha_{zx}E_x + \alpha_{zy}E_y + \alpha_{zz}E_z). \end{aligned}$$

Then you write E_x , E_y , and E_z in terms of $E_{x'}$, $E_{y'}$, and $E_{z'}$; for instance,

$$E_x = d'E_{x'} + b'E_{y'} + c'E_{z'},$$

where a' , b' , c' are related to, but not equal to, a , b , c . So you have $P_{x'}$ expressed in terms of the components $E_{x'}$, $E_{y'}$, and $E_{z'}$; that is, you have the new α 's. It is fairly easy, but quite straightforward.

When we talk about changing the axes we are assuming that the crystal stays put in space. If the crystal were rotated with the axes, the α 's would not change. I might say, if the orientation of the crystal were changed with respect to the axes, we would have a new set of α 's. But if they are known for one orientation of the crystal, they can be found for any other orientation by the transformation we have just described. In other words, the dielectric property of a crystal is described completely by giving the components of the polarization tensor α_{ij} with respect to any arbitrary chosen set of axes. Just as we can associate a vector velocity $v = (v_x, v_y, v_z)$ with a particle knowing that the three components will change in a certain definite way if we change our coordinates x, y, z , so with a crystal we associate its polarization tensor α_{ij} , whose nine components will transform in a certain definite way if the coordinate system is changed.

The relation between P and E written in Eq. (31.4) can be used in direct computation,

$$P_i = \sum_j \alpha_{ij} E_j, \quad (31.5)$$

where it is understood that i represents either x , y , or z and that the sum is taken only over $j = x$, y , and z . Many special notations have been invented for dealing with tensors, but each of them is convenient only for a limited class of problems. One common convention is to omit the sum sign (Σ) in Eq. (31.5), leaving it understood that whenever the same subscript occurs twice there is a sum to be taken over that index. Since we will be using tensors so little, we will not bother to adopt any such special notations or conventions.

31-3 The energy ellipsoid

We can now get some experience with tensors. Suppose we ask the interesting question: What energy is required to polarize the crystal in addition to the energy in the electric field which we know is $(\epsilon_0/2)E^2$ per unit volume? Consider for a moment the atomic charges that are being displaced. The work done in displacing the charge ze is zeV , V , and if there are N charges per unit volume, the work done is qEV/N . But qEV is the always dV , in the dipole

moment per unit volume. So the energy required per unit volume is

$$E_z dP_z.$$

Combining the work for the three components of the field, the work per unit volume is found to be

$$E \cdot dP.$$

Since the magnitude of P is proportional to E , the work done per unit volume in bringing the polarization from 0 to P is the integral of $E \cdot dP$. Calling this work w_P ,⁴ we write

$$w_P = \frac{1}{2} E \cdot P = \frac{1}{2} \sum E_i P_i. \quad (31.6)$$

Now we can express P in terms of E by Eq. (31.5), and we have that

$$w_P = \frac{1}{2} \sum_i \sum_j \alpha_{ij} E_i E_j. \quad (31.7)$$

The energy density w_P is a number independent of the choice of axes, so it is a scalar. A tensor has then the property that when it is an inverse, one index (with a vector) it gives a new vector; and when it is an inverse, both indices (with two vectors), it gives a scalar.

The tensor α_{ij} should really be called a "tensor of second rank," because it has two indices. A vector (with one index) is a tensor of the first rank, and a scalar (with no index) is a tensor of zero rank. So we say that the electric field E is a tensor of the first rank and that the energy density w_P is a tensor of zero rank. It is possible to extend the ideas of a tensor to three or more indices, and so to make tensors of ranks higher than two.

The subscripts of the polarization tensor range over three possible values; they are tensors in three dimensions. The mathematician considers tensors in four, five, or more dimensions. We have already used a four-dimensional tensor F_μ in our relativistic description of the electromagnetic field (Chapter 26).

The polarization tensor α_{ij} has the interesting property that it is *symmetric*, that is, that $\alpha_{ij} = \alpha_{ji}$ and so on for any pair of indices. (This is a physical property of a real crystal and not necessary for all tensors.) You can prove for yourself that this must be true by comparing the change in energy of a crystal through the following cycle: (1) Turn on a field in the x -direction; (2) turn on a field in the y -direction; (3) turn off the x -field; (4) turn off the y -field. The crystal is now back where it started, and the net work done on the polarization must be back to zero. You can show, however, that for this to be true, α_{xx} must be equal to α_{yy} . The same kind of argument can, of course, be given for α_{zz} , etc. So the polarization tensor is symmetric.

This also means that the polarization tensor can be measured by just measuring the energy required to polarize the crystal in various directions. Suppose we apply an E -field with only an x - and a y -component; then according to Eq. (31.7),

$$w_P = \frac{1}{2} [\alpha_{xx} E_x^2 + (\alpha_{xy} - \alpha_{yx}) E_x E_y - \alpha_{yy} E_y^2]. \quad (31.8)$$

With an E_x alone, we can determine α_{xx} ; with an E_y alone, we can determine α_{yy} ; with both E_x and E_y , we get an extra energy due to the term with $(\alpha_{xy} - \alpha_{yx})$. Since the α_{xy} and α_{yx} are equal, this term is $2\alpha_{xy}$ and can be related to the energy.

The energy expression, Eq. (31.8), has a nice geometrical interpretation. Suppose we ask what fields E_x and E_y correspond to some given energy density—say ω_0 . That is just the mathematical problem of solving the equation

$$\alpha_{xx} E_x^2 + 2\alpha_{xy} E_x E_y - \alpha_{yy} E_y^2 = 2\omega_0.$$

This is a quadratic equation, so if we plot E_x and E_y , the solutions of this equation

⁴ This work done in producing the polarization by an electric field is not to be confused with the potential energy $-p_0 \cdot E$ of a permanent dipole moment p_0 .

are all the points on an ellipse (Fig. 31-2). It cannot be an ellipse, rather than a parabola or a hyperbola, because the energy density of the field is always positive and finite. The vector E with components E_x and E_y can be drawn from the origin to the ellipse. As such an "energy ellipse" is a nice way of "visualizing" the polarization tensor.

If we now generalize to include all three components, the electric vector E is any direction required to give a unit energy density over a point which will be at the surface of an ellipsoid, as shown in Fig. 31-3. The shape of this ellipsoid of constant energy密度 characterizes the tensor polarizability.

Now an ellipsoid has the nice property that it can always be described simply by giving the directions of three "principal axes" and the diameters of the ellipsoid along these axes. The "principle axes" are the directions of the longest and shortest diameters and the corresponding right angles to them. They are indicated by the axes a , b , and c in Fig. 31-3. With respect to these axes, the ellipsoid has the particularly simple equation

$$\alpha_{xx}E_x^2 + \alpha_{yy}E_y^2 + \alpha_{zz}E_z^2 = 1 \text{ unit.}$$

So with respect to these axes, the polarization tensor has only three components that are non-zero: α_{xx} , α_{yy} , and α_{zz} . That is to say, no matter how complicated a crystal is, it is always possible to choose a set of axes (not necessarily the crystal axes) for which the polarization tensor has only three components. With such a set of axes, Eq. (31.2) becomes simply

$$P_x = \alpha_{xx}E_x, \quad P_y = \alpha_{yy}E_y, \quad P_z = \alpha_{zz}E_z. \quad (31.9)$$

An electric field along any one of the principal axes produces a polarization along the same axis, but the coefficients for the three axes may, of course, be different.

Often, a tensor is described by listing the nine coefficients in a table inside of a pair of brackets:

$$\begin{bmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{bmatrix}. \quad (31.10)$$

For the principal axes a , b , and c only the diagonal terms are not zero, we say here that "the tensor is diagonal." The complete tensor is

$$\begin{bmatrix} \alpha_{aa} & 0 & 0 \\ 0 & \alpha_{bb} & 0 \\ 0 & 0 & \alpha_{cc} \end{bmatrix}. \quad (31.11)$$

The important point is that any polarization tensor (in fact, any symmetric tensor of rank two in any number of dimensions) can be put in this form by choosing a suitable set of coordinate axes.

If the three elements of the polarization tensor in diagonal form are all equal, that is,

$$\alpha_{aa} = \alpha_{bb} = \alpha_{cc} = \alpha, \quad (31.12)$$

the energy ellipsoid becomes a sphere, and the polarizability is the same in all directions. The material is isotropic. In the tensor notation,

$$\alpha_{ij} = \alpha \delta_{ij}, \quad (31.13)$$

where δ_{ij} is the unit tensor.

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}. \quad (31.14)$$

That means, of course,

$$\delta_{ii} = 1, \quad \text{if } i = j, \quad (31.15)$$

$$\delta_{ij} = 0, \quad \text{if } i \neq j.$$

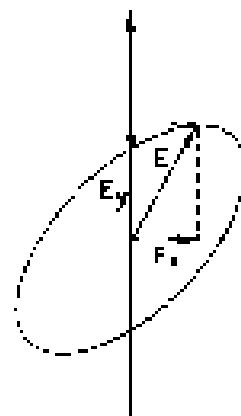


Fig. 31-2. Locus of the vector $E = (E_x, E_y)$ that gives a constant energy of polarization.

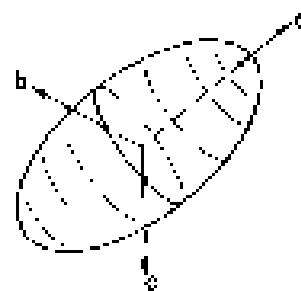


Fig. 31-3. The energy ellipsoid of the polarization tensor.

The tensor δ_{ij} is often called the "Kronecker delta." You may amuse yourself by proving that the tensor (31.14) has exactly the same form if you change the coordinate system to any other rectangular one. The polarization tensor of Eq. (31.13) gives

$$P_i = \sigma \sum_j \delta_{ij} E_j = \sigma E_i$$

which means the same as our old result for isotropic dielectrics:

$$P = \sigma E.$$

The shape and orientation of the polarization ellipsoid can sometimes be related to the symmetry properties of the crystal. We have said in Chapter 30 that there are 320 different possible internal geometries of a three-dimensional lattice and that they can, for many purposes, be conveniently grouped into seven classes, according to the shape of the unit cell. Now the ellipsoid of polarizability must share the internal geometric symmetries of the crystal. For example, a cubic crystal has low symmetry. Its dipole moment will have random axes, and its orientation will not, in general, be aligned with the crystal axes. On the other hand, a monoclinic crystal has the property that its properties are unchanged if the crystal is rotated 180° about one axis. So the polarization tensor must be the same after such a rotation. It follows that the ellipsoid of polarizability must return to itself after a 180° rotation. That can happen only if one of the axes of the ellipsoid is in the same direction as the symmetry axis of the crystal. Otherwise, the orientation and dimensions of the ellipsoid are interchanged.

For an orthorhombic crystal, however, the axes of the ellipsoid must correspond to the crystal axes, because a 180° rotation about any one of the three axes repeats the same lattice. If we go to a tetragonal crystal, the ellipsoid must have the same symmetry, so it must have two equal dimensions. Finally, for a cubic crystal, all three dimensions of the ellipsoid must be equal; it becomes a sphere, and the polarizability of the crystal is the same in all directions.

There is a big game of figuring out the possible kinds of relations for all the possible symmetries of a crystal. It is called "group-theoretical" analysis. But for the simple case of the polarizability tensor, it is relatively easy to see what the relations must be.

31-4 Other tensors: The tensor of inertia

There are many other examples of tensors appearing in physics. For example, in a metal, or in any conductor, one often finds that the current density j is approximately proportional to the electric field E ; the proportionality constant is called the conductivity σ :

$$j = \sigma E$$

For crystals, however, the relation between j and E is more complicated: the conductivity is not the same in all directions. The conductivity is a tensor, and we write

$$\sigma = \sum_{i,j} \sigma_{ij} \delta_{ij}$$

Another example of a physical tensor is the moment of inertia. In Chapter 18 of Volume 1 we saw that a solid object rotating about a fixed axis has angular momentum L , proportional to the angular velocity ω , and we called the proportionality factor I , the moment of inertia:

$$L = I\omega.$$

For an arbitrarily shaped object, the moment of inertia depends on its orientation with respect to the axis of rotation. For instance, a rectangular block will have different moments about each of its three orthogonal axes. Now angular velocity ω and angular momentum L are both vectors. For rotations about one of the axes of symmetry, they are parallel. But if the moment of inertia is different for the

Two principal axes, both ω and α , are, in general, not in the same direction (see Fig. 31-4). They are related in a way analogous to the relation between E and P . In general, we must write

$$\begin{aligned} I_x &= I_{xx}\omega_x + I_{xy}\omega_y + I_{xz}\omega_z \\ I_y &= I_{yx}\omega_x + I_{yy}\omega_y + I_{yz}\omega_z \\ I_z &= I_{zx}\omega_x + I_{zy}\omega_y + I_{zz}\omega_z \end{aligned} \quad (31.16)$$

The three coefficients I_i are called the tensor of inertia. Following the analogy with the propagation, the kinetic energy for any angular momentum must be some quadratic form in the components ω_x , ω_y , and ω_z :

$$KE = \frac{1}{2} \sum_i I_i \omega_i^2 \quad (31.17)$$

We again use the energy to define the ellipsoid of inertia. Also, energy arguments can be used to show that the tensor is symmetric—that $I_{ij} = I_{ji}$.

The tensor of inertia for a rigid body can be worked out, if the shape of the object is known. We need only to write down the total kinetic energy of all the particles in the body. A particle of mass m and velocity v has the kinetic energy $\frac{1}{2}mv^2$, and the total kinetic energy is just the sum

$$\sum m v^2$$

over all of the particles of the body. The velocity v of each particle is related to the angular velocity ω of the solid body. Let's assume that the body is rotating about its center of mass, which we take to be at rest. Then if r is the displacement of a particle from the center of mass, its velocity v is given by $v = \omega \times r$. So the total kinetic energy is

$$KE = \sum m(\omega \times r)^2 \quad (31.18)$$

Now if we have to do a little $\omega \times r$ out in terms of the components ω_x , ω_y , ω_z , and x_1 , y_1 , z_1 , and compare the result with Eq. (31.17), we find I_{ij} by elementary terms. Carrying out the algebra, we write

$$\begin{aligned} (\omega \times r)^2 &= (\omega \times r)_x^2 + (\omega \times r)_y^2 + (\omega \times r)_z^2 \\ &= (\omega_x x - \omega_y y)^2 + (\omega_x y - \omega_z z)^2 + (\omega_x z - \omega_y y)^2 \\ &= \omega_x^2 x^2 - 2\omega_x \omega_y xy + \omega_y^2 y^2 \\ &\quad + \omega_x^2 y^2 - 2\omega_x \omega_z xz + \omega_z^2 z^2 \\ &\quad + \omega_y^2 z^2 - 2\omega_y \omega_z yz - \omega_z^2 x^2. \end{aligned}$$

Multiplying this equation by $m/2$, summing over all particles, and comparing with Eq. (31.17), we see that I_{xx} , for instance, is given by

$$I_{xx} = \sum m(r^2 - x^2)$$

This is the formula we have had before (Chapter 19, Vol. I) for the moment of inertia of a body about the x -axis. Since $r^2 = x^2 + y^2 + z^2$, we can also write this term as

$$I_{xx} = \sum m(r^2 - x^2).$$

Working out all of the other terms, the tensor of inertia can be written as

$$I_{ij} = \left[\begin{array}{ccc} \sum m(r^2 - x^2) & -\sum mxy & -\sum mxz \\ -\sum myx & \sum m(r^2 - y^2) & -\sum myz \\ -\sum mzx & -\sum myz & \sum m(r^2 - z^2) \end{array} \right]. \quad (31.19)$$

If you wish, this may be written in "tensor notation" as

$$I_{ij} = \sum m(r^2 \delta_{ij} - x_i x_j). \quad (31.20)$$

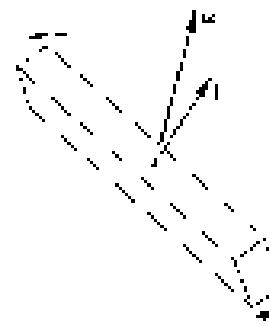


Fig. 31-4. The angular momentum L of a solid object is not, in general, parallel to its angular velocity ω .

where the r_i are the components (x, y, z) of the position vector of a particle and the Σ means to sum over all the particles. The moment of inertia, then, is a tensor of the second rank whose terms are a property of the body and relate E to ω by

$$E = \sum_i I_i \omega_i \quad (3.21)$$

For a body of any shape whatever, we can find the tensor of inertia, that is, the three principal axes. Reduced to three axes, the tensor will be diagonal, so for a rigid, there are always three principal axes for which the components of the angular velocity are zero. They are called the principal axes of inertia.

31-5 The stress product

We should point out that we have been using tensors of the second rank since Chapter 20 of Volume I. There, we defined a "vector in a plane," such as τ_{xy} , by

$$\tau_{xy} = xF_y - yF_x.$$

Generalized to three dimensions, we could write

$$\tau_{ij} = r_i t_j - r_j t_i. \quad (3.22)$$

The quantity τ_{ij} is a tensor of the second rank. One way to see this is to go by obtaining τ_{ij} with some vector, say the unit vector \hat{e}_k , according to

$$\sum_i \tau_{ij} e_i.$$

If this quantity is a vector, then τ_{ij} must be homogeneous (homogeneous is our definition of a tensor). Substituting in (3.22), we have

$$\begin{aligned} \sum_j \tau_{ij} e_i &= \sum_j r_i t_j e_i - \sum_i r_j t_i e_i \\ &= r_i (\hat{e}_j \cdot \hat{e}_i) - r_j (\hat{e}_i \cdot \hat{e}_j). \end{aligned}$$

Since the dot products are scalars, the two terms on the right-hand side are vectors, and likewise their difference. So τ_{ij} is a tensor.

But τ_{ij} is a special kind of tensor: it is *antisymmetric*, that is,

$$\tau_{ij} = -\tau_{ji}$$

so it has only three nonzero terms: τ_{12}, τ_{23} , and τ_{13} . We were able to show in Chapter 20 of Volume I that these three terms, although "by accident," transform like the three components of a vector, so that we could define

$$\tau = (\tau_{12}, \tau_{23}, \tau_{13}) = (r_1, r_2, r_3) \cdot \hat{e}_k.$$

We say "by accident," because it happens only in three dimensions. In four dimensions, for example, an analogous metric tensor of the second rank has six nonzero terms, up to symmetry, and must be replaced by a vector and four components.

Just as the usual vector $\tau = \tau \cdot \hat{e}_k$ is a tensor, so also is every τ its product of two vector vectors, and the same argument applies. By luck, however, they are also represented by vectors (already pseudovectors), so our mathematics has been made easier! i.e., τ .

Mathematically, if τ and θ are any two vectors, the nine quantities appearing in a tensor (although it may have no useful physical purpose). Thus, for the position vector $r_i = r_i \hat{e}_i$, is a tensor, and since β_i is also, we see that the right side of Eq. (3.22) is indeed a tensor. Likewise Eq. (3.22) is a tensor, since the two terms on the right-hand side are tensors.

31-6 The tensor of stress

The symmetric tensors we have described so far arose as coefficients in relating one vector to another. We would like to look now at a tensor which has a different physical significance—the tensor of stress. Suppose we have a solid object with various forces on it. We say that there are various "stresses" inside, by which we mean that there are internal forces between neighboring parts of the material. We have talked a little about such stresses in a two-dimensional case when we discussed the surface tension in a stretched diaphragm in Section 12-3. We will now see that the internal forces in the material of a three-dimensional body can be described in terms of a tensor.

Consider a body of some elastic material—say a block of jello. If we make a cut through the block, the material on each side of the cut will, in general, get displaced by the internal forces. Before the cut was made, there must have been force between the two parts of the block that kept the material in place; we can define the stresses in terms of these forces. Suppose we look at an imaginary plane, perpendicular to the x -axis—like the planes in Fig. 31-5—some way above the front, as shown in Fig. 31-5(a). The material on the left of the plane exerts the force ΔF_x on the material to the right of the plane, as shown in part (b) of the figure. There is, of course, the opposite reaction force $-\Delta F_x$ exerted on the material to the left of the surface. If the area is small enough, we expect that ΔF_x is proportional to the area Δx .

You are already familiar with one kind of stress—the pressure in a static liquid. There the force is equal to the pressure times the area and is at right angles to the surface segment. For solids—also for viscous liquids in motion—the force need not be normal to the surface; there are shear forces involved, as pressures (positive or negative). By a "shear" force we mean the tangential components of the force across a surface. All that we say about the force might be taken literally. Notice also that if we strike you out on a point with your fist, the resulting forces will be different. A complete description of the internal stress requires a tensor.

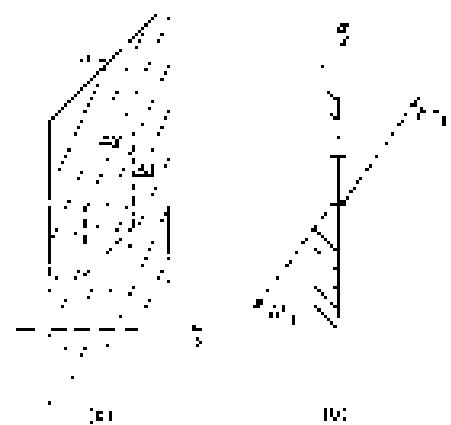


Fig. 31-5. The material to the left of the plane x exerts across the area $\Delta x \Delta y$ the force ΔF_x on the material to the right of the plane.

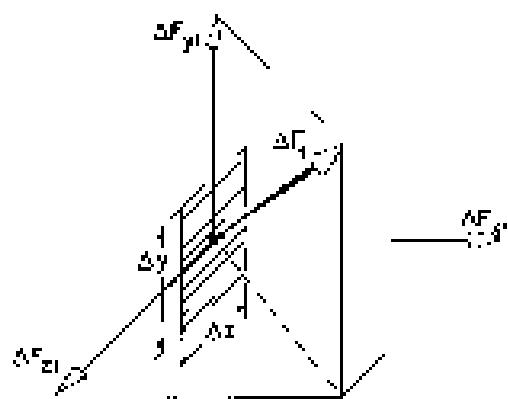


Fig. 31-6. The force ΔF_x acting on a small element $\Delta x \Delta y$ perpendicular to the x -axis is resolved into the three components ΔF_{xy} , ΔF_{xz} , and ΔF_{yl} .

We define the stress tensor in the following way: First, we imagine a cut perpendicular to the x -axis and resolve the force ΔF_x across the cut into its own components ΔF_{xy} , ΔF_{xz} , ΔF_{yl} , as in Fig. 31-6. The ratio of these forces to the area $\Delta x \Delta y$ we call S_{xy} , S_{xz} , and S_{yl} . For example,

$$S_{xy} = \frac{\Delta F_{xy}}{\Delta x \Delta y}.$$

The first index y refers to the direction-force component; the second index x is normal to the area. If you wish, you can write the area $\Delta x \Delta y$ as Δx , meaning an element of area perpendicular to x . Then

$$S_{xy} = \frac{\Delta F_{xy}}{\Delta x}.$$

Next, we think of an imaginary cut perpendicular to the y -axis. Across a small

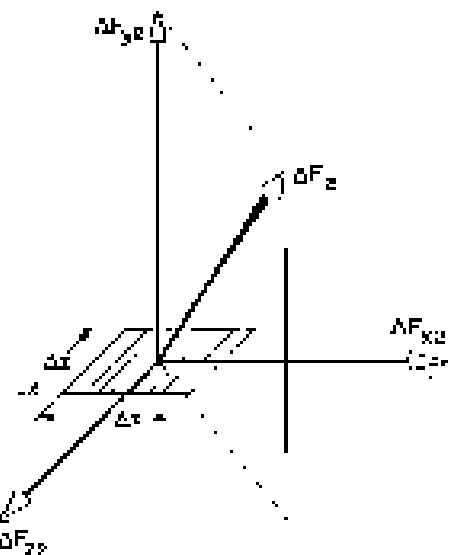


Fig. 31-7. The force across an element of area ΔA perpendicular to y is resolved into three rectangular components.

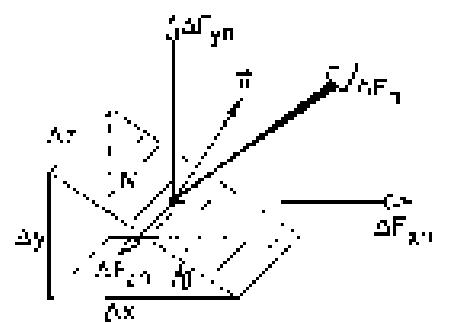


Fig. 31-8. The force F_z across the face N whose unit normal is n is resolved into components.

area ΔA as there will be a force ΔN . Again we resolve this force into three components, as shown in Fig. 31-7, and define the three components of the stress S_{xx} , S_{yy} , S_{zz} as the force per unit area in the three directions. Finally, we choose an imaginary cut perpendicular to y and define the three components S_{xy} , S_{yz} , and S_{xz} ; we have the nine numbers:

$$S_{ij} = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{bmatrix}. \quad (31-2)$$

We want to show now that these nine numbers are sufficient to describe completely the internal state of stress, and that S_{ij} is indeed a tensor. Suppose we want to know the force across a surface inclined at some arbitrary angle θ from the x -axis. From S_{ij} ? Yes, in the following way: We imagine a little solid triangle, which has one face N in the new surface and the other two parallel to the x -y-z axes. If the face N happened to be parallel to the xy -plane, we would have the triangular piece shown in Fig. 31-8. (This is a somewhat special case, but will illustrate well enough the general method.) Now the stress forces at the little solid triangle in Fig. 31-8 are in equilibrium (at least in the limit of infinitesimal changes in θ), so the total force on it must be zero. We know the forces on the edges parallel to the x -axis come directly from S_{xx} . These forces are just equal in magnitude, so ΔF_{xe} is zero and we can express this condition in terms of S_{ij} :

The assumption that the surface forces on the small triangular volume are in equilibrium implies any one "body" forces that might be present, such as gravity or surface forces of different magnitude systems, are in mutual balance. Notice, however, that each body force is parallel to the x -axis of the little triangle and therefore, to Δx , Δy , Δz , whereas all the surface forces are proportional to the areas such as ΔF_x , ΔF_y , ΔF_z , etc. So if we take the scale of the little wedge small enough, the body forces can always be neglected in comparison with the surface forces.

Let's now add up forces on the little wedge. We can forget the x -component, which is the sum of the forces from each face. However, if Δz is small enough, the forces on the triangular faces (perpendicular to the z -axis) will be equal and opposite, so we can forget them. The y -component of the force on the bottom rectangle is

$$\Delta F_{ye} = S_{yy} \Delta y \Delta z.$$

The x -component of the force on the vertical rectangle is

$$\Delta F_x = S_{xx} \Delta x \Delta z.$$

These two must be equal to the y -component of the force summed across the face N . (It's either the unit vector method for forces ΔF and the force on ΔF , or the δ -function)

$$\Delta F_{yy} = S_{yy} \Delta z \Delta x + S_{yz} \Delta x \Delta z.$$

The x -component S_{xy} of the stress across this plane is equal to ΔF_x divided by $\Delta x \Delta z$, which is $\Delta F_x / \Delta x^2 = \Delta F_x / \Delta A$,

$$S_{xy} = S_{yy} \frac{\Delta y}{\sqrt{\Delta x^2 + \Delta z^2}} = S_{yy} \frac{\Delta x}{\sqrt{3} \sqrt{\Delta x^2 + \Delta z^2}}.$$

Now $\Delta x / \sqrt{\Delta x^2 + \Delta z^2}$ is the cosine of the angle θ between Δx and its y -axis, as shown in Fig. 31-8, so Δ can also be written as θ , the y -component of $\pi \cdot S$; namely, $\Delta x / \sqrt{\Delta x^2 + \Delta z^2}$ is an $\theta = \theta_y$. We can write

$$S_{xy} = S_{yy} \theta_y + S_{yz} \theta_x.$$

If we now generalize to an arbitrary surface element, we would get the

$$S_{ij} = S_{ii} \theta_i + S_{iy} \theta_y + S_{iz} \theta_z,$$

or, in general,

$$S_{ij} = \sum_k S_{ik} \delta_{kj}. \quad (31.24)$$

We can find the force across any surface element in terms of the S_{ij} , so it does describe completely the state of internal stress of the material.

Equation (31.24) says that the tensor S_{ij} relates the force S_i to the unit vector n_j , just as σ_i relates F to E . Since n and S are vectors, the components of S_{ij} must transform as a tensor with changes in coordinate axes. So S_{ij} is indeed a tensor.

We can also show that S_{ij} is a symmetric tensor by looking at the forces on a little cube of material. Suppose we take a little cube, oriented with its faces parallel to our coordinate axes, and look at its six faces separately, as shown in Fig. 31-9. If we let the x -axis of the cube be one axis, the six components of the forces on the faces normal to the x -axis would simply be as shown in the figure. To the cube it seems that the stresses do not change approximately from one side of the cube to the other (unless there are other components, like a shear component S_{xy} , which there must be since one of the edges of it would start opening). In fact, if we chart the center of $(S_{xx}, -S_{yy})$ versus the front edge of the cube, and since the point is zero, S_{yy} is equal to S_{xx} , and the stress tensor is symmetric.

Since S_{ij} is a symmetric tensor, it can be represented by an ellipsoid which will have three principal axes. But surfaces normal to these axes, the stresses are particularly simple—they correspond to planes or paths perpendicular to the surfaces. There are also shear balances along these faces. For any stress, we can always choose our axes so that the shear components are zero. If the ellipsoid is a sphere, there are only normal forces in any direction. This corresponds to a hydrostatic pressure (positive or negative). So for a hydrostatic pressure, the tensor is diagonal and all three components are equal; they are, in fact, just equal to the pressure p . We can write

$$S_{ij} = p \delta_{ij}. \quad (31.25)$$

The stress tensor—and also its ellipsoid—will, in general, vary from point to point in a block of material; to describe the whole block we need to give the values of each component of S_{ij} as a function of position. So the stress tensor is a field. We have had scalar fields like the temperature $T(x, y, z)$, which give one number for each point in space, and vector fields like $E(x, y, z)$, which give three numbers for each point. Now we have a tensor field which gives nine numbers for each point in space—or really six for the symmetric tensor S_{ij} . A complete description of the total field needs six arbitrary discrete values (or nine if S_{ij} is not zero).

31-7 Tensors of higher rank

The stress tensor S_{ij} describes the internal forces of matter. If the material is elastic, it is convenient to describe the internal stresses in terms of another tensor T_{ijkl} called the strain tensor. For a simple object like a bar of metal, you know that the change in length, ΔL , is approximately proportional to the force, so we say “Energy Method”; we

$$\Delta L = \gamma F.$$

For a solid elastic body with arbitrary distortions, the strain T_{ijkl} is related to the stress S_{ij} by a set of linear equations:

$$T_{ijkl} = \sum_k S_{ijk} \delta_{lj}. \quad (31.26)$$

Also, you know that the potential energy of a spring (or bar) is

$$\frac{1}{2} F^2 \Delta L = \frac{1}{2} k F^2.$$

The generalization for the elastic energy density in a solid body is

$$U_{\text{elastic}} = \sum_{ijkl} \gamma_{ijkl} S_{ij} S_{kl}. \quad (31.27)$$

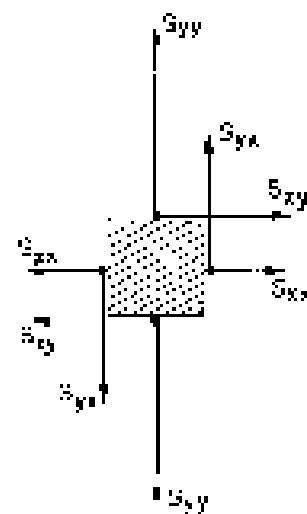


Fig. 31-9. The x - and y -faces on four faces of a small unit cube.

The complete description of the elastic properties of a crystal must be given in terms of the coefficients S_{ijkl} . This introduces us to a new beast. It is a tensor of the fourth rank. Since each index can take on any one of three values, i, j, k , or l , there are $3^4 = 81$ coefficients. But there are really only 21 different numbers. First, since S_{ij} is symmetric, i , has only $\frac{1}{2}k$ different values, and only $\frac{1}{2}k$ different coefficients are needed in Eq. (31.27). But also, S_{ij} can be interchanged with S_{kl} without changing the energy, so S_{ijkl} must be symmetric if we interchange i and k . This reduces the number of different coefficients to 21. So to describe the elastic properties of a crystal of the lowest possible symmetry requires 21 elastic constants! This number is, of course, reduced for crystals of higher symmetry. For example, a cubic crystal has only three elastic constants, and an isotropic substance has only two.

That the latter is true can be seen as follows. How can the components of S_{ijkl} be independent of the direction of the axes, as they must be if the material is isotropic? On the other hand, they can be independent only if they are expressible in terms of the tensor δ_{ij} . There are two possible expressions, $\delta_{ik}\delta_{jl}$ and $\delta_{il}\delta_{jk} + \delta_{ik}\delta_{lj}$, which have the required symmetry, so S_{ijkl} must be a linear combination of them. Therefore, for isotropic materials,

$$S_{ijkl} = a(\delta_{ik}\delta_{jl}) - b(\delta_{il}\delta_{jk} + \delta_{ik}\delta_{lj}),$$

and the material requires two constants, a and b , to describe its elastic properties. We will leave it for you to show that a cubic crystal needs only three.

As a final example, this time of a third-rank tensor, we take the piezoelectric effect. Under stress, a crystal possesses an electric field property called P_{ijk} , hence, "piezoelectric." The law is

$$E_i = \sum_{j,k} P_{ijk} S_{jk},$$

where E_i is the electric field, and the P_{ijk} are the piezoelectric coefficients of the piezoelectric tensor. Can you show that if the crystal has a center of inversion under $(x, y, z) \rightarrow (-x, -y, -z)$ the piezoelectric coefficients are all zero?

31-8 The four-vector of electromagnetic induction

All the tensors we have looked at so far in this chapter relate to the three dimensions of space; they are defined to have a certain transformation property under spatial rotations. In Chapter 26 we had occasion to use a tensor in the four dimensions of relativistic space-time—the electromagnetic field tensor $A_{\mu\nu}$. The components of such a four-tensor transform under a Lorentz transformation of the coordinates in a special way that we worked out. (Although we did not do it that way, we could have considered the Lorentz transformation as a “rotation” in a four-dimensional “space” called Minkowski space; then the analogy with what we are doing here would have been closer.)

As our last example, we want to consider another tensor in the four dimensions (x, y, v, z) of relativity theory. When we wrote the stress tensor, we defined S_{ij} as a component of a force across a unit area. But a force is equal to the rate of change of a momentum. Therefore, instead of saying “ S_{xy} is the x component of the force across a unit area perpendicular to y ,” we could equally well say, “ S_{xy} is the rate of flow of the x -component of momentum through a unit area perpendicular to y .” In other words, each term of S_{ij} also represents the flow of the i -component of momentum through a unit area perpendicular to the j -direction. There are pure space components, but they are parts of a “larger” tensor S_{ij} in four dimensions (x and $y = (i, x, y, z)$) containing additional components like S_{xx} , S_{yy} , S_{zz} , etc. We will now try to find the physical meaning of these extra components.

We know that the space components represent flow of momentum. We can get a clue on how to extend this to the time dimension by studying another kind of “flow”—the flow of electric charge. For the scalar quantity, charge, the rate of flow (per unit area perpendicular to the flow) is a space vector—the current density

were). We have seen that the time component of this four-vector is the density of the stuff that is flowing. For instance, you can combine with a time component, $j_0 = \rho$, the charge density, to make the four-vector $k = (\rho, j)$; that is, the ω in k takes on the values ρ, j, x, y, z to mean "density, rate of flow in the x -direction, rate of flow in y , rate of flow in z " of the scalar charge.

Now by analogy with our statement about the time component of the flow of a scalar quantity, we might expect that with S_{xx}, S_{yy} , and S_{zz} , describing the flow of the x -component of momentum, there should be a time component S_{tt} , which would be the density of whatever is flowing; that is, S_{tt} should be the density of a momentum. So we can extend our tensor by inserting to include a t -component. We have

$$\begin{aligned} S_{tt} &= \text{density of } x\text{-momentum,} \\ S_{tx} &= x\text{-flow of } x\text{-momentum,} \\ S_{ty} &= y\text{-flow of } x\text{-momentum,} \\ S_{tz} &= z\text{-flow of } x\text{-momentum.} \end{aligned}$$

Similarly, for the y -component of momentum we have the three components of flow: S_{yy}, S_{ty}, S_{zy} . In a similar way shall we find the z -momentum terms:

$$S_{zz} = \text{density of } y\text{-momentum.}$$

And, of course, to S_{tx}, S_{ty}, S_{tz} we would add

$$S_{tx} = \text{density of } x\text{-momentum.}$$

In four dimensions there is also a component of momentum, which is, we know, energy. So the tensor S_{ij} should be extended vertically with S_{tt}, S_{yy} , and S_{zz} , where

$$\begin{aligned} S_{tt} &= x\text{-flow of energy,} \\ S_{tx} &= x\text{-flow of energy,} \\ S_{ty} &= y\text{-flow of energy,} \\ S_{tz} &= z\text{-flow of energy.} \end{aligned} \quad (11.28)$$

That is, S_{tt} is the flow of energy per unit area and per unit time across a surface perpendicular to the x -axis, and so on. Finally, to complete our tensor we need S_{vv} , which would be the density of energy. We have extended our stress tensor S_{ij} of three dimensions to the four-dimensional stress-energy tensor S_{ijv} . The index v can take on the four values t, x, y , and z , meaning, respectively, "density," "flow per unit area in the x -direction," "flow per unit area in the y -direction," and "flow per unit area in the z -direction." In this we may v takes on the four values t, x, y, z to label other flows, namely, "energy," "momentum," "heat current," "current density," "charge density," and "momentum in the v -direction."

As an example, we will discuss the electric field in optics, but in a region of free space (in which there is no electric or magnetic field). We know that the flow of energy is the Poynting vector $S = \mathbf{E} \times \mathbf{B}$. So the x , y , and z -components of S are, from the relativistic sum of squares, the components S_{tx}, S_{ty} , and S_{tz} of our four-dimensional stress-energy tensor. The fourth part of the tensor S_{ijv} carries over to the time components as well, so the four-dimensional tensor S_{ijv} is symmetric:

$$S_{ij} = S_{ji}. \quad (11.29)$$

In other words, the components S_{tx}, S_{ty}, S_{tz} , which are the densities of x , y , and z -momentum, are also equal to the x , y , and z -components of the Poynting vector S , the energy flow, as we have already shown it, in either of space or a different kind of language.

The remaining components of the electromagnetic stress tensor S_{ij} can also be expressed in terms of the electric and magnetic fields E and B . That is to say, we must admit there is, to put it very seriously, flow of momentum in the electromagnetic field. We discussed this in Chapter 27 in connection with Eq. (27.21), but did not work out the details.

For a given charge density their power is given in two dimensions by
 ω_0 to see the formula for \mathcal{E}_0 and details of the fields:

$$\lambda_{\mu\nu} = \frac{q}{4\pi} \left(\sum_{\alpha} \delta_{\mu\alpha} \delta_{\nu\alpha} - \frac{1}{3} \epsilon_{\mu\nu} \sum_{\alpha\beta\gamma} \delta_{\alpha\beta} \delta_{\gamma\mu} \right).$$

where $\epsilon_{\mu\nu}$ is the epsilon tensor, $\delta_{\mu\nu}$ is the Kronecker delta function and we have used a special meaning for the sum sign \sum_{α} to be equal to the symbol $\delta_{\mu\nu}$ in the case where the terms are not to be differentiated and $\delta_{\mu\nu} = 1$, while $\delta_{\mu\nu} = \delta_{\mu\mu} = \delta_{\nu\nu} = 1$ and $\delta_{\mu\nu} = 0$ for $\mu \neq \nu$ for $\mu = \nu$. One can verify that it gives the charge density $\lambda_{\mu\nu} = q\epsilon_{\mu\nu}/4\pi r^2$. If one the Boyer-Lindquist metric $\times R^2$ then you know that ϵ is an eigenfunction of ∂_R . This implies that ϵ is orthogonal to the direction of the field, i.e. $\epsilon \cdot E = 0$. There is an equal number of ϵ 's having perpendicular to the field direction.

Refractive Index of Dielectric Materials

32-1 Polarization of matter

We want now to discuss the polarization of the detection of light— and 3.5. We have the description of the dielectric materials. In Chapter 31 of Vol. 1 we discussed the theory of the index of refraction n , because it is the refractive index of the material at the time ω but to correct corrected to adding the index only for materials of low density, like gases. A physical process that produces densities here (3.5), is not clear. The electric field E of light wave induces the polarization of the gas, producing existing dipole moments. The acceleration of the oscillating charges radiates new waves of ripples. This new field, interacting with the old field, produces a change in field which is equivalent to a phase shift of the original wave. However, the phase shift is proportional to the density of the gas. At the effect is equivalent to having a different phase velocity in the material. While we looked at the subject before, we neglected the polarization due to the form factor effect on the new wave changing the ends of the scattering dipole. We assumed that the forces on the charges in the atoms came from the incoming wave. It was, in fact, the result of the wave only by the incoming wave to be seen by the reflected wave of the previous issues. It would now have to think, what is in their turn to account this effect, so we started only the ray for the wave and didn't account.

Now, there are several ways that it is very convenient to be polarization. One of different experiments. This method requires the physical origin of the index of refraction from the reflected wave with respect to the original wave, can determine the density of these materials more simply. This chapter will bring together arguments and proofs from our earlier work. We will let you practically everything we will review the material only for really non-idealized materials. Since you may need one of the previous chapters when we are going to read again. Table 32-1 lists some equations we are going to use together with a reference to the place where they can be found. In most instances, we will get more details by giving the physical argument again, and will give just the equation.

Table 32-1

The work in this chapter will be based on the following material, already covered in earlier chapters

Subject	Reference	Equation
Damped oscillations	Vol. 1, Chap. 17	$\ddot{\mathbf{r}}(t) = -\omega_0^2 \mathbf{r}(t) - \frac{2\pi}{\tau} \mathbf{r}(t) + \mathbf{F}$
Index of glass	Vol. 1, Chap. 17	$n = 1 + \frac{1 - \omega_0^2}{2\omega_0^2(\omega_0^2 - \omega^2)}$ $n = c = n^{1/2}$
Motion	Vol. 1, Chap. 17	$\ddot{\mathbf{r}}(t) = -\omega_0^2 \mathbf{r}(t) + \mathbf{F}$
Electric conductivity	Vol. 1, Chap. 18	$\mathbf{j} = \sigma \mathbf{E} = \frac{q}{m} \mathbf{v} = \frac{q}{m} \mathbf{E}$
Polarization	Vol. 1, Chap. 19	$\mathbf{p}_s = -\nabla \times \mathbf{B}$
Induced dipoles	Vol. 1, Chap. 19	$\mathbf{p}_{sd} = \epsilon_0 \mathbf{E} = \frac{1}{\epsilon_0} \mathbf{F}$

32-2 Polarization of matter

32-2 Maxwell's equations in a dielectric

32-3 Waves in a dielectric

32-4 The complex index of refraction

32-5 The index of a medium

32-6 Waves in metals

32-7 Low-frequency and high-frequency approximations, the skin depth and the plasma frequency

Review. See Table 32-1.

We begin by recalling the machinery of the index of refraction for a gas. We suppose that there are N particles per unit volume and that each particle behaves as a harmonic oscillator. We use a model of an atom or molecule in which the electron is bound with a force proportional to its displacement (as though the electron were held in place by a spring). We emphasize that this was not a legitimate classical model of an atom, but we will show later that the current quantum mechanical theory gives results equivalent to this model (in simple cases). In our earlier treatment, we did not include the possibility of damping force in the atomic oscillators, but we will do so now. Such a force corresponds to a resistance to the motion, that is, to a force proportional to the velocity of the electron. Then the equation of motion is

$$\ddot{x} + \omega_0^2 x = \omega_0^2 E - m^{-1} \omega_0^2 \dot{x} \quad (32.1)$$

where x is the displacement parallel to the direction of E . (We are assuming an isotropic medium where damping force is the same in all directions.) Also, we are taking, for the moment, a linearly polarized wave, so that E doesn't change direction. If the electric field oscillates, the x -values sinusoidally with time, we write

$$E = E_0 e^{i\omega t} \quad (32.2)$$

The displacement will then oscillate with the same frequency, and we write

$$x = x_0 e^{i\omega t}$$

Substituting $\ddot{x} = i\omega x_0$ and $\dot{x} = -\omega x_0$, we can solve Eq. 32.1 for x_0 :

$$x_0 = \frac{i\omega^2 E_0}{\omega^2 + i\omega\gamma + \omega_0^2} \quad (32.3)$$

Knowing the displacement, we can calculate the acceleration \ddot{x} and find the forces responsible for the index. This was the way we computed the index in Chapter 31 of Volume I.

Now, however, we want to take a different approach. The induced dipole moment p is related to x as follows, using Eq. (32.3).

$$p = \frac{q_e^2 m}{-\omega^2 + i\omega\gamma + \omega_0^2} E \quad (32.4)$$

Since p is proportional to E , we write

$$p = \kappa_0 \epsilon_0 \omega E \quad (32.5)$$

where κ_0 is called the *atomic polarization*. With this definition, we have

$$\kappa_0 = \frac{q_e^2 m \epsilon_0}{-\omega^2 + i\omega\gamma + \omega_0^2} \quad (32.6)$$

The quantum mechanical solution is the motion of electrons in atoms gives a similar answer except with the following modifications. The atoms have several natural frequencies, each frequency with its own dissipation constant γ . Also the effective "strength" of each mode is different, when we end represent by multiplying the polarizability for each frequency by a strength factor f , which is a number we expect to be of the order of 1. Representing the three parameters ω , γ , and f by ω_i , γ_i , and f_i for each mode of oscillation, and summing over the

¹ Throughout this chapter we follow the notation of Chapter 31 of Volume I, and Eq. 32.1 represents the same situation as we defined there. In our last example we used ω to represent the relative permeability—the ratio of μ to μ_0 . In the notation of this chapter $\mu = \kappa_0 \epsilon_0 \omega$; μ (Eq. 32.6).

various modes, we modify Eq. (32.6) to read

$$\epsilon(\omega) = \frac{N}{C\omega} \sum_p \frac{\rho_p}{\omega_p^2 - \omega^2 + i\omega\tau_p} \quad (32.7)$$

If ΔE is the induced voltage per unit volume in the material, the polarization P is just $\Delta E = \epsilon_0 \Delta E/E$, and is proportional to E :

$$P = \epsilon_0 \epsilon(\omega) E. \quad (32.8)$$

In other words, when there is a sinusoidal electric field acting in a material, there is a induced dipole moment per unit volume which is proportional to the electric field—i.e., proportionality constant is that, for simplicity, depends upon the frequency. At very high frequencies ω is small; there is no net response. However, at low frequencies ω can be so strong response. Also, the proportionality constant is complex in order which means that the polarization does not exactly follow the electric field, but may vary with time as shown before. At any rate there is $\epsilon(\omega)$ a total polarizability or susceptibility proportional to the strength ω^2 of the electric field.

32-2 Maxwell's equations in a dielectric

The existence of such nonlinear materials that have an polarization-charging density only inside of the material, and these must be put into Maxwell's equations in order to find the fields. We can ignore the free charges since Maxwell's equations describe situations in which the charge and currents are not zero, as in a vacuum, but it is given implicitly by the polarization vector. Our first step is to find explicitly the charge density ρ and current density j averaged over a small volume of the size ΔV ; we look in order to see what we get for P . Then the primary we need can be obtained from the polarization.

We have seen, in Chapter 13 that, why the polarization P varies from place to place, there is a conductivity given by

$$\rho_s = -\nabla \cdot P. \quad (32.9)$$

At that time, we were dealing with static fields, and the same formula is valid also for time-varying fields. However, when P varies with time, there are charges in motion, so there is also a polarization current. Each of the oscillating charges contributes a current equal to its charge q_{av} times its velocity v . With N such charges per unit volume, the current density j is

$$j = Nq_v v$$

Since we know that $v = \dot{r}$ divide, then $j = Nq \dot{r}$ (velocity) which is just $\partial P/\partial t$. Therefore the current density from the varying polarization is

$$j_{av} = \frac{\partial P}{\partial t}. \quad (32.10)$$

Our problem is now clear and simple. We write Maxwell's equations with the charge density and current density expressed in terms of P , using Eqs. (32.8) and (32.10). (We assume that there are no other currents and charges in the material.) We then relate P to E with Eq. (32.5), and we solve the equation on E and B —looking for the wave solutions.

Before we do this, we would like to make an historical note. Maxwell originally wrote his equations in a form which was different from the one we have been using. Because the equations were written in this different form for many years—and are still written that way by many people—we will explain the difference. In the early days, the mechanism of the dielectric constant was not fully understood, nor that there was a polarization of the material. So people did not agree to the fact there was a contribution

to the charge density ρ and $\nabla \cdot P$. They thought only in terms of charges that were far from the sources (such as the charges for flow in wires or in coated air surfaces).

Today, we prefer to keep separate the total charge density, including the part from the bound atomic charges. It was well that part ρ_{ext} , we can write

$$\rho = \rho_{\text{ext}} + \rho_{\text{atom}}$$

where ρ_{atom} is the charge density considered by Maxwell and refers to the charges and binding forces of atoms. We will then write

$$\nabla \cdot E = \frac{\rho_{\text{ext}} + \rho_{\text{atom}}}{\epsilon_0}$$

Substituting ρ_{ext} from Eq. (32.9),

$$\begin{aligned} \nabla \cdot E &= \frac{\rho_{\text{atom}} - \frac{1}{\mu_0} \nabla \cdot P}{\epsilon_0} \\ \text{or} \quad \nabla \cdot (\epsilon_0 E + P) &= \rho_{\text{atom}}. \end{aligned} \quad (32.11)$$

The current density in the Maxwell equations for $\nabla \times B$ also has unperturbed contributions from bound atomic currents. We can therefore write

$$j = j_{\text{ext}} + j_{\text{atom}}$$

and the Maxwell curling becomes

$$\epsilon_0 \nabla \times B = j_{\text{ext}} + \frac{j}{\mu_0} + \frac{\partial H}{\partial t}. \quad (32.12)$$

Using Eq. (32.10), we get

$$\epsilon_0 c^2 \nabla \times B = j_{\text{atom}} + \frac{P}{\mu_0} (\epsilon_0 E + P). \quad (32.13)$$

Now given, as in Fig. 1, we want to define a new vector D by

$$D = \epsilon_0 E + P, \quad (32.14)$$

the two field equations would become

$$\nabla \cdot D = \rho_{\text{atom}}, \quad (32.15)$$

and

$$\epsilon_0 c^2 \nabla \times B = j_{\text{atom}} - \frac{\partial D}{\partial t}. \quad (32.16)$$

This is actually the form that Maxwell used for dielectrics. His two remaining equations were

$$\nabla \times E = -\frac{\partial H}{\partial t},$$

and

$$\nabla \cdot B = 0,$$

which are the same as we have been using.

Maxwell and the other early workers also had a problem with magnetic materials (which we will take up soon). Because they did not know about the circulating currents responsible for atomic magnetism, they used a current density that was missing still another part. Instead of Eq. (32.16), they actually wrote

$$\nabla \times B = P + \frac{\partial D}{\partial t}, \quad (32.17)$$

where H differs from $\epsilon_0^{-1}B$ because it reduces the effects of surface currents (They represent what is left of the currents). So Maxwell had four field vectors— E , B , D , and H . He gave little, if any, passing attention to what H is.

was going on inside the material. You will find the equations written this way in many places.

To solve the equations, it is necessary to relate D and B to the other fields, and perhaps to ω as well:

$$D = \epsilon E \quad \text{and} \quad B = \mu H \quad (32.18)$$

However, these relations are only approximately true for some materials, and even then only if the fields are not changing rapidly with time. (For sinusoidally varying fields one often can't use the equations this way by making ϵ and μ complex functions of the frequency, but not for an arbitrary time variation of the fields.) So there used to be all sorts of cheating in solving the equations. We think the right way is to keep the equations in terms of the fundamental quantities as we now understand them—and that's how we have done it.

A2.3 Waves in a dielectric

We want now to find out what kind of electromagnetic waves exist in a dielectric material in which there are no extra charges other than those bound in atoms. So we take $\rho = -\nabla \cdot P$ and $j = \partial P/\partial t$. Because there are no free

$$\begin{aligned} (\text{a}) \quad \nabla \cdot B &= -\frac{\rho}{\epsilon_0} & (\text{b}) \quad c^2 \nabla \times B &= \frac{\partial}{\partial t} \left(\frac{P}{\epsilon_0} + E \right) \\ (\text{c}) \quad \nabla \times E &= -\frac{\partial B}{\partial t} & (\text{d}) \quad \nabla \cdot B &= 0 \end{aligned} \quad (32.19)$$

We will solve these equations as we have done before. We start by taking the curl of Eq. (32.19c)

$$\nabla \times (\nabla \times B) = -\frac{d}{dt} \nabla \times B.$$

Next, we make use of the vector identity

$$\nabla \times (\nabla \times E) = \nabla(\nabla \cdot E) - \nabla^2 E,$$

and also substitute for $\nabla \times B$, using Eq. (32.19b), we get

$$\nabla(\nabla \cdot E) - \nabla^2 E = -\frac{1}{\epsilon_0 c^2} \frac{\partial^2 P}{\partial t^2} - \frac{1}{c^2} \frac{\partial^2 E}{\partial t^2}.$$

Using Eq. (32.19a) for $\nabla \cdot E$, we get

$$\nabla^2 E + \frac{1}{c^2} \frac{\partial^2 E}{\partial t^2} = -\frac{1}{\epsilon_0} \nabla(\nabla \cdot P) = \frac{1}{\epsilon_0 c^2} \frac{\partial^2 P}{\partial t^2}. \quad (32.20)$$

So instead of the wave equation, we now get that the differentiation of E is equal to two terms involving the polarization P .

Since P depends on E , however, Eq. (32.20) can still have wave solutions. We will now limit ourselves to isotropic dielectrics, so that P is always in the same direction as E . Let's try to find a solution for a wave going in the x -direction. Then, the electric field might vary as $e^{i(kx-\omega t)}$. We will also suppose that the wave is polarized in the x -direction—that the electric field has only an x -component. We write

$$E_x = E_0 e^{i(kx-\omega t)}. \quad (32.21)$$

You know that any function of $(x - vt)$ represents a wave that travels with the speed v . The exponent of Eq. (32.21) can be written as

$$ik \left(x - \frac{\omega}{c} t \right),$$

so Eq. (32.21) represents a wave with the phase velocity

$$v_{ph} = \omega/c$$

The index of refraction is defined (see Chapter 31, Vol. I) by setting

$$\frac{c}{v} = \frac{\epsilon'}{\epsilon_0}$$

Thus Eq. (32.21) becomes

$$P_x = \epsilon_0 \epsilon' \omega^2 E_x$$

So we can find α by finding what value of ϵ' is required if Eq. (32.21) is to satisfy the given field equations, and then using

$$\alpha = \frac{\epsilon'}{\epsilon_0} \quad (32.22)$$

In an isotropic material, there will be only an x -component of the polarization, then P has no variation with the x coordinate, so $\nabla \cdot P = 0$, and we get rid of the first term on the right-hand side of Eq. (32.21). Also, since we are assuming a linear dielectric, ϵ_0 will vary as ω^2 , and $\epsilon_0^2 \partial_x / \partial t^2 = -\alpha^2 P_x$. The Laplacian in Eq. (32.20) becomes simply $\partial^2 E_x / \partial x^2 = -\alpha^2 P_x$, so we get

$$-\alpha^2 E_x = \frac{\epsilon'^2}{c^2} E_x = -\frac{\epsilon'^2}{c^2} P_x \quad (32.23)$$

Now let us assume for the moment that since E is varying sinusoidally, we can set P proportional to E , as in Eq. (32.5). (We'll come back to discuss this assumption later.) We write

$$P_x = \epsilon_0 \epsilon' E_x$$

Then P_x is a solution of Eq. (32.23), and we find

$$\alpha^2 = \frac{m^2}{c^2} (1 - \alpha \alpha_0) \quad (32.24)$$

We have found that a wave like Eq. (32.21), with the wave number k given by Eq. (32.24), will satisfy the field equations. Using Eq. (32.23), the answer to question 1 of

$$\alpha^2 = 1 + \beta_{\text{eff}} \quad (32.25)$$

Let's compare this formula with what we obtained in our theory of the index of a gas (Chapter 31, Vol. I). There, we got Eq. (31.29), which is

$$\alpha = 1 + \frac{\lambda_0^2}{2 \pi \alpha_0} \frac{1}{1 - \omega^2 + \omega_0^2} \quad (32.26)$$

Taking α from Eq. (32.5), Eq. (32.25) would give us

$$\alpha' = 1 + \frac{\lambda_0^2}{m c} \frac{1}{1 - \omega^2 + (\pi \alpha_0 + \omega_0^2)} \quad (32.27)$$

First, we have the new term in α' , because we are including the dissipation of the oscillators. Second, the left-hand side is a fraction of α^2 , and there is an extra factor of $1/2$. But notice that if α is small enough so that ω is close to one (as it is for a gas), then Eq. (32.27) says that α^2 is one plus a small number: $\alpha^2 = 1 + \alpha$. We can then write $\alpha = \sqrt{1 + \alpha} \approx 1 + \alpha/2$, and the two expressions are equivalent. Thus our new method gives for a gas the same result we found earlier.

Now you might think that Eq. (32.27) would give the index of refraction for dense materials also. It needs to be modified, however, for several reasons. First, the derivation of this equation assumes that the polarizing field on each atom is the field E_x . That assumption is not right, however, because in dense materials there is also the field produced by other atoms in the vicinity, which may be comparable to E_x . We considered a similar problem when we studied the static fields in dielectrics. (See Chapter 11.) You will remember that we estimated the field at a single atom by imagining that it sat in a spherical hole in the surrounding dielectric. The field in such a hole—which we called the local field—is increased

over the average field E by the amount $\rho/\epsilon_0 c$. (Remember, however, that this result is only strictly true in *isotropic* materials—including the special case of a static crystal.)

The latter approximation will hold for the electric field in a wave, so long as the wavelength of the wave is much longer than the spacing between atoms. Limiting ourselves to such cases, we write

$$E_{\text{local}} = E - \frac{\rho}{\epsilon_0 c}. \quad (32.25)$$

This local field is the one that should be used for E in Eq. (32.3); that is, Eq. (32.3) should be rewritten:

$$P = \epsilon_0 N \mu_0 E_{\text{local}}. \quad (32.26)$$

Using E_{local} from Eq. (32.25), we find

$$\mu = \epsilon_0 N \mu_0 \left(E - \frac{\rho}{\epsilon_0 c} \right)$$

or

$$\mu = \frac{\epsilon_0}{1 - (\lambda \rho / \epsilon_0 c)} \epsilon_0 E. \quad (32.27)$$

In other words, the local field E is only proportional to E (for a material "at rest"). However, the constant of proportionality is not ϵ_0 , as we wrote above Eq. (32.27), but should be $\epsilon_0 N \mu_0 / (1 - (\lambda \rho / \epsilon_0 c))$. So we should rewrite Eq. (32.25) instead

$$\eta^2 = 1 + \frac{N \rho}{1 - (\lambda \rho / \epsilon_0 c)}. \quad (32.28)$$

It will be more convenient if we rewrite this equation as

$$\frac{\eta^2 - 1}{\eta^2 + 2} = N \rho, \quad (32.29)$$

which is algebraically equivalent. This is known as the Clausius-Mosotti equation.

There is another complication in dense materials. Because neighboring atoms are so close, there are strong interactions between them. The individual modes of oscillation are, therefore, modified. The natural frequencies of the atomic oscillators are屏ed off by the interactions, and they are usually quite rapidly damped—the relaxation well could become quite large. So that ω_0 's and γ_0 's of the ω_0 's will be quite different from those of the free atoms. With these reservations, we can still represent η^2 at least approximately by Eq. (32.29). We have then that

$$\frac{\eta^2 - 1}{\eta^2 + 2} = \frac{N \rho^2}{4 \pi n_0} \sum_i \frac{1}{-\omega_i^2 + i \gamma_i \omega_i + \omega_{i0}^2}. \quad (32.30)$$

One final complication. If the dense material is a mixture of several components, each will contribute to the polarization. The total η^2 will be the sum of the contributions from each component of the mixture [except for the inaccuracy of the local field approximation, Eq. (32.28), in ordered crystals—effect we discussed when analyzing ferroelectrics]. Writing N_i as the number of atoms of each component per unit volume, we should replace Eq. (32.30) by

$$\frac{\eta^2 - 1}{\eta^2 + 2} = \sum_i N_i \eta_i^2, \quad (32.31)$$

where each η_i^2 will be given by an expression like Eq. (32.27). Equation (32.31) complements theory of the index of refraction. The quantity $(\eta^2 - 1)/(\eta^2 + 2)$ is called by some authors "mean" or frequency, which is the mean atomic polarizability $\alpha(\omega)$. The precise evaluation of $\alpha(\omega)$ (that is, finding f_i , ω_i , and ω_{i0}) in dense substances is a difficult problem of quantum mechanics. It has been done from first principles only for a few especially simple substances.

32-4 The complex index of refraction

We want to look now at the consequences of our result, Eq. (32.33). First we notice that ϵ is complex, so the index is going to be a complex number. What does that mean? Let's say that ω is the angular frequency of a real and an imaginary part

$$\omega = \omega_R - i\omega_I \quad (32.35)$$

where ω_R and ω_I are real functions of ω . We would like with a minus sign so that ω will be a positive quantity in all ordinary optical theories—the ordinary insulating materials—that are real. Like ω_R , ω_I is connected to ϵ_0 since ϵ_0 is a positive number, and that makes ω_I negative if ω_R is positive. Our picture here of Eq. (32.21) is $E = E_0 e^{-i\omega R t + \omega_I t}$.

$$E = E_0 e^{-i\omega R t + \omega_I t}$$

Writing ω as in Eq. (32.35), we would have

$$E = E_0 e^{-(\omega_R t + \omega_I t)} e^{i\omega_R t} \quad (32.36)$$

The term $e^{-(\omega_R t + \omega_I t)}$ represents a wave travelling with the speed c/ϵ_0 , so ω_R represents what we normally think of as the index of refraction. But the amplitude of this wave is

$$E_0 e^{-\omega_I t},$$

which decreases exponentially with t . A graph of the strength of the electric field at some instant as a function of t is shown in Fig. 32-4, for $\omega_I < \omega_R/2\pi$. The imaginary part of the index represents the attenuation of the wave due to the energy losses in the atomic oscillators. The intensity of the wave is proportional to the square of the amplitude, so

$$\text{Intensity} \propto e^{-2\omega_I t}.$$

This is often written as

$$\text{Intensity} \propto e^{-ct},$$

where $c = 2\omega_I/c$ is called the absorption coefficient. Thus we have in Eq. (32.33) not only the theory of the index of refraction of materials, but the theory of their absorption of light as well.

In most materials considered to be transparent materials, the quantity c —which has the dimensions of a length—is quite large in comparison with the thickness of the material.

32-5 The index of a mixture

There is another prediction of our theory of the index of refraction that we can check against experiment. Suppose we consider a mixture of two materials. The index of the mixture is not the average of the two indices, but should be given in terms of the sum of the two polarizabilities, as in Eq. (32.34). If we ask about the index of, say, a sugar solution, the total polarizability is the sum of the polarizability of the water and that of the sugar. Each can, of course, be evaluated using for N the number per unit volume of the molecules of the particular kind. In other words, if a given solution has N_1 molecules of water, whose polarizability is α_1 , and N_2 molecules of sucrose ($C_{12}H_{22}O_{11}$), whose polarizability is α_2 , we should have that

$$2 \left(\frac{\alpha_1}{N_1} + \frac{\alpha_2}{N_2} \right) = N_1 \alpha_1 + N_2 \alpha_2. \quad (32.37)$$

We can use this formula to test our theory against experiment, by measuring the index for various concentrations of sucrose in water. We are making several assumptions here, however. Our formula assumes that there is no chemical action when the sucrose is dissolved in the water. No disturbance to the individual atoms

Table 32-2
Refractive index of sucrose solutions, and comparison with predictions of Eq. (32.37).

Data from Haileman's			D Moles of sucrose per liter, N_1/N_2	E Moles of water per liter, N_2/N_3	F $\beta \left(\frac{N_1}{N_2} + \frac{1}{2} \right)$	G N_{expt}	H N_{pred}	I N_{pred} (g. n. index)
A Fraction of sucrose by weight	B Density (g/cm ³)	C at 589 nm						
0	0.9982	1.332	0	55.5	0.717	0.612	0	
0.20	1.1263	1.3811	0.079	45.3	0.708	0.647	0.211	0.213
0.40	1.2293	1.4200	0.298	32.5	0.750	0.704	0.380	0.311
0.60	1.4452	1.5013	0.59	17.0	0.885	0.1005	0.757	0.717
1.00	1.88	1.5977	1.41	7	0.960	0	0.950	0.207

* pure water

** sucrose (crystallized)

† molar-molecular weight of water = 18

*** sucrose crystal

**** molecular weight of sucrose = 342

oscillations are not too different for various concentrations. So our result is certainly only approximate. Anyway, let's see how good it is.

We have picked the example of a sugar solution because there is a good table of measured values of the index of refraction in the *Handbook of Chemistry and Physics*, and also because sugar is a molecular crystal that goes into solution without changing or otherwise changing its chemical state.

We give in the first three columns of Table 32-2 the data from the handbook. Column A is the percent of sucrose by weight, column B is the measured density (g/cm³), and column C is the measured index of refraction (at light-wavelength is 589.3 millimicrons). For pure sugar we have taken the measured index of sugar crystals. The crystals are not isotropic, so the measured index is different along different directions. The handbook gives three values:

$$n_1 = 1.5275, \quad n_2 = 1.5651, \quad n_3 = 1.5715.$$

We have taken the average.

Now we could try to compute α for each concentration, but we don't know what value to have for n_1 or n_2 . Let's test the theory this way: We will assume that the polarizability of water (α_1) is the same at all concentrations and compute the polarizability of sucrose by using the experimental values for α and solving Eq. (32.27) for α_2 . If the theory is correct, we should get the same α_2 for all concentrations.

First, we need to know N_1 and N_2 ; let's express them in terms of Avogadro's number, N_A . Let's take one liter (1000 cm³) for our unit of volume. Then N_1/N_2 is the weight per liter divided by the gram-molecular weight. And the weight per liter is the density multiplied by 1000 to get grams per liter. So the fractional weight of either the sucrose or the water. In this way, we get N_1/N_2 and N_1/N_3 , as in columns D and E of the table.

In column F we have computed $\beta(n^2 - 1)/(n^2 + 2)$ from the experimental values of n in column C. For pure water, $\beta(n^2 - 1)/(n^2 + 2)$ is 0.612, which is equal to just N_1/N_3 . We can then fill in the rest of Column G, since for each row G/F may be the same ratio, namely, 0.612/55.5. Subtracting column G from column F, we get the contribution N_{expt} of the sucrose, shown in column H. Dividing these entries by the values of N_1/N_3 in column D, we get the values of N_{pred} shown in column I.

From our theory we would expect all the values of N_{pred} to be the same. They are not exactly equal, but pretty close. We can conclude that our ideas are fairly correct. Even more, we find that the polarizability of the sugar molecule doesn't seem to depend much on its surroundings—it's polarizability is nearly the same in a dilute solution as it is in the crystal.

32-6 Waves in metals

The theory we have worked out in this chapter for solid materials can also be applied to poor conductors, like metals, with very little modification. In metals some of the electrons have no binding force holding them to any particular atom; these are "free" electrons which are responsible for the conductivity. There are other electrons which are bound, and the theory shown is directly applicable to them. Their influence, however, is usually swamped by the effects of the conduction electrons. We will consider now only the effects of the free electrons.

There is an accelerating force on an electron—but still some resistance to its motion. Its equation of motion differs from Eq. (32.1) only because the term in a_E is missing. So if we have been except of $a_E = 0$ in the rest of our derivations—except that there's one more difference. The reason that we had to distinguish between the average field and the local field in a dielectric is that in an insulator each of the dipoles is fixed in position, so that it has a definite relationship to the position of the others. But because the conduction electrons in a metal move around, however, the field on top of the average is just the average field E . By the correction we made in Eq. (32.5) by using Eq. (32.28) should now be used for conduction electrons. Therefore the formula for the index of refraction for metals should look like Eq. (32.27), except with one set of units, namely,

$$n^2 = 1 - \frac{c_0}{2\pi\epsilon_0} \frac{1}{\omega^2 + \omega_{pe}^2}. \quad (32.48)$$

This is only the contribution from the conduction electrons, which we will assume is the major term, for metals.

Now we even know how to find what value to use for ω_p , because it is related to the conductivity of the metal. In Chapter 3 of Volume I we discussed how the conductivity of a metal comes from the flow of the free electrons through the crystal. The electrons go on a jagged path, from one scattering to the next, and between scatterings they move freely except for an acceleration due to any average electric field (as shown in Fig. 32-2). We found in Chapter 10 of Volume I that the average drift velocity is just the acceleration times the average time τ between collisions. The acceleration is $q_0 E/m$, so

$$v_{drift} = \frac{q_0 E}{m} \tau. \quad (32.49)$$

Now for such an acceleration, E was constant, so that v_{drift} was a steady velocity. Since there is no average acceleration, the drag force is equal to the applied force. We have defined "drag" since F_{drag} is the drag force (see Eq. (32.11)), which is $q_0 E$; therefore we have that

$$\tau = \frac{1}{\gamma}. \quad (32.50)$$

Although we cannot easily measure v_{drift} directly, we can determine it by measuring the conductivity of the metal. It is known physically that an electric field E in a metal produces a current with the density j proportional to E (for most materials):

$$j = \sigma E.$$

The proportionality constant σ is called the *conductivity*. This is just what we expect from Eq. (32.32) if we set

$$j = Nq v_{drift} e.$$

Then

$$\sigma = \frac{Nq^2}{m} \tau. \quad (32.51)$$

Since v_{drift} and τ can be related to the observed electrical conductivity, using Eqs. (32.49) and (32.51), we can rewrite our formula for the index, Eq. 32.48



Fig. 32-2. The motion of a free electron.

(32.38), in the following form:

$$n^2 = 1 + \frac{\sigma/\tau}{\omega(1 + \omega\tau)} \quad (32.42)$$

where

$$\tau = \frac{1}{\gamma} = \frac{m_e}{Nq_e}. \quad (32.43)$$

This is a convenient formula for the index of refraction of metals.

32-7 Low-frequency and high-frequency approximations: the skin depth and the plasma frequency

Our result, Eq. (32.42), for the index of refraction for metals predicts quite different characteristics for wave propagation at different frequencies. Let's first consider waves at very low frequencies. If ω is small enough, we can approximate Eq. (32.42) by

$$n^2 = 1 + \frac{\sigma}{\omega\tau}. \quad (32.44)$$

Now, as you can check by taking the square root,

$$\sqrt{-i} = \frac{1 - i}{\sqrt{2}},$$

so for low frequencies,

$$n = \sqrt{2}/2 \exp(i(1 - \omega\tau)). \quad (32.45)$$

The real and imaginary parts of n have the same magnitude. With such a large imaginary part $\text{Im } n$, the wave is rapidly attenuated in the metal. Referring to Eq. (32.36), the amplitude of a wave going in the $-z$ direction decreases as

$$\exp[-\sqrt{\sigma\epsilon_0/\omega\tau} z]. \quad (32.46)$$

Let's write this as

$$e^{-\alpha z}, \quad (32.47)$$

where δ is just the distance in which the wave amplitude decreases by the factor $e^{-1} = 1/2.72$ —or roughly one-third. The amplitude of such a wave as a function of z is shown in Fig. 32-3. Since electromagnetic waves will penetrate into a metal only this distance, δ is called the *skin depth*. It is given by

$$\delta = \sqrt{\sigma\epsilon_0/\omega\tau}. \quad (32.48)$$

Now what do we mean by "low" frequencies? Looking at Eq. (32.47), we see that it can be approximated by Eq. (32.44) only if $\omega\tau$ is small. This being true, one condition is also much less than one—that is, our low-frequency approximation applies when

$$\omega \ll \frac{1}{\tau}$$

and

$$\omega \ll \frac{\sigma}{\epsilon_0}. \quad (32.49)$$

Let's see what frequencies these correspond to for a typical metal like copper. We compute ϵ by using Eq. (32.43), and σ/ϵ_0 by using the measured conductivity. We take the following data from a handbook:

$$\sigma = 5.76 \times 10^7 \text{ (ohm-meter)}^{-1},$$

$$\text{atomic weight} = 63.5 \text{ grams},$$

$$\text{density} = 8.9 \text{ g/cm}^3 = 8.9 \times 10^{-3} \text{ kg/m}^3,$$

$$\text{Avogadro's number} = 6.02 \times 10^{23} \text{ (gram atomic weight)}^{-1}.$$

¹ On writing $-i = \omega\tau/2$, $\sqrt{-i} = e^{-\omega\tau/4} = \cos(\omega\tau/4) - i \sin(\omega\tau/4)$, which gives the same result.

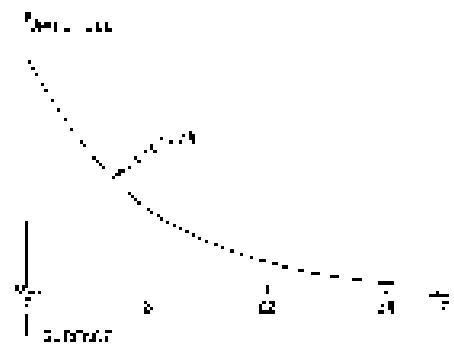


Fig. 32-3. The amplitude of an incident electromagnetic wave as a function of distance in a metal.

If we assume that there is one free electron per atom, then the number of electrons per cubic meter is

$$N = 5.5 \times 10^{28} \text{ meter}^{-3}.$$

Using

$$q_e = 1.6 \times 10^{-19} \text{ coulombs},$$

$$\epsilon_0 = 8.85 \times 10^{-12} \text{ farad/meter}^{-1},$$

$$m = 9.11 \times 10^{-31} \text{ kg},$$

we get

$$\tau = 2.4 \times 10^{-13} \text{ sec},$$

$$\frac{i}{\tau} = 2.1 \times 10^{11} \text{ sec}^{-1},$$

$$\frac{\sigma}{\epsilon_0} = 6.5 \times 10^{18} \text{ sec}^{-1}.$$

At low frequencies less than about 10^{12} cycles per second, copper will have the "low-frequency" behavior we describe (that means for waves whose free-space wavelength is longer than 0.3 millimeters - waves from radio waves).

For these waves, the skin depth in copper is

$$d = \sqrt{\frac{0.026 \text{ m}^2 \cdot \text{sec}}{\omega}}$$

For microwaves of 10,000 megacycles per second (lambda waves)

$$d = 6.7 \times 10^{-4} \text{ cm}.$$

The wave penetrates a very small distance.

We can see from this why in studying cavities (or waveguides) we need to worry only about the fields inside the cavity, and not in the metal outside the cavity. Also, we see why the losses in a cavity are reduced by a thin plating of silver or gold. The losses come from the current which appears only in a thin layer equal to the skin depth.

Suppose we look now at the index of a metal like copper at high frequencies. For very high frequencies ω is much greater than ω_p , and Eq. (32.42) is well approximated by

$$n^2 = 1 + \frac{\epsilon_0}{\epsilon_r \omega^2 \tau}. \quad (32.50)$$

For waves at high frequencies the index of a metal becomes real and less than unity. This is also evident from Eq. (32.48) if the dissipative term α is neglected. As can be done for very large ω , Equation (32.48) gives

$$n^2 = 1 - \frac{\omega_p^2}{\omega^2 \tau \epsilon_0}, \quad (32.51)$$

which is, of course, the same as Eq. (32.49). We have seen before the quantity $\omega_p^2/\epsilon_0 \tau \epsilon_0$, which we called the square of the plasma frequency (Section 7-5)

$$\omega_p^2 = \frac{\epsilon_0 \eta^2}{\epsilon_r \tau \epsilon_0},$$

so we can write Eq. (32.49) or Eq. (32.51) as

$$n^2 = 1 - \left(\frac{\omega_p}{\omega} \right)^2$$

The plasma frequency is a kind of "critical" frequency.

For $\omega < \omega_p$ the index of a metal has no imaginary part, and waves are attenuated; but for $\omega > \omega_p$ the index is real, and the metal becomes transparent. You know, of course, that metals are reasonably transparent to x-rays - *BUT*, some metals are even transparent in the ultraviolet. In Table 23-2 we give the ω_p

several metals the experimental observed wavelength at which they begin to become transparent. In the second column we give the calculated critical wavelength $\lambda_c = c/\omega_p$. Considering that the experimental wavelength is not too well defined, the fit of the theory is fairly good.

You may wonder why the plasma frequency ω_p should have anything to do with the propagation of electromagnetic waves in metals. The plasma frequency comes up in Chapter 7 as the natural frequency of density oscillations of the free electrons. (A cloud of electrons is repelled by electric forces, and the inertia of the electrons leads to an oscillation of density.) So longitudinal plasma waves are resonant at ω_p . But we are now talking about transverse electromagnetic waves, and we have found that transverse waves are absorbed for frequencies below ω_p . (It's no coincidence since ω_p is an accidental coincidence!)

Although we have been talking about wave propagation in metals, you appear to by this time the inaccuracy of the observation of physics—that it doesn't make any difference whether the free electrons are in a metal, or whether they are in the plasma of the interplanetary space, or in the atmosphere of a star. To understand radio propagation in the atmosphere, or even radio communications—using, of course, the proper values for λ and ω . We can see now why long radio waves are absorbed or reflected by the ionosphere, whereas short waves get right through. (Short waves must be used for communications with satellites.)

We have to look ahead the high-end, low-frequency extreme for wave propagation in metals. For the cut-off between transparent and fully opaque, the following formula¹ (Eq. 32.42) must be used. In general, the index will have real and imaginary parts; the wave is attenuated as it propagates into the metal. For very thin layers, metals are somewhat transparent even at optical frequencies. As an example, special gildes for people who work around high-temperature furnaces are made by evaporating a thin layer of gold on glass. The visible light is transmitted fairly well—with a strong green tinge—but the infrared is strongly absorbed.

Finally, it cannot have escaped the reader that many of these formulas resemble in some ways those for the dielectric constant ϵ discussed in Chapter 10. The dielectric constant ϵ measures the response of the material to a constant field, that is, for $\omega = 0$. If you look carefully at the definition of ϵ and ν you see that ν is simply the limit of ϵ^2 as $\omega \rightarrow 0$. Indeed, placing $\omega = 0$ and $\epsilon^2 = \nu$ in equations of this chapter will reproduce the equations of the theory of the dielectric constant of Chapter 11.

Table 32-3*

Wavelengths below which the metal becomes transparent

Metal	λ_c (in micrometers)	$\lambda_c = c/\omega_p$
T	1550 Å	1550 Å
Na	2130	2130
K	2150	2870
Rb	3400	4220

* From: G. Kino, *Introduction to Solid State Physics*, John Wiley and Sons, Inc., New York, 2nd ed., 1958, p. 266.

Reflection from Surfaces

33-1 Reflection and refraction of light

The subject of this chapter is the reflection and refraction of light—or electromagnetic waves in general—at surfaces. We have already discussed the laws of reflection and refraction in Chapter 35 of Volume I. Recall what we found out there:

1. The angle of reflection is equal to the angle of incidence. With the angles defined as shown in Fig. 33-1,

$$\theta_r = \theta_i \quad (33.1)$$

2. The product $n \sin \theta$ is the same for the incident and transmitted beam. (Snell's law)

$$n_i \sin \theta_i = n_s \sin \theta_s \quad (33.2)$$

3. The intensity of the reflected light depends on the angle of incidence and also on the direction of polarization. For E perpendicular to the plane of incidence, the reflection coefficient R_{\perp} is

$$R_{\perp} = \frac{I_r}{I_i} = \frac{\tan^2(\theta_i - \theta_s)}{\tan^2(\theta_i + \theta_s)} \quad (33.3)$$

For E parallel to the plane of incidence, the reflection coefficient R_{\parallel} is

$$R_{\parallel} = \frac{I_r}{I_i} = \frac{\tan^2(\theta_i - \theta_s)}{\tan^2(\theta_i + \theta_s)} \quad (33.4)$$

4. For normal incidence (any polarization),

$$\frac{I_r}{I_i} = \left(\frac{n_2 - n_1}{n_2 + n_1} \right)^2 \quad (33.5)$$

(Earlier, we used i for the "incident angle" and r for the refracted angle. Since we can't use i for both "refracted" and "reflected" angles, we are now using θ_i = incident angle, θ_r = reflected angle, and θ_s = transmitted angle.)

Our earlier discussion is really about what anyone would normally meet up with the subject, but we are going to go it all over again in a different way. Why? One reason is that we assumed before that the indexes were real for absorption in the materials. But another reason is that you should know how to deal with what happens to waves at surfaces from the point of view of Maxwell's equations. We'll get the same answers as before, but now from a straightforward solution of the wave problem, rather than by some clever arguments.

We want to emphasize that the amplitude of a surface reflector is not a property of the material, as is the index of refraction. It's a "surface property," one that depends precisely on how the surface is made. A thin layer of extraneous junk on the surface between two materials of index n_1 and n_2 will usually change the reflector. (There are all kinds of possibilities of interference here—like the colors of oil films. Sufficient thickness can even reduce the reflected amplitude to zero for a given frequency; that's how coated lenses are made.) The formulas we will derive are correct only if the change of index is sudden—within a distance very small compared with one wavelength. For light, the wavelength is about 5000 Å, so by a "smooth" surface we mean one in which the conditions change in

33-1 Reflection and refraction of light

33-2 Waves in dense materials

33-3 The boundary conditions

33-4 The reflected and transmitted waves

33-5 Reflection from metals

33-6 Total internal reflection

Review: Chapter 35, Vol. I, Polarization

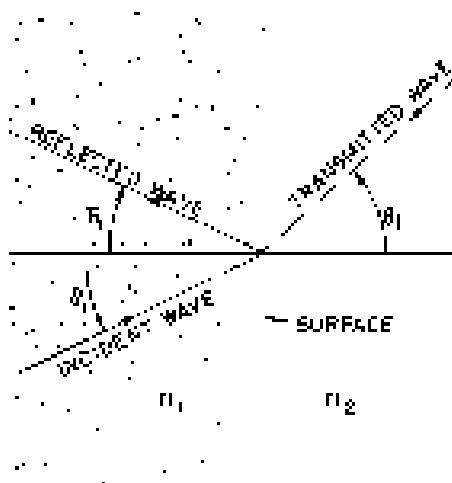


Fig. 33-1. Reflection and refraction of light waves at a surface. (The wave directions are normal to the wave fronts.)

going a distance of only a few atoms (or a few angstroms). Our equations will work for light for highly polished surfaces. In general, if the index changes gradually over a distance of several wavelengths, there is very little reflection at all.

33-2 Waves in dense materials

Just as we showed you about the convenient way of describing a sinusoidal plane wave we used in Chapter 26 of Volume I, Any field component in the wave we use E as an example can be written in the form

$$E = E_0 e^{i(kx - \omega t)}, \quad (33.6)$$

where E represents the amplitude at the point x (from the origin) at the time t . The vector k points in the direction the wave is travelling, and the magnitude $|k| = k = 2\pi/\lambda$ is the wave number. The phase velocity of the wave is $v_p = \omega/k$, for a light wave in a material the value $v_p = c/n = 300,000$

$$t = \frac{v_p}{c} x, \quad (33.7)$$

Suppose k is in the \hat{x} -direction, i.e. $k = k\hat{x}$, just as we have often used it. For k in some other direction, we should replace k by k_r , the distance from the origin in the \hat{k} -direction; that is, we should replace k^2 by k_r^2 , which is just k^2 . (See Fig. 33.2.) So Eq. (33.6) is a convenient representation of a wave in any direction.

We must remember, of course, that

$$k^2 r = k_x x + k_y y + k_z z,$$

where k_x , k_y , and k_z are the components of k along the three axes. In fact, we pointed out once that (k_x, k_y, k_z, k) is a four-vector, and that its scalar product with $(x, y, z, 1)$ is an invariant. So the phase of a wave is an invariant, and Eq. (33.6) could be written

$$E = E_0 e^{i k \cdot r},$$

But we don't need to be that fancy now.

For a sinusoidal E , as in Eq. (33.6), $\partial E / \partial t$ is the wave velocity v , and $i k \partial E / \partial t = -ik_x E$, and so on for the other components. You know already it is very convenient to use the form in Eq. (33.6) when working with the other equations. Differentiations are replaced by multiplications. One little useful point: The operation $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ gets replaced by the three in it ∂ 's, $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$. But these three act as the components of the vector k , so the operator ∇ gets replaced by multiplying each with $-ik$:

$$\begin{aligned} \frac{\partial}{\partial x} &\rightarrow ik_x, \\ \nabla &\rightarrow -ik. \end{aligned} \quad (33.8)$$

This remains true for any ∇ operation—whether it is the gradient, or the divergence, or the curl. For instance, the x component of $\nabla \times E$ is

$$\frac{\partial E_y}{\partial z} - \frac{\partial E_z}{\partial y}.$$

If both E_y and E_z vary as $e^{i k \cdot r}$, then we get

$$ik_x E_y - ik_y E_z,$$

which is, you see, the x -component of $-ik \times E$.

So we have the very useful fact that, whenever you have to take the gradient of a vector field, such as a wave function, in a medium (they are an important part of physics), you can always take the differentiation quickly and cleanly without thinking by means of the fact that the operation ∇ is equivalent to multiplication by $-ik$.

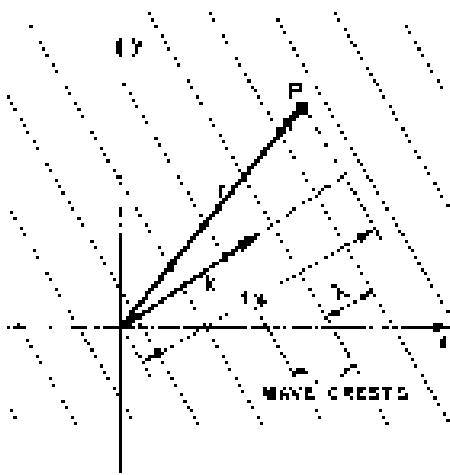


Fig. 33.2. For a wave moving in the direction k , the phase at any point P is $\text{int} = \mathbf{k} \cdot \mathbf{r}_P$.

For instance, the Faraday equation

$$\nabla \times E = -\frac{\partial B}{\partial z}$$

becomes for a wave

$$\nabla \times E = -ikB.$$

This tells us that

$$B = \frac{k}{w} \frac{E}{z}, \quad (33.9)$$

which corresponds to the result we found earlier for waves in free space—that B , in a wave, is at right angles to E and to the wave direction. (In free space, $w/k = c$.) You can remember the sign in Eq. (33.9) from the fact that B is in the direction of Poynting's vector, $S = \text{curl } E \times B$.

If you use Faraday's law with the other Maxwell equations, you get again the results of the last chapter, and, in particular, (3.1)

$$k \cdot k = c^2 + \frac{w^2 n^2}{c^2}. \quad (33.10)$$

But since we know that, we won't do it again.

If you would like to entertain yourself, you can try the following tantalizing problem that was the ultimate test for graduate students back in 1930: solve Maxwell's equations for plane waves in an *anisotropic* crystal, that is, when the polarization P is related to the electric field E by a tensor of polarizability. You should, of course, choose your axes so that the principal axes of the tensor, so that the relations are simplest (then $P_x = n_1 E_x$, $P_y = n_2 E_y$, and $P_z = n_3 E_z$), but let the waves have an arbitrary direction and polarization. You should be able to find the relations between E and B , and how k varies with directions in wave polar space. Then you will understand the optics of an anisotropic crystal. It would be best to start with the simplest case of a birefringent crystal, like calcite, in which two of the polarizabilities are equal (say, $n_1 = n_2$), and see if you can understand why you see double when you look through such a crystal. If you are not that bold, then try the hardest case, in which all three n 's are different. (Even you will know whether you are up to the level of a graduate student in 1930.) In this chapter, however, we will consider only isotropic substances.

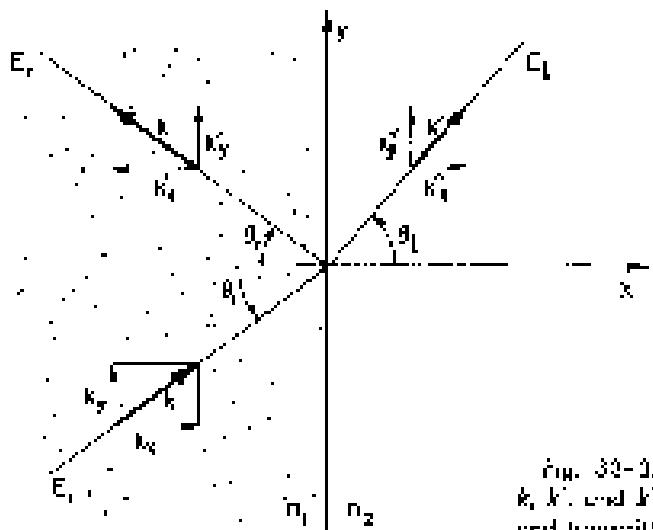


Fig. 33-3. The propagation system k , k' , and k'' for the incident, reflected, and transmitted waves.

We know from experience that when a plane wave arrives at the boundary between two different materials—say, air and glass, or water and oil—there is a wave reflected and a wave transmitted. Suppose we assume no more than that, and see what we can work out. We choose our axes with the z -axis in the surface, and the x -axis perpendicular to the incident wave velocity, as shown in Fig. 33-3.

The electric vector of the incident wave can then be written as

$$E_i = E_{i0} e^{j(kz - \omega t)} \quad (33.11)$$

Since \mathbf{A} is perpendicular to the z -axis,

$$k_x r = k_{y1} + k_{y2} \quad (33.12)$$

We add the reflected wave as

$$E_r = E_{r0} e^{j(kz + \omega t)} \quad (33.13)$$

so that its frequency is ω , its wave number is k' , and its amplitude is E_{r0} . (We know, of course, that the frequency is the same, the magnitude of k is the same as for the incident wave, but we are not going to assume even that. We will let k come out of the normalized equations.) Finally, we write for the transmitted wave,

$$E_t = E_{t0} e^{j(kz - \omega t)} \quad (33.14)$$

We know that one of Maxwell's equations gives Eq. (33.3), so for each of the waves we have

$$B_z = \frac{k}{\omega} \times E_{z0}, \quad B_r = \frac{k'}{\omega'} \times E_r, \quad B_t = \frac{k''}{\omega''} \times E_t. \quad (33.15)$$

Also, if we call the indices of the two media n_1 and n_2 , we have from Eq. (33.20)

$$k'^2 = k_1^2 + k_y^2 = \frac{\omega^2 n_1^2}{c^2}. \quad (33.16)$$

Since the reflected wave is in the same material, then

$$k'^2 = \frac{\omega^2 n_1^2}{c^2}, \quad (33.17)$$

whereas for the transmitted wave,

$$k''^2 = \frac{\omega^2 n_2^2}{c^2}. \quad (33.18)$$

33-3 The boundary conditions

All we have done so far is to describe the three waves; our problem now is to work out the parameters of the reflected and transmitted waves in terms of those of the incident wave. How can we do that? The three waves we have described satisfy Maxwell's equations in the uniform material, but Maxwell's equations must also be satisfied at the boundary between the two different materials. So we must now look at what happens right at the boundary. We will find that Maxwell's equations demand that the three waves fit together in a certain way.

As an example of what we mean, the x -component of the electric field E must be the same on both sides of the boundary. This is required by Faraday's law,

$$\nabla \times \mathbf{E} = - \frac{\partial \mathbf{B}}{\partial z}, \quad (33.19)$$

as we can see in the following way. Consider a little rectangular loop Γ which straddles the boundary, as shown in Fig. 33-4. Equation (33.19) says that the line integral of \mathbf{E} around Γ is equal to the rate of change of the flux of \mathbf{B} through the loop:

$$\oint_{\Gamma} \mathbf{E} \cdot d\mathbf{s} = - \frac{\partial}{\partial z} \int_{\Gamma} \mathbf{B} \cdot d\mathbf{n}$$

Now imagine that the z -direction is very narrow, so that the loop encloses an infinitesimal area. If \mathbf{B} is zero & finite (and that's the reason it should be infinite at the boundary) the flux through the loop is zero. So the line integral of \mathbf{E} must be zero.

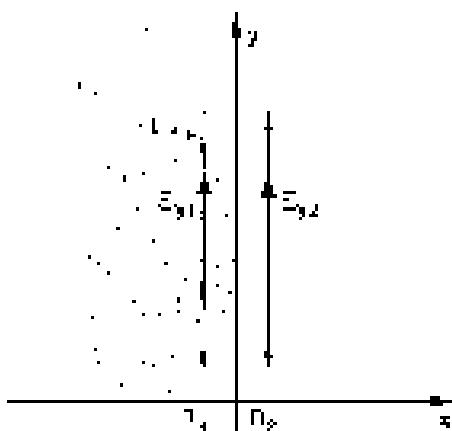


Fig. 33-4. A boundary condition $\oint_{\Gamma} \mathbf{E} \cdot d\mathbf{s} = 0$ is obtained from $\oint_{\Gamma} \mathbf{B} \cdot d\mathbf{n} = 0$.

by $\nabla \cdot E$, $\nabla \times B$, and E_{ext} are the components of the field on the two sides of the boundary and of $\partial \Omega$, $\partial \Omega$ being the boundary, we have

$$\begin{aligned} \nabla \cdot E &= E_{\text{ext}} = 0 \\ \nabla \times B &= B_{\text{ext}} \end{aligned} \quad (33.20)$$

as we have said. This gives us some relation among the fields at the three waves.

The procedure of "working out the consequences of Maxwell's equations of the continuity of each field along the boundary conditions." Fortunately, this can be done by finding as many derivatives (the up to (33.20) as one can), by moving a point about little rotating each (in Fig. 33.4, or by taking little positive surfaces that straddle the boundary). Although that is a perfectly good way of proceeding, it gives the impression that the problem of dealing with a boundary is different for every different physical problem.

For example, in a problem of heat flow across a boundary, how are the temperatures on the two sides related? Well, you could argue, for one thing, that the heat flow to the boundary from one side would have to equal the flow from the other side. It is usually possible and generally quite useful to work out the boundary conditions by making such physical arguments. There may be times, however, when in working on some problem you have only some equations, and you may not see right away what physical arguments to use. So although we are at the moment interested only in an electromagnetic problem, where we can use the physical arguments, we want to show you a method that can be used for any problem—a general way of finding what happens at a boundary directly from the differential equations.

We begin by writing all the Maxwell equations for a dielectric—see this to see we are very specific and write out explicitly all the components:

$$\nabla \cdot F = -\frac{\epsilon_0}{c^2} \frac{\partial B}{\partial t} + \frac{1}{\epsilon_0} \left(\frac{\partial E_x}{\partial y} + \frac{\partial E_y}{\partial z} + \frac{\partial E_z}{\partial x} \right) = -\frac{1}{\epsilon_0} \left(\frac{\partial P_x}{\partial y} + \frac{\partial P_y}{\partial z} + \frac{\partial P_z}{\partial x} \right) \quad (33.21a)$$

$$\begin{aligned} \nabla \times E &= -\frac{\partial B}{\partial t} \\ \frac{\partial E}{\partial y} - \frac{\partial E}{\partial z} &= \frac{\partial B}{\partial x} \quad (33.21b) \\ \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} &= \frac{\partial B_y}{\partial z} \quad (33.21b) \\ \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} &= \frac{\partial B_z}{\partial x} \quad (33.21c) \end{aligned}$$

$$\nabla \cdot B = 0 \quad (33.22) \\ \frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0$$

$$\nabla \times B + \frac{1}{c^2} \frac{\partial P}{\partial t} = \frac{\partial E}{\partial t} \\ c^2 \left(\frac{\partial B_x}{\partial y} - \frac{\partial B_y}{\partial x} \right) = \frac{1}{\epsilon_0} \frac{\partial P_x}{\partial y} + \frac{\partial E_x}{\partial y} \quad (33.24a)$$

$$c^2 \left(\frac{\partial B_y}{\partial z} - \frac{\partial B_z}{\partial y} \right) = \frac{1}{\epsilon_0} \frac{\partial P_y}{\partial z} + \frac{\partial E_y}{\partial z} \quad (33.24b)$$

$$c^2 \left(\frac{\partial B_z}{\partial x} - \frac{\partial B_x}{\partial z} \right) = \frac{1}{\epsilon_0} \frac{\partial P_z}{\partial x} + \frac{\partial E_z}{\partial x} \quad (33.24c)$$

Now first derivatives must be 0 in region 1 (to the left of the boundary) and in region 2 (to the right of the boundary). We have already written the second in regions 1 and 2. Finally, they might be no derivative of the boundary, which we call region 3. Although we usually think of the boundary as being simply a boundary, in reality it's not. The physical properties change very rapidly but not infinitely fast. In any case, we can imagine that there is a very rapid, yet non-jerky, transition at the interface between region 1 and 2. In a short distance we can tell regions 1, 2, etc. Any field quantity like E_x , or B_z , etc., will make a similar kind of transition in region 3. In this region, the differential equations must still be satisfied, and it is by following the differential equations in this region that we can arrive at the needed "boundary conditions."

For instance, suppose that we have a boundary between vacuum (region 1) and glass (region 2). There is nothing to polarize in the vacuum, so $P_x = 0$. Let's say there is some polarization P_x in the glass. Between the vacuum and the glass there is a smooth, but rapid, transition. If we look at any component of P_x , say P_{xz} , it might vary as drawn in Fig. 33-5(a). Suppose now we take the last of our equations, Eq. (33.21). It involves derivatives of the components of P with respect to x , y , and z . The y - and z -derivatives are not interesting; nothing spectacular is happening in those directions. But the x -derivative of P_x will have some very large values in region 2, because of the tremendous slope of P_x . The derivative dP_{xz}/dx will have a sharp spike at the boundary, as shown in Fig. 33-5(b). If we manage to squeeze the boundary to an even sharper focus, the spike would get much bigger. If the boundary is really sharp for the waves we are interested in, the magnitude of dP_{xz}/dx in region 3 will be much, much greater than any other derivative we might look at. Look at the x -derivative of P in the wave near the boundary, so we ignore any variations of P other than those due to the boundary.

Now how can Eq. (33.21) be satisfied if there is a sharp corner space on the right-hand side? Only if we can somehow wrap up space on the either side. Something on the left-hand side must also be big. The only candidate is $\epsilon_0 E_{zx}$, because the vacuum water wave is the only place where there's no effect on the wave we just calculated. So $-\epsilon_0 \partial E_{zx}/\partial x$ must be as shown in Fig. 33-5(c), just a copy of Fig. 33-5(b). We then find

$$\epsilon_0 \frac{\partial E_{zx}}{\partial x} = -\frac{\partial P_{xz}}{\partial x}. \quad (33.25)$$

If we plug this equation with respect to x into Eq. (33.21), we conclude that

$$\epsilon_0(E_{zx} + P_{xz}) = -(\partial_{xz}^2 + P_{zz}). \quad (33.25)$$

In other words, the polarization, moving from region 1 to region 2, must be equal to the jump in P_{xz} .

We can rewrite Eq. (33.25) as

$$(\epsilon_0 E_{zx} + P_{xz}) = \epsilon_0 E_{zz} - P_{zz}, \quad (33.26)$$

which says that the quantity $(\epsilon_0 E_{zx} + P_{xz})$ has equal values in region 2 and region 1. People say: the quantity $(\epsilon_0 E_{zx} + P_{xz})$ is continuous across the boundary. We have, in this way, one of our boundary conditions.

Although we took as an illustration the case in which P was zero because region 1 was a vacuum, it is clear that the same argument applies for any two materials in the two regions, so Eq. (33.26) is true in general.

Let's now go through the rest of Maxwell's equations and see what each of them tells us. We take next Eq. (33.22a). There are no x -derivatives, so it doesn't tell us anything. (Remember that the fields themselves do not get especially large at the boundary; only the derivatives with respect to x can become so large that they dominate the equations.) Next, we look at Eq. (33.22b). Ah! There is an x -derivative! We have $\partial E_{yy}/\partial x$ on the left-hand side. Suppose it has a huge derivative. But wait a moment! There is nothing, on the right-hand side to cancel it with the derivative of E_y , except for one tiny jump in going from region 1 to region 2. If it did, it would be a spike on the left of Eq. (33.22b) but none on the right,

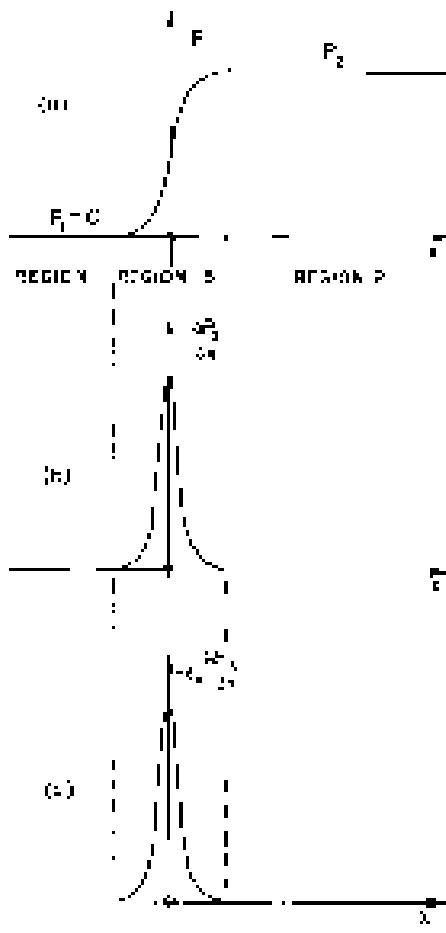


Fig. 33-5. The fields in the transition region (3), between two different materials (1) and (2).

and the equation would be false. So we have a contradiction.

$$E_{12} = E_{21}. \quad (33.27)$$

By the same argument, Eq. (33.26) gives

$$F_{12} = F_{21}. \quad (33.28)$$

This last result is just what we got in Eq. (33.20) by a time integral argument.

We go on to Eq. (33.21). The only term that could have a spike is $\partial B_1 / \partial t$. But there's nothing on the right to match it, so we conclude that

$$B_{12} = B_{21}. \quad (33.29)$$

On to the last of Maxwell's equations! Equation (33.24) gives nothing, because there are no ϵ -elements. Equation (33.25) has one, $-\epsilon' \partial B_1 / \partial x$, but again there is nothing to match it with. We get

$$B_{12} = B_{21}. \quad (33.30)$$

The last equation is quite similar, and gives

$$B_{12} = B_{21}. \quad (33.31)$$

The last three equations give us that $B = B_1$. We want to emphasize, however, that we get this result only when the materials on both sides of the boundary are magnetically—say, μ_1 —when we can neglect any magnetic effects at the interface. This is simply because there are, basically, except for magnetic ones. (We don't want to use the properties of materials in some other context.) Our argument is essentially the same as the one in Section 2 and those in Section 3. We leave out E and H together in Table 33-1. We can now use them to match the vectors in the two regions. We want to emphasize, however, that when we have just two materials we can always find a situation in which you have differential equations and you won't be able to find a sharp boundary between two regions where some property changes. For our present purposes, we will have easily derived the same equations by using arguments about the fluxes of current across the boundaries. You might say whether you can get the same result that way? But if you do get a result like that, well, we know you must check and don't see any reason to say something like “The physics of wave propagation at the boundary” you can just work with the equations.

33-4 The reflected and transmitted waves

Now we're ready to apply our boundary conditions to the waves we wrote down in Section 33-2. We find:

$$E_r = E_1 e^{i(k_1 z - \omega t + \delta_1)} \quad (33.32)$$

$$E_t = E_1 e^{i(k_1 z - \omega t + \delta_1)} \quad (33.33)$$

$$E_r = E_0 e^{i(k_1 z - \omega t + \delta_1)} \quad (33.34)$$

$$H_r = \frac{k'}{\omega} \times E_r \quad (33.35)$$

$$H_r = \frac{k'}{\omega} \times E_r \quad (33.36)$$

$$H_r = \frac{k''}{\omega} \times E_r \quad (33.37)$$

We have one further bit of knowledge: E is perpendicular to the propagation vector k for such waves.

Table 33-1

Boundary conditions at the surface of a dielectric

$$\begin{aligned} (\mu_1 + \mu_2) &= (\epsilon_1 + \epsilon_2) \\ (\mathbf{E}_1)_n &= (\mathbf{E}_2)_n \\ (\mathbf{E}_{1r})_n &= (\mathbf{E}_{2r})_n \\ \mathbf{B}_1 &= \mathbf{B}_2 \end{aligned}$$

(The surface is in the z -plane)

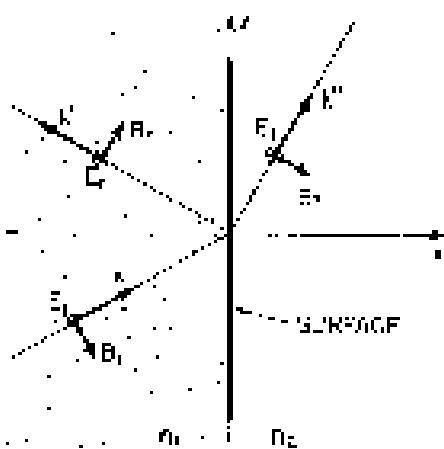


Fig. 33-8. Polarization of the reflected and transmitted waves when the E -field of the incident wave is perpendicular to the plane of incidence.

The results will depend on the direction of the E -vector (the "polarization" of the incoming wave). The analysis is simplest if we look separately at the s-polarized wave with its E -field E_s parallel to the "plane of incidence" (that is, the x - y -plane) and the p-polarized wave with the E -vector perpendicular to the plane of incidence. A wave of longitudinal polarization is just a linear combination of these two waves. In other words, the reflected and transmitted intensities are different for different polarizations, and this causes us to pick the two simplest cases and treat them separately.

We will carry through the analysis for an incoming wave polarized perpendicular to the plane of incidence and then just give you the result for the other. We are cheating a little by taking the simplest case, but the principle is the same for both. So we take that E_i has only a E_x -component, and since all the E -vectors are in the same direction we can leave off the vector signs.

So long as both materials are isotropic, this induces oscillations of charges in the material will also be in the x direction, and the E field of the transmit, loss, and reflected waves will have only E_x components. So for all the waves, E_y and B_z , and P_x and P_y are zero. The waves will have their E and H vectors as drawn in Fig. 33-8. (We are cutting a corner here in our original plan of getting everything from the equations. This result would also come out of the boundary conditions, but we can save a lot of algebra by using the physical argument. When you have some spare time, see if you can get the same result from the equations. It is clear that what we have said agrees with the equations, it is just that we have not shown that there are no other possibilities.)

Now our boundary conditions, Eqs. (33.26) through (33.31), give relations between the components of E and B in regions 1 and 2. In region 2 we have only two terms in total waves, but in region 1 we have two waves. Which one do we use? The fields in region 1 are, of course, the superposition of the fields of the incident and reflected waves. (Since each satisfies Maxwell's equations, so does the sum.) So when we set the boundary conditions, we must use that

$$E_1 = E_r \quad E_x = E_r$$

and similarly for the B 's.

For the polarization we are considering, Eqs. (33.26) and (33.28) give us no new information; only Eq. (33.27) is useful. It says that

$$E_1 + E_r = E_t$$

at the boundary, that is, for $x = 0$. So we have that

$$E_0 e^{i(kx - \omega t)} + E_0 e^{i(kx + \omega t)} = E_0 e^{i(kx - \omega t)} + E_0 e^{i(kx + \omega t)}, \quad (33.38)$$

where E_0 is not zero for all x and for all t . Suppose we look instead at $x = 0$. Then we have

$$E_0 e^{i\omega t} + E_0 e^{-i\omega t} = E_0 e^{i\omega t}.$$

This equation says that two non-harmonic waves, $e^{i\omega t}$ and $e^{-i\omega t}$, add to zero. That can happen only if all the oscillations have the same frequency. (It is impossible for three—or any number—of such waves with different frequencies to add to zero for all times.) So

$$\omega^T = \omega^I = \omega. \quad (33.39)$$

As we know already, the intensities of the reflected and transmitted waves are the same as that of the incident wave.

We should really have said "or" rather than "so" since some trouble by putting that in at the beginning, but we wanted to show you that it can also be got out of the equations. When you are doing a problem, it's usually the best thing to put everything you know into the waves right at the start and save yourself a lot of trouble.

By definition, the amplitude A^T is given by $A^T = \sqrt{A_x^2 + A_y^2}$, so we have also that

$$\frac{\omega^T}{\omega^I} = \frac{A^T}{A^I} = \frac{\sqrt{A_x^2 + A_y^2}}{\sqrt{A_x^2}} \quad (33.40)$$

Now look at Eq. (33.38) for $k = k_1$. Using again the same kind of argument we have just made, but this time based on the fact that the equation must hold for all values of n , we get that

$$k_y^2 = k_1^2 - k_n^2. \quad (33.41)$$

From Eq. (33.30), $k^{yy} = k_y^2$, so

$$k_y^2 = k_y^2 = k_1^2 + k_n^2.$$

Combining this with Eq. (33.41), we note that

$$k_y^2 = -k_1^2,$$

so that $k_y^2 = -k_1^2$. The positive sign makes no sense; that would just give a reflected wave, but another incident wave, and we said at the start that we were solving the problem of only one incident wave. So we have

$$k_y^2 = -k_1^2. \quad (33.42)$$

The two equations (33.41) and (33.42) give us that the angle of reflection is equal to the angle of incidence, as we expected. (See Fig. 33-3.) The reflected wave is

$$E_r = E_0 e^{i(k_1 x_1 - \omega_1 t)} \quad (33.43)$$

For the transmission wave we already have found

$$k_y^2 = k_0^2$$

and

$$\frac{k_y^2}{k_1^2} = \frac{k_0^2}{n_1^2}, \quad (33.44)$$

so we can divide these to find k_y^2 . We get

$$k_y^2 = k^{yy} = k_p^2 = \frac{k_0^2}{n_1^2} k_1^2 = k_0^2. \quad (33.45)$$

Suppose for a moment that n_1 and n_0 are real numbers (that the imaginary parts of the indices are very small). Then all the k 's are also real numbers, and from Fig. 33-3 we find that

$$\frac{k_0}{k} = \tan \theta_1, \quad \frac{k_y}{k} = \tan \theta_2. \quad (33.46)$$

From (33.46) we get that

$$n_1 \sin \theta_1 = n_0 \sin \theta_2, \quad (33.47)$$

which is, since θ_1 and θ_2 are right angles, something we already know. If the indices are not real, the tangent functions are complex, and we cannot trust Eq. (33.47). (We could still define the angles θ_1 and θ_2 by Eq. (33.46) and Snell's law, Eq. (33.45), would not this be general? But I do the "complex" also as complex numbers, thereby losing their simple geometric interpretation, as in Fig. 33-3. It is best to leave to the behavior of the waves by themselves, i.e., on E_p unless.)

So far we have learned nothing new. We have just had the simple-minded delight of getting some obvious answers from a complicated mathematical machinery. Now we are ready to find the amplitudes of the waves which we have not yet known. Using our results for the k 's and θ 's, the exponential factors in Eq. (33.36) can be canceled, and we get

$$E_0 = E_1 = E_0^T. \quad (33.48)$$

Since both E_0 and E_0^T are unknown, we need one more relationship. We must remember & use the boundary conditions. Our equations for E_2 and E_0 are no help because all the E 's are varying in x -space. So we must use the condition in Eq. (33.32), Eq. (33.49)

$$E_{2A} = E_{2B}$$

From Eqs. (33.44) through (33.50),

$$B_x = \frac{k_x E_0}{\omega}, \quad B_{x0} = \frac{k_x E_0}{\omega'}, \quad B_{z0} = \frac{k_z E_0}{\omega'}$$

Recalling that $\omega' = \omega - k_z$ and $k_z' = k_z + k_x$, we get that

$$E_0 + k_x B_{z0} = E_0'$$

But this is just Eq. (33.48) all over again! We've just wasted time when we could have stopped.

We could try Eq. (33.49), $B_{y0} = B_{z0}$, but there are two more unknowns, E_0' & B_y' . So there's only one equation left, Eq. (33.44), $B_{x0} = B_{y0}$. But the terms vanish,

$$B_y = -\frac{k_y E_0}{\omega}, \quad B_{y0} = -\frac{k_y E_0}{\omega'}, \quad B_{x0} = \frac{k_x E_0}{\omega'}$$
 (33.49)

Putting in E_0 , E_0' , and B_y , the wave expression for $x = 0$ (to level the boundary), the boundary condition is

$$\frac{k_x}{\omega} E_{0x} + \omega B_{y0} = \frac{k_x}{\omega'} E_{0x}' + \omega' B_{y0}' + \frac{k_x}{\omega'} E_{0x}'' + \omega' B_{y0}''$$

Again all ω 's and k_x 's are equal, so this reduces to

$$\kappa E_0 + \kappa' E_0' = \kappa'' E_0''$$
 (33.50)

This gives us an equation for the E 's that is different from Eq. (33.48). With the two, we can solve for E_0' and E_0'' . Recalculating that $\kappa'' = k_x$, we get

$$E_0' = \frac{\kappa}{k_x} - \frac{k_x^2}{\omega'^2} E_0$$
 (33.51)

$$E_0'' = \frac{2\kappa}{k_x} - \frac{\omega'^2}{\omega^2} E_0$$
 (33.52)

This, together with Eq. (33.5) or Eq. (33.46) for κ'' , give us what we wanted to know. We will discuss the consequences of this result in the next section.

If we begin with a wave polarized with its E -vector oriented in the plane of incidence, E will have both x - and y -components, as shown in Fig. 33-7. The analysis is straightforward but more complicated. (The work can be somewhat reduced by expressing things in terms of the magnetic fields, which are in the x -direction in the figure.)

$$|E'| = \frac{\kappa k_x - \omega'_x k_x}{\omega'^2 k_x + \omega^2 k_x^2} E_0$$
 (33.53)

and

$$E_0'' = \frac{2\kappa k_x}{\omega'^2 k_x - \omega^2 k_x^2} E_0$$
 (33.54)

Let's see whether our results agree with those we got earlier. Equation (33.3) is the result we worked out in Chapter 32 of Volume I for the ratio of the intensity of the reflected wave to the intensity of the incident wave. Then, however, we were considering only real indices. For real indices (and κ'), we can write

$$I_r = I_0 \cos \theta_i = \frac{n_{r1}}{n} \cos \theta_i$$

$$I_0' = I_0' \cos \theta_i = \frac{n_{r2}}{n} \cos \theta_i$$

Substituting in Eq. (33.51), we have

$$\begin{aligned} E_0' &= A_0 \cos \theta_i - A_0' \cos \theta_i \\ E_0 &= A_0 \cos \theta_i + A_0' \cos \theta_i \end{aligned}$$
 (33.55)

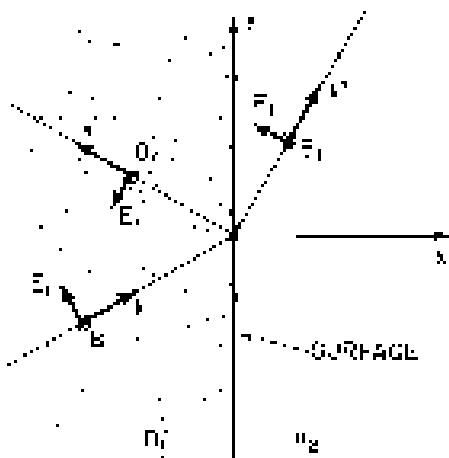


Fig. 33-7. Polarization of the waves when the E field of the incident wave is parallel to the plane of incidence.

which, very get, look the same as Eq. (33.3). To well, however, if we use Snell's law to get the α 's. Setting $n_1 = n$, $\sin \theta_1 / \sin \theta_2$, and multiplying the numerator and denominator by $\sin \theta_2$, we get

$$\frac{E_0}{E_1} = \frac{\cos \theta_1 \sin \theta_2}{\cos \theta_2 \sin \theta_1} = \frac{\sin \theta_1 \cos \theta_2}{\sin \theta_2 \cos \theta_1}.$$

The numerator and denominator are just the sines of $(\delta_1 - \delta_2)$ and $(\delta_1 + \delta_2)$; we get

$$\frac{E_0}{E_1} = \frac{\sin(\delta_1 - \delta_2)}{\sin(\delta_1 + \delta_2)}. \quad (33.56)$$

Since E_1 and E_0 are in the same material, the intensities are proportional to the squares of the electric fields, and we get the same result as before. Similarly, Eq. (33.57) is the same as Eq. (33.2).

For waves which travel in normal media say, $\delta_1 = 0$ and $\delta_2 = 0$, Equation (33.56) gives 0.0, which is not very useful. We can, however, go back to Eq. (33.55), which gives

$$\frac{I_0}{I_1} = \left(\frac{E_0}{E_1} \right)^2 = \left(\frac{\theta_1 - \theta_2}{\theta_1 + \theta_2} \right)^2. \quad (33.57)$$

This result, just as (33.55), applies for "either" polarization, since for normal incidence there is no system "phase of incidence."

33-6 Reflection from metals

We can now use our results to understand the interesting phenomenon of reflection from metals. Why is it that metals are shiny? We saw in the last chapter that metals have an index of refraction n which, for some frequencies, has a large imaginary part. Let's see what we would get for the reflected intensity when light comes from air with $n = 1$ and to a metal with $n = -i\kappa_0$. Then Eq. (33.55) gives (for normal incidence)

$$\frac{E_0}{E_1} = \frac{1 + i\kappa_0}{1 - i\kappa_0}.$$

For the intensity of the reflected wave, we want the square of the absolute value of E_0 and E_1 :

$$\frac{I_0}{I_1} = \frac{|E_0|^2}{|E_1|^2} = \frac{(1 + i\kappa_0)^2}{(1 - i\kappa_0)^2},$$

or

$$\frac{I_0}{I_1} = \frac{1 + 2i\kappa_0 + \kappa_0^2}{1 + 2i\kappa_0 + \kappa_0^2} = 1. \quad (33.58)$$

For a material with an index which is a pure imaginary number, there is 100 percent reflection.

Metals do not reflect 100 percent, but many do reflect visible light very well. In other words, the imaginary part of their indexes is very large. But we have seen that a large imaginary part of the index means a strong absorption. So there is a general rule that: any material gets to be a very good absorber at any frequency, the waves are strongly reflected at the surface and very little gets made to be absorbed. You can see this effect with strong dyes. Pure crystals of the strongest dyes have a "metallic" luster. Probably you have noticed that as the edge of a bottle of purple ink (the dried dye will give a golden metallic reflection) or the dried red ink will sometimes give a greenish metallic reflection. Red ink absorbs most the energy of "unabsorbed" light, so if the ink is very concentrated it will act just as strong surface reflector for the frequencies of green light.

You can easily show this effect, by covering a glass plate with red ink and letting it dry. If you direct a beam of white light at the back of the plate, as shown in Fig. 33-8, there will be a transmitted beam of red light and a reflected beam of green light.

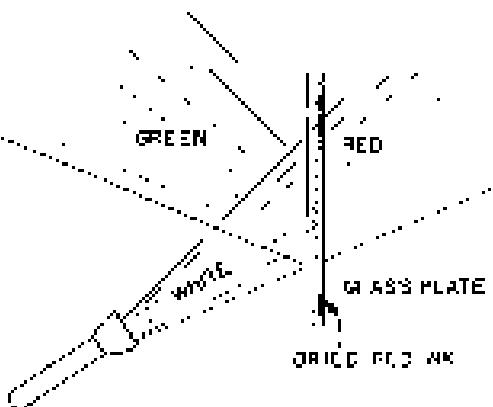


Fig. 33-8. A material which absorbs light strongly at the frequency ω also reflects light at that frequency.

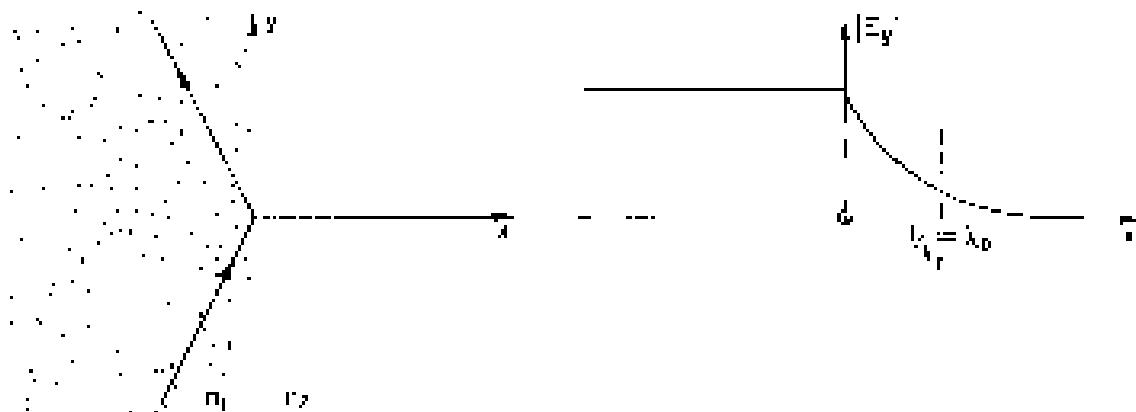


Fig. 22-9. Total internal reflection.

22-6 Total internal reflection

If light goes from a material like glass, with a real index of greater than 1, toward, say, water, an index n_2 equal to 1.33, Snell's law says that

$$\sin \theta_2 = n_2 \sin \theta_1.$$

The angle θ_2 of the transmitted wave becomes 90° when the incident angle θ_1 is equal to the "critical angle" ϕ_c , given by

$$\sin \phi_c = 1/n_2. \quad (22.59)$$

What happens for θ_1 greater than the critical angle? You know that there is total internal reflection. But now does that mean above?

Let's go back to Eq. (22.43) which gives the wave number k' for the transmitted wave. We would have

$$k'^2 = \frac{k^2}{n^2} - k_p^2.$$

Now $k_p = k \sin \phi_c$ and $k = \omega/c_0 \sin$

$$k'^2 = \frac{\omega^2}{c_0^2} (1 - n^2 \sin^2 \phi_c).$$

If $\phi_c < \theta_1$ is greater than one, k'^2 is negative and k' is a pure imaginary, try $-ik'$. You know by now what that means! The "transmitted" wave (Eq. 22.34) will have the form

$$E_r = E_0 e^{-ik' z} e^{i\omega t + k' x},$$

The wave amplitude either grows or decays exponentially with increasing x . Clearly, we want here is the negative sign. Then the amplitude of the wave to the right of the boundary will go as shown in Fig. 22-8. Notice that k_f is of the order of ω/c_0 , the free-space wavelength of the light. When light is totally reflected from the inside of a glass/rubber surface, the electric fields in the air, and they extend beyond the surface only a distance of the order of the wavelength of the light!

We can now see how to answer the following question. If a light wave in glass arrives at the surface at a large enough angle, it is reflected. If a thin piece of glass is brought up to the surface (so that the "reflectance" is at least disappears) the light is transmitted. Exactly when does this happen? Since there must be a continuous change from total reflection to no reflection, the answer, of course, is that if the air gap is so small that the exponential tail of the wave in the air has an appreciable strength at the second piece of glass, it will shake the electrons there and generate a new wave, as shown in Fig. 22-10. So in light it is the transient that occurs, our solution is now update, we should solve all the equations again for a thin layer of air between two regions of glass.)

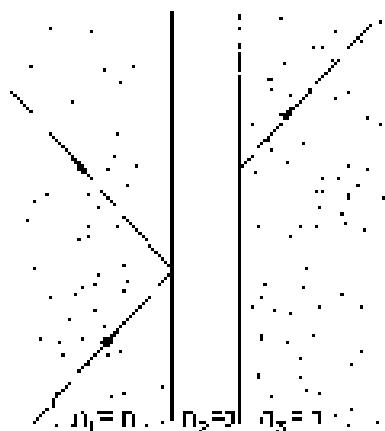


Fig. 22-10. If there is a small gap, "total" reflection is not "total"; a transmitted wave appears beyond the gap.

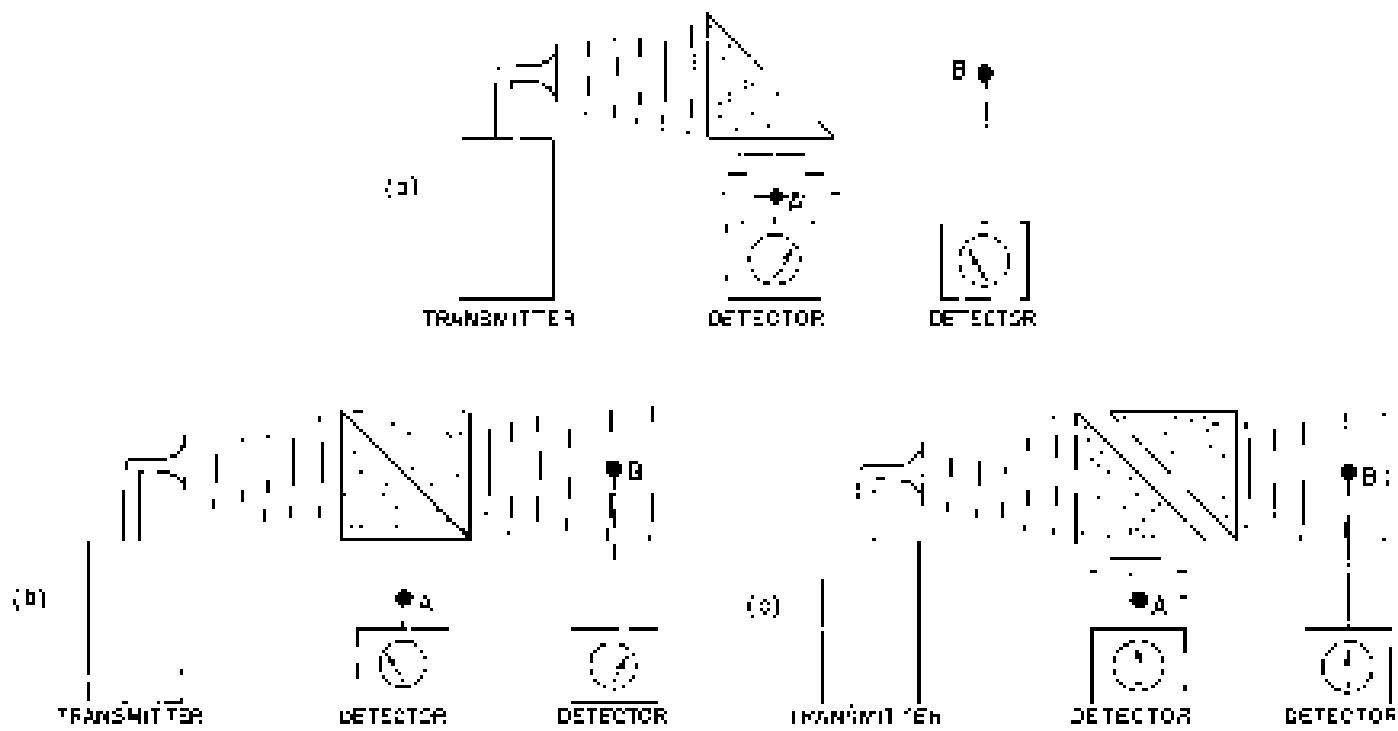


Fig. 30-11. A demonstration of the generation of internally reflected waves.

This transmission effect can be observed with ordinary light only if the air gap is very small (of the order of the wavelength of light, like 10^{-3} cm), but it is easily demonstrated with three centimeter waves. Then the exponentially decreasing field extends several centimeters. A microwave apparatus that shows the effect is drawn in Fig. 30-11. Waves from a small three centimeter transmitter are directed at a 45° prism of paraffin. The index of refraction of paraffin for these frequencies is 1.50, and therefore the critical angle is 41.5°. So the wave is readily reflected from the 45° face and is picked up by detector A, as indicated in Fig. 30-11(a). If a second paraffin prism is placed in contact with the first, as shown in part (b) of the figure, the wave passes straight through and is picked up at detector B. If a gap of a few centimeters is left between the two prisms, as in part (c), there are both transmitted and reflected waves. The electric field outside the 45° face of the prism in Fig. 30-11(c) can also be shown by bringing detector B to within a few centimeters of the surface.

The Magnetism of Matter

34-1 Diamagnetism and paramagnetism

In this chapter we are going to talk about the magnetic properties of materials. The material which has the most striking magnetic properties is iron, steel, cobalt, and tin; but magnetic properties are shared also by the elements nickel, vanadium, and others at sufficiently low temperatures (below 16°C) by gadolinium, as well as by a number of peculiar alloys. This kind of magnetism, called ferromagnetism, is sufficiently striking and complicated that we will discuss it in a special chapter. However, all ordinary substances do show some magnetic effects, although very small ones—a thousand to a million times less than the effects in ferromagnetic materials. Here we are going to describe ordinary magnetism, that is to say, the magnetism of substances other than the ferromagnetic ones.

This small magnetism is of two kinds. Some materials are attracted toward magnetic fields; others are repelled. Unlike the electrical effect in matter, which always causes dipoles to be attracted, there are two signs to the magnetic effect. These two signs can be easily shown with the help of a strong electromagnet which has one sharply pointed pole piece and one flat pole piece, as drawn in Fig. 34-1. The magnetic field is much stronger near the pointed pole than near the flat pole. If a small piece of material is fastened to a long string and suspended between the poles, there will, in general, be a small force on it. This small force can be seen by the slight displacement of the hanging material when the current is turned on. The few ferromagnetic materials are attracted very strongly toward the pointed pole; all other materials feel only a very weak force. Some are weakly attracted to the pointed pole; and some are weakly repelled.

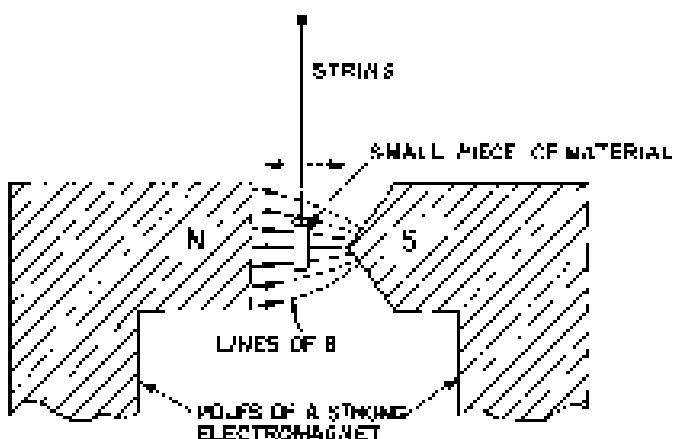


Fig. 34-1. A small spinner of beryllium is weakly repelled by the sharp pole; a piece of aluminum is attracted.

This effect is most easily seen with a small cylinder of beryllium, which is repelled from the high-field region. Substances which are repelled in this way are called diamagnetic. Bismuth is one of the strongest diamagnetic materials, but even with it, the effect is still quite weak. The magnetism is always very weak. If a small piece of aluminum is suspended between the poles, there is also a weak force, but never to the pointed pole. Substances like aluminum are called paramagnetic. (In such an experiment, oddly enough, it takes a tug when the magnet is turned on and off, and there can also be strong vibrations. You must be careful to look for the net displacement after the hanging object settles down.)

34-1 Diamagnetism and paramagnetism

34-2 Magnetic moments and angular momentum

34-3 The precession of atomic magnets

34-4 Demagnetism

34-5 Curie's theorem

34-6 Classical physics likes neither diamagnetism nor paramagnetism

34-7 Angular momentum in quantum mechanics

34-8 The magnetic energy of atoms

Review: Section 15-1, "The forces on a charged body moving at a charge."

We want now to discuss briefly the mechanisms of these two effects. First, in many substances the atoms have no permanent magnetic moments; either all the moments within each atom, because of it, so that the net moment of the atom is zero. The electron spins are excited, however, all exactly balance out, so that any particular atom has an average magnetic moment. In these circumstances, when a current or magnetic field-like environments are generated inside the atom by a dipole, according to Lenz's law, these currents flow in such a direction as to oppose the increasing field. So the induced magnetic moments of the atoms are oriented opposite to the magnetic field. This is the origin of diamagnetism.

Then there are some substances for which the atoms do have permanent magnetic moments, in which the electron spins and moments have a net total dipole moment that is not zero. Besides the diamagnetic field (which is always present), there is also the possibility of having up the individual atomic magnetic moments in the case, the moments try to line up with the magnetic field. In this way the permanent dipoles of a dipole are lined up by the external field, and the related magnetic field can induce the magnetic field. These are the paramagnetic substances. Paramagnetism is generally fairly weak, its susceptibility (magnetization) is relatively small compared with the forces from the valence electrons which try to align the order. It is as follows that paramagnetism is usually sensitive to the temperature. (The paramagnetism arising from the spins of the electrons responsible for conduction in a metal contributes an exception. We will not be discussing this phenomenon here.) For ordinary paramagnetism, the lower the temperature, the stronger the effect. There is more lining up at low temperatures when the aligning effects of the collisions are less. Diamagnetism, on the other hand, is more or less independent of the temperature. In any substance with atomic magnetic moments, there is a diamagnetic as well as a paramagnetic effect, but the paramagnetic effect usually dominates.

In Chapter 11 we described a ferroelectric material, in which all the electric dipoles get lined up by their own mutual electric fields. It is also possible to imagine the converse analog of ferroelectricity, in which all the atomic moments would line up and lock together. If you make calculations of how this should happen you will find that because the magnetic forces are so much smaller than the electric forces, thermal motions should knock out this alignment even at temperatures as low as a few tenth of a degree Kelvin. So it would be impossible at room temperature to have any permanent lining up of the magnets.

On the other hand, this is exactly what does happen in iron—it does get lined up. There is an attractive force between the magnetic moments of the different atoms of iron which is much, much greater than the direct magnetic interaction. It is an interaction of $\alpha \cdot \beta \cdot \gamma$, which can be explained only by quantum mechanics. It is about 10^3 times stronger than the direct magnetic interaction, and is what lines up the moments in ferromagnetic materials. We discuss this special interaction in the next chapter.

Show that we have tried to give you a qualitative explanation of diamagnetism and magnetostatics, we must confess ourselves and say that we are probably to understand the complete behavior of materials in any honest way from the point of view of classical physics. Such complete effects are a completely nonclassical phenomena. It is, however, possible to make some partially classical arguments and to get some general hint of what is going on. We might, in this way, You can make some classical arguments and get access to the behavior of the material, but these arguments are not "exact" in any sense because it is absolutely essential that quantum mechanics be involved in every one of these magnetic phenomena. On the other hand, there are situations, such as in a plasma or a region of space with many free electrons, where the electrons obey the laws of classical mechanics. And in those circumstances, some of the descriptions of classical magnetism are trustworthy. Also, the classical arguments are of some value for historical reasons. The first theories that people were able to guess at the meaning and behavior of magnetic materials, they used classical arguments. Finally, as we have already mentioned, classical mechanics can give us some useful guides.

so what might happen—even though the real terms, way to study this subject would be to learn quantum mechanics first; and then to understand the connection in terms of quantum mechanics.

On the other hand, we don't want to wait until we begin quantum mechanics inside out to understand a simple thing like this magnetism. We will have to lean on the classical mechanics as long as half showing what happens, realizing, however, that the arguments are really not yet set. We therefore make a series of theories about classical magnetism that will be three years apart so they will prove different things. Except for the last theorem, every one of them will be π true. Furthermore, they will all be wrong as a description of the physical world, because quantum mechanics is π true.

34-2 Magnetic moments and angular momentum

The first theorem we want to prove from classical mechanics is the following: If an electron is moving in a circular orbit (for example, revolving around a nucleus under the influence of a central force), there is a definite ratio between the magnetic moment and the angular momentum. Let's call J the angular momentum and μ the magnetic moment of the electron in the orbit. The magnitude of the angular momentum is the mass of the electron times the velocity times the radius. (See Fig. 34-2.) It is directed perpendicular to the plane of the orbit.

$$J = m v a. \quad (34.2)$$

(This is, of course, exactly right for nuclei, but it's a poor approximation for atoms, because for the electrons, v varies $\sim R$ is generally of the order $m^2/R = \pi/10$, or about 1 percent.)

The magnetic moment of the same orbit is the current times the area. (See Section 34-5.) The current is the charge per unit time which passes any point on the orbit, namely, the charge q times the frequency of rotation. The frequency is the velocity divided by the circumference of the orbit: so

$$f = q \frac{1}{2\pi}.$$

The area is πr^2 , so the magnetic moment is

$$\mu = \frac{q f}{2}. \quad (34.3)$$

It is also directed independently of the plane of the orbit. So J and μ are in the same direction:

$$\mu = \frac{q}{2m} J \text{ (orbit).} \quad (34.4)$$

There also happens to be another quantity, called the radius vector. For any particle moving up a coordinate axis, the magnetic moment is always along the line of the angular momentum. But just about the only exception is the electron. In the electron, the charge is negative, so the spin is $-q$; so the moment

$$\mu = -\frac{q}{2m} J \text{ (electron orbit).} \quad (34.4)$$

That's where we would expect classically and, in a fairly enough, it is also true quantum-mechanically. It's one of three things. However, if you keep going with the classical physics, you find other places where it goes the wrong direction, and it's a great game to try to remember which things are right and which things are wrong. We might as well give you immediately what is true as given in quantum mechanics. First, Eq. (34.4) is true for orbital motion, but that's not the only magnetism that exists. The electron also has a spin rotation about its own axis (something like the earth rotating on its axis), and as a result of that spin it has both an angular momentum and a magnetic moment. But for reasons that are purely quantum mechanical, there is no classical explanation—the ratio of μ

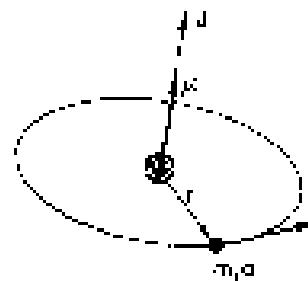


Fig. 34-2 For any circular orbit the magnetic moment μ is $\pi/2m$ times the angular momentum J .

to it for the electron spin is twice as large as it is for orbital motion of the spinning electron:

$$\mu = -\frac{e}{m} f(\text{orbital spin}), \quad (34.5)$$

In our atoms there are, generally speaking, several electrons and some combination of spin and orbit motions which builds up a total angular momentum and a total magnetic moment. Although there is no classical reason why it should be so, it is always true in quantum mechanics that (for an isolated atom) the direction of the magnetic moment is exactly opposite to the direction of the angular momentum. The ratio of the two is of course either $-g/m$ or $-g/2m$, but somewhere in between, because there is a mixture of the contributions from the orbits and the spins. We can write

$$\mu = -g \left(\frac{S+L}{2m} \right) J, \quad (34.6)$$

where g is a factor which is characteristic of the state of the atom. It would be 1 for a pure orbital motion, or 2 for a pure spin motion, or some other number in between for a complicated system like an atom. This formula does not, of course, tell us very much. It says that the magnetic moment is parallel to the angular momentum, but can have any magnitude. The form of Eq. (34.6) is convenient, however, because g —called the “Lande g -factor”—is a dimensionless constant whose magnitude is of the order of one. It is one of the joys of quantum mechanics to predict the g -factor for any particular atomic state.

You might also be interested in what happens in nuclei. In nuclei there are protons and neutrons which may have around an s or p orbital of order 10^{-14} of the same size, like an electron, have an intrinsic spin. Again the magnetic moment is parallel to the angular momentum. Only now the order of magnitude of the ratio of the two is what you would expect for a proton (or up quark) at a radius r_0 to be in Eq. (34.5) equal to the previous ones. That is, g is about the same for nuclei.

$$\mu = z \left(\frac{e r_0}{2m_p} \right) J, \quad (34.7)$$

where r_0 is the radius of the proton, and z —called the nuclear g -factor—is a number, about one, to be determined for each nucleus.

Another important difference for a nucleus is that the spin (or magnetic moment) of the proton does not have a spin-orbit of $\frac{1}{2}$, as the neutron does. For a proton, $g = 1/2m$. Surprisingly enough, the neutron then has a spin magnetic moment and its magnetic moment relative to its angular momentum is $S_n = 1.0$ e. The neutron, in other words, is not exactly neutral in the magnetic sense. That is, like a little magnet, it has the kind of magnetic moment that a turning negative charge would have.

34-3 The precession of atomic magnets

One of the consequences of having the magnetic moment proportional to the angular momentum is that an atomic magnet, plus whatever magnetic field it will produce, will rotate as it moves. Suppose that we have the magnetic moment μ suspended freely in a uniform magnetic field B . It will feel a torque τ , equal to $\mu \times B$, which tends to bring it in line with the field direction B . But the outside magnet is in general not at right angles the momentum J . Therefore the torque due to the magnetic field will not bring the magnet to line up. Instead, the magnet will precess, as we saw when we analyzed a gyroscope in Chapter 20 of Volume I. The angular momentum J and with it the magnetic moment μ precess about a axis parallel to the magnetic field. You can find the rules of precession by the same method as we used in Chapter 20 of the first volume.

Suppose that in a small time Δt the angular momentum changes from J to J' , as drawn in Fig. 34.3, staying always at the same angle θ with respect to the direction of the magnetic field B . Let's call ω_p the angular velocity of the precession, so that in the time Δt the angle of precession is $\omega_p \Delta t$. From the geometry of the figure

lighter, we see that the change of angular momentum in the time Δt is

$$\Delta I = (I \sin \theta) \omega_0 \Delta t.$$

So the rate of change of the angular momentum is

$$\frac{dI}{dt} = \omega_0 I \sin \theta, \quad (34.6)$$

which must be equal to the torque

$$\tau = \mu B \sin \theta. \quad (34.7)$$

The angular velocity of precession is then

$$\omega_p = \frac{\mu}{J} B. \quad (34.8)$$

Substituting μ/J from Eq. (34.6), we see that for an atomic system

$$\omega_p = g \frac{\mu B}{J_m}, \quad (34.9)$$

the precession frequency is proportional to B . It is handy to remember that for an atom (or electron)

$$f_p = \frac{\omega_p}{2\pi} = (1.1 \text{ megacycles/gauss})B, \quad (34.10)$$

and that for a nucleus

$$f_p = \frac{\omega_p}{2\pi} = (0.76 \text{ kilocycles/gauss})B. \quad (34.11)$$

(The formulas for atoms and nuclei are different only because of the different conversions for g for the two cases.)

According to the classical theory, then, the electron orbits and spins in an atom would precess in a magnetic field. Is it also true quantum mechanically? It is essentially true, but the meaning of the "precession" is different. In quantum mechanics one cannot talk about the *direction* of the angular momentum in the same sense as one does classically; nevertheless, there is a very close analogy and close that we continue to call it "precession." We will discuss it later when we talk about the quantum-mechanical point of view.

34-4 Diamagnetism

Next we want to look at diamagnetism from the classical point of view. It will be most apparent in several ways. In one of the nice ways is the following. Suppose that we already turn on a magnetic field in the vicinity of an atom. As the magnetic field changes an electric field E generated by magnetic induction. From Faraday's law, the line integral of E around any closed path is the rate of change of the magnetic flux through the path. Suppose we pick a path P which is a circle of radius r concentric with the center of the atom, as shown in Fig. 34-3. The average tangential electric field E around this path is given by

$$\delta E_{\text{av}} = - \frac{\delta}{\delta t} (\partial \Phi / \partial r),$$

and there is a circulating electric field whose strength is

$$E = - \frac{r}{2} \frac{\delta \Phi}{\delta t}$$

The tangential electric field acting on an electron in the atom precesses in torque equal to $-e_E r \epsilon$, which must equal the rate of change of the angular momentum dI/dt :

$$\frac{dI}{dt} = \frac{e_r^2 r^2}{2} \frac{d\Phi}{dt}. \quad (34.12)$$

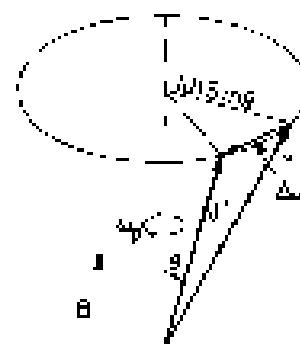


Fig. 34-3. An object with angular momentum J and a parallel magnetic moment μ placed in a magnetic field B precesses with the angular velocity ω_p .

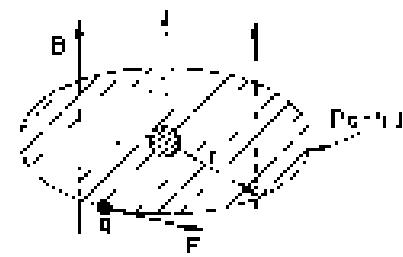


Fig. 34-4. The induced electric fields on the electrons in an atom.

Integrating with respect to time from zero field, we find that the change in angular momentum due to turning on the field is

$$\Delta J = \frac{q\epsilon^2}{2} R. \quad (34.15)$$

This is the extra angular momentum due to the turning on the field, as required.

The scaled step function makes the value magnetic moment which, as we know, is an orbital angular momentum times the angular momentum. The induced diamagnetic moment is

$$\Delta \mu = -\frac{q\epsilon}{2m} \Delta J = -\frac{q\epsilon^2}{4m} R. \quad (34.16)$$

The minus sign for $\Delta \mu$ can see is right by using Lenz's law; since that the added moment is opposite to the magnetic field.

We would like to write Eq. (34.16) a little differently. The ϵ which appears is the radius from z -axis through the atom parallel to B , so if B is along the z -direction, it is $\epsilon^2 = r^2$. If we consider spherically symmetric atoms (or average over atoms with their natural axes in all directions) the average of $x^2 + y^2$ is $2/3$ of the average of the square of the true radial distance from the center point of the atom. It is therefore usually more convenient to write Eq. (34.16) as

$$\Delta \mu = -\frac{q^2}{6m} \langle r^2 \rangle_{atom} B. \quad (34.17)$$

In any case, we have found an induced atomic moment proportional to the magnetic field B and opposing it. This is diamagnetism of matter. It is this magnetic effect that is responsible for the small force on a piece of metal in a non-uniform magnetic field. (You could compute the force by working out the energy of the induced moments in the field and seeing how the energy changes as the metal is moved into or out of the high-field region.)

We are still left with the problem: What is the mean square radius, $\langle r^2 \rangle_{atom}$? Classical mechanics cannot supply an answer. We must go back and start over with quantum mechanics. In an atom we cannot really say where an electron is, only know the probability that it will be at some place. If we interpret $\langle r^2 \rangle_{atom}$ as an average of the square of the distance from the center for the probability distribution, the definition is not given by symmetry reasons, as is just the same as for $\langle r \rangle$ (Eq. 34.17). The exception, of course, is the moment by one electron. The total moment is given by the sum over all the electrons in the atom. The surprising thing is that the classical argument and quantum mechanics give the same result. (Through, as was half said, the class argument that gives Eq. (34.17) is not really valid in classical mechanics.)

The second magnet's effect occurs even when an atom already has an angular momentum. Then the system will precess in the magnetic field. As the whole atom precesses, it takes up an additional small update velocity, and that some turning gives a small current which represents a correction to the magnet moment. This is just the diamagnetic effect represented in another way. But we don't really have to worry about that when we talk about paramagnetism. If the diamagnetic effect is (as computed, as we have done here), we don't have to worry about the fact that there is an extra little current from the precession. That has already been included in the diamagnetic term.

34-5 Larmor's theorem

We can already conclude something from our results so far. First of all, in the classical theory the moment μ was always proportional to J , with a given constant of proportionality for a particular atom. There wasn't any spin of the electrons and the constant of proportionality was always $-q/mc$; that is to say, in Eq. (34.6) we should set $g = 1$. The ratio of μ to J was independent of the internal motion of the electrons. Thus according to the classical theory, all systems

of electrons would precess with the same angular velocity. (This is not true in quantum mechanics.) This result is related to a theorem in classical mechanics that we would now like to prove. Suppose we have a group of electrons which are all held together by attraction toward a central point, so the electrons are attracted by a nucleus. The electrons will also be interacting with each other, and can, in general, have complicated motions. Suppose you have solved for the motions with no magnetic field and then want to know what the motions would be with a weak magnetic field. The theorem says that the motion with a weak magnetic field is always one of the no-field solutions with an added rotation, about the axis of the field, with the angular velocity $\omega_0 = qB/2m$. (This is the same as ω_0 , if $B = 1$.) There are, of course, many possible motions. The point is that for any \dot{r} motion without the magnetic field there is a corresponding motion in the field, which is the original motion plus a uniform rotation. This is called Larmor's theorem, and ω_0 is called the Larmor frequency.

We would like to show how the theorem can be proved, but we will let you work out the details. Take, first, one electron in a central force field. The force on it is just $-F(r)$, directed toward the center. If we now turn on a uniform magnetic field, there is an additional force, $qv \times B$; so the total force is

$$\mathbf{F}(r) = -q\mathbf{v} \times \mathbf{B}. \quad (34.18)$$

Now let's look at the same system from a coordinate system rotating with angular velocity ω about an axis through the center of force and parallel to \mathbf{B} . This is no longer an inertial system, so we have to put in the proper pseudoforces—the centrifugal and Coriolis forces we talked about in Chapter 19 of Volume I. We found there that in a frame rotating with angular velocity ω , there is an apparent radial force proportional to v_r , the radial component of velocity:

$$F_r = -2mv_r\omega, \quad (34.19)$$

And there is an apparent radial force which is given by

$$F_\theta = mv^2/r = 2mv_r\omega, \quad (34.20)$$

where v is the tangential component of the velocity, measured in the rotating frame. (The radial component v_r for rotating and inertial frames is the same.)

Now for small enough angular velocities (that is, if $2\omega \ll v_r$), we can neglect the first term (centrifugal) in Eq. (34.20) in comparison with the second (Coriolis). Then Eqs. (34.19) and (34.20) can be written together as

$$\mathbf{F} = -2mv_r\omega \hat{e}_r + \omega \hat{e}_\theta. \quad (34.21)$$

If we now combine a rotation and a magnetic field, we must add the forces in Eq. (34.21) to that in Eq. (34.18). The total force is

$$\mathbf{F}(r) = -q\mathbf{v} \times \mathbf{B} - 2mv_r\omega \hat{e}_r + \omega \hat{e}_\theta. \quad (34.22)$$

[we reverse the cross product and the sign of Eq. (34.21) to get the \hat{e}_θ term]. Adding up our results, we see that all

$$2mv_r\omega = -qB$$

the two terms on the right cancel, and in the moving frame the only force is $\mathbf{F}(r)$. The motion of the electron is just the same as with no magnetic field: $\omega_0\hat{e}_\theta$, of course, no rotation. We have proved Larmor's theorem for one electron. Since the proof assumes a small ω , it also means that the theorem is true only for weak magnetic fields. The only thing we could ask you to improve on is to do the case of many electrons mutually interacting with each other, but still in the same central field, and prove the same theorem. So no matter how complex our system, if it has a central field the theorem is true. But that's the end of the classical mechanics, because it isn't true in fact that the motions precess in that way. The precession frequency ω_0 of Eq. (34.11) is only equal to ω_0 if g happens to be equal to 1.

34-6 Classical physics gives neither diamagnetism nor paramagnetism

Now we would like to demonstrate that according to classical mechanics there can be no diamagnetism and no paramagnetism at all. It sounds crazy. First, we have proven that, if you take particles, the magnetism, interacting outside, can exist, and now we are going to prove that it is a tautology. Yet! We're going to prove that if you follow the classical mechanics framework, there are no such magnetic effects as diamagnetism. If you start a classical argument in a certain place and thought it through, you can get any answer you want. But the only logical and correct proof shows that there is no magnetic effect whatever.

It is a consequence of classical mechanics that if you have any kind of system: a gas with electrons, motions, and whatever—keep it in tact so that the whole thing can't turn, there will be no magnetic effect. It is possible to have a magnetic effect if you have an isolated system, like a star held together by itself, which can start rotating when you act on the magnetic field. But if you have a piece of material that's held in place so that it can't start spinning, then there will be no magnetic effects. What we mean by holding down the spin is summarized this way: At a given temperature we suppose that there is only one state of thermal equilibrium. The Postulate then says that if you turn on a magnetic field and wait for the system to get into thermal equilibrium, there will be no paramagnetism or diamagnetism—there will be no induced magnetic moment. (Postulate) According to statistical mechanics the probability that a system will have any given state of motion is proportional to $e^{-E/kT}$, where E is the energy of that motion. Now what is the energy of motion? For a particle moving in a constant magnetic field, the energy is the ordinary potential energy plus $mv^2/2$, with nothing additional for the magnetic field. [You know that the forces from electromagnetic fields are $q(E - v \times B)$, and that the rate of work $F \cdot v$ is just $qE \cdot v$, which is not affected by the magnetic field.] So the energy of a system, whether it is in a magnetic field or not, is always given by the kinetic energy plus the potential energy. Since the probability of any motion depends only on the energy—that is, on the velocity and position—it is the same whether or not there is a magnetic field. For thermal equilibrium, therefore, the magnetic field has no effect. If we have one system in a box, and then have another system in a second box, this time with a magnetic field, the probability of any particular velocity at any point in the first box is the same as in the second. If the first box has an average circulating current (which it will not have if it is in equilibrium with the stationary walls), there is no average magnetic moment, since in the second box all the motions are the same. There is no average magnetic moment there either. Hence, if the temperature is kept constant and thermal equilibrium is re-established after the field is turned on, there can be no magnetic moment induced by the field—according to classical mechanics. We can only get a satisfactory understanding of magnetic phenomena from quantum mechanics.

Unfortunately, we cannot assume that you have a thorough understanding of quantum mechanics, so this is hardly the place to discuss the matter. On the other hand, we don't always have to learn something first by learning the exact rules and then by learning how they are applied in different cases. Almost every subject that we have taken up in this course has been treated in a different way. In the case of electrodynamics we wrote the Maxwell equations on "Page One," and that did not tell all the consequences. That's one way. But we will not now try to begin a new "Page One," writing the equations of quantum mechanics and deducing everything from them. We will just have to tell you some of the consequences of quantum mechanics, before you begin to see they come from. So here we go.

34-7 Angular momentum in quantum mechanics

We have already given you a relation between the magnetic moment and the angular momentum. That's pleasant. But what is the magnetic moment and the angular momentum *now* in quantum mechanics? In quantum mechanics it turns out to be best to define things like angular momentum in terms of the other concepts, such as energy, in order to make sure that one knows what it means. Now,

It is easy to define a magnetic moment in terms of energy, because the energy of a moment in a magnetic field is, in the classical theory, $\mu \cdot B$. Therefore, the following definition has been taken in quantum mechanics: If we calculate the energy of a system in a magnetic field and we find that it is proportional to the field strength (for small fields), the coefficient is called the component of magnetic moment in the direction of the field. (We don't have to get so elegant for our work now; we can still think of the magnetic moment in the ordinary, or some extent classical, sense.)

Now we would like to discuss the idea of angular momentum in quantum mechanics—or rather, the characteristics of what in quantum mechanics is called angular momentum. You see, when you go to new kinds of laws, you can't just assume that each word is going to mean exactly the same thing. You may think, say, "Oh, I know what angular momentum is. It's that thing that is always a torque." But what's a torque? In quantum mechanics we have to give new definitions of old quantities. It would, therefore, be really best to call it by some other name such as "quantum angular momentum," or something like that, because it is the angular momentum as defined in quantum mechanics. But I want a single quantity in quantum mechanics which is identical to our old idea of angular momentum when the system becomes large enough, that is, for something as extreme as. We might as well just call it angular momentum. With that understanding, this next thing that we're about to do is very simple. In mechanics, it is nothing which in a large system we recognize as angular momentum or class mechanics.

First, we take a system in which angular momentum is conserved, such as an atom all by itself in empty space. Now such a thing (like the earth spinning on its axis) could, in the ordinary sense, be spinning around any axis one wished to choose. And for a given spin, there could be many different "states," all of the same energy, each "state" corresponding to a particular direction of the axis of the angular momentum. So in the classical theory, with a given angular momentum, there is an infinite number of possible states, all of the same energy.

If we look at quantum mechanics, however, that's several strange things happen. First, the number of states in which such a system can exist is limited—there is only a finite number. If the system is small, the finite number is very small, and if the system is large, the finite number gets very, very large. Second, we cannot directly a "state" by giving the direction of its angular momentum, or, only by giving the component of the angular momentum along some direction—say in the z-direction. Classically, an object with a given total angular momentum J could have, for its z-component, any value $J_{z,\text{classical}} = J \sin \theta$. But, qualitatively, for the general of angular momentum, we have only certain discrete values. Any given system—a particle, atom, molecule, or anything—with a given energy, has a also a certain number j , and its component of angular momentum can only be one of the following set of values:

$$\begin{aligned} & j \\ & j_z = j\hbar \\ & j_z = 2j\hbar \\ & \dots \\ & -(j - 2)\hbar \\ & (j - 1)\hbar \\ & -j\hbar \end{aligned} \tag{Q4.23}$$

The largest z-component is j times \hbar ; the next smaller is one unit of \hbar less, and so on down to $-j\hbar$. The number j is called "the spin of the system." Some people call it the "total angular momentum quantum number"; but we'll call it the "spin."

You may be worried that what we are saying can only be true for some "special" axis. But that is not so. For a system whose spin is j , the component of angular momentum along any axis can have only one of the values in (Q4.23). Although it is quite mysterious, we ask you just to accept it for the moment. We

will come back and discuss this point later. You may at least be pleased to see that the z-component goes from some number to minus the same number; so that we at least don't have to decide which is the plus direction of the axis. (Actually, it was this that it went from + to minus a different amount, that would be just as mysterious, because we wouldn't have been able to define the axis, pointing the other way.)

Now if the z-component of angular momentum never goes down by integers from $-j$ to $+j$, then j must be an integer. Not. Not quite: j must be an integer. It's only the difference between j and $-j$ that must be an integer. So, in general, the spin j is either an integer or a half-integer, depending on whether $2j$ is even or odd. Take, for instance, a nucleus like lithium, which has a spin of three-halves, $j = 3/2$. Then the angular momentum around the z axis, in units of \hbar , is one of the following:

$$\begin{aligned} &+3/2 \\ &+1/2 \\ &0 \\ &-1/2 \\ &-3/2 \end{aligned}$$

These are four possible states, each of the same energy, if the nucleus is in empty space with no external fields. If we have a system whose spin is two, then the z-component of angular momentum has only the values, in units of \hbar ,

$$\begin{aligned} &2 \\ &1 \\ &0 \\ &-1 \\ &-2. \end{aligned}$$

If you count now many states (including degeneracy), there are $(2j+1)$ possibilities. In other words, if you tell me the energy and also the spin j , it turns out that there are exactly $(2j+1)$ states with that energy, each state corresponding to one of the different possible values of the z-component of the angular momentum.

We would like to add one other fact. If you pick out any state of knowing at random and measure the z-component of the angular momentum, then you may get any one of the possible values, and each of the values is equally likely. All of the states are in fact single states, and each is just as good as any other. Each one has the same "weights" in the world. (We are assuming that nothing has tendency to sort out a special sample.) This fact has, incidentally, a simple classical analog. If you ask the same question classically: What is the line broadening of a particular z-component of angular momentum if you take a random sample of systems, all with the same total angular momentum? - the answer is that all values from the maximum to the minimum are equally likely. (You can easily work this out.) The classical result corresponds to the equal probability of the $(2j+1)$ possibilities in quantum mechanics.

From what we have so far, we can get another interesting and somewhat surprising conclusion. In certain classical calculations the quantity that appears in the final result is the square of the magnitude of the angular momentum J ; i.e. other words, $J \cdot J$. It turns out that it is often possible to get at the correct quantum-mechanical formula by using the classical calculation and the following simple rule: Replace $J^2 = J \cdot J$ by $\sigma_J = 1/6^2$. This rule is commonly used, and actually gives the correct result, but not always. We can give the following argument to show why you might expect this rule to work.

The scalar product $J \cdot J$ can be written as

$$J \cdot J = J_x^2 + J_y^2 + J_z^2$$

Since it is a scalar, it should be the same for any orientation of the spin. Suppose we pick samples of any given atomic system at random and make measurements of J_x^2 or J_y^2 or J_z^2 , the average value should be the same for each. (There is no special distinction for any one of the directions.) Therefore, the average of $J \cdot J$ is just

equal to three times the average of any component squared, say of J_z ,

$$\langle J \cdot J \rangle_{av} = 3\langle J_z^2 \rangle_{av}.$$

But since $J \cdot J$ is the same for all orientations, its average is, of course, just its constant value; we have

$$J \cdot J = 3\langle J_z^2 \rangle_{av}. \quad (Q4.24)$$

If we now say that we will use the same equation for quantum mechanics, we can easily find $\langle J_z^2 \rangle_{av}$. We just have to take the sum of the $(j_y + 1)$ possible values of J_z^2 , and divide by the total number;

$$\langle J_z^2 \rangle_{av} = j^2 - (j+1)^2 + \frac{(-j)^2 + (-j+1)^2}{2j+1}. \quad (Q4.25)$$

For a system with a spin of $1/2$, it goes like this:

$$\langle J_z^2 \rangle_{av} = (3/2)^2 + (1/2)^2 - \frac{(-1/2)^2 + (-3/2)^2}{2} = \frac{3}{4} \text{ J}^2.$$

We conclude that

$$J \cdot J = 3\langle J_z^2 \rangle_{av} = 3j^2 \approx 3(j+1)^2.$$

We will leave it to you to show that Eq. (Q4.25), together with Eq. (44.24), gives the general result

$$J \cdot J = j(j+1)h^2. \quad (Q4.26)$$

Although we would think classically that the largest possible value of the component of J is just the magnitude of J —namely, $\sqrt{J \cdot J}$ —quantum mechanically the maximum of J_z is always a little less than that, because j_h is always less than $\sqrt{(j+1)h}$. The angular momentum is said “completely along the z-direction.”

34-8. The magnetic energy of atoms

Now we want to talk again about the magnetic moment. We have said that in quantum mechanics the magnetic moment of a particular atomic system can be written in terms of the angular momentum by Eq. (31.8);

$$\mu_z = -q \left(\frac{\hbar}{2me} \right) J_z, \quad (Q4.27)$$

where $-q$ and m are the charge and mass of the electron.

An atomic magnet placed in an external magnetic field will have an excess magnetic energy which depends on the component of its magnetic moment along the field direction. We know that

$$U_{\text{mag}} = -\mu \cdot \theta. \quad (G4.28)$$

Choosing our basis along the direction of θ ,

$$U_{\text{mag}} = -\mu_z \theta_z. \quad (G4.29)$$

Using Eq. (Q4.27), we see that

$$U_{\text{mag}} = q \left(\frac{\hbar}{2me} \right) J_z \theta_z.$$

Quantum mechanics says that J_z can have only certain values: j_h , $(j+1)h$, ..., $-jh$. Therefore, the magnetic energy of an atomic system is not arbitrary; it can have only certain values. Its maximum value, for instance, is

$$q \left(\frac{\hbar^2}{2me} \right) j_h h.$$

The quantity μ_B can be usually given the name "the Bohr magneton" and written as:

$$\mu_B = \frac{e\hbar}{2mc}$$

The possible values of the magnetic energy are

$$E_{m,\pm} = \mu_B m_s \frac{\hbar c}{2}$$

where m_s takes on the possible values j_z , $j_z = +\frac{1}{2}\hbar$, $j_z = -\frac{1}{2}\hbar$, $j_z = +\frac{3}{2}\hbar$, etc.

In other words, the energy of an atomic system is changed when it is put in a magnetic field by an amount that is proportional to the field, and proportional to j_z . We say that the energy of an atomic system is "split into $(2j+1)$ levels" by a magnetic field. For instance, an atom whose energy is E_0 outside a magnetic field and whose j is $3/2$, will have four possible energies when placed in a field. We can show these energies by an energy-level diagram just as shown in Fig. 31-5. Any particular atom can have only one of the four possible states in a given field B . That's what quantum mechanics says about the behavior of an atomic system in a magnetic field.

The simplest "atomic" system is a single electron. The spin of an electron is $1/2$, so there are two possible states, $j_z = +\frac{1}{2}\hbar$ and $j_z = -\frac{1}{2}\hbar$. For an electron at rest (no orbital motion), the spin magnetic moment has a p -value of μ , so the magnetic energy can be either $\pm \mu B$. The possible energies in a magnetic field are shown in Fig. 34-6. Speaking loosely we say that the electron either has its spin "up" (along the field) or "down" (opposite the field).

For systems with higher spins, there are more states. We can think that the spin is "up" or "down" or "clockwise," some "angle," in between, depending on the value of j_z .

We will use these quantum mechanical results to discuss the magnetic properties of materials in the next chapter.

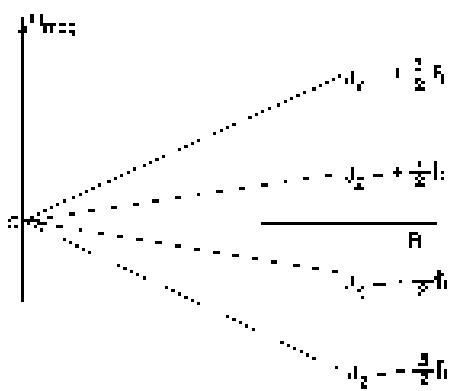


Fig. 34-5. The possible magnetic energies, E_m , of an atomic system with a spin of $3/2$ in a magnetic field B .

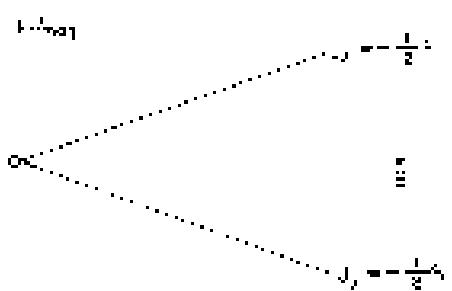


Fig. 34-6. The two possible energy states of an electron in a magnetic field B .

Paramagnetism and Magnetic Resonance

35-1 Quantized magnetic states

In the last chapter we described how in quantum mechanics the angular momentum of a thing does not have an arbitrary direction, but its component along a given axis can take on only certain discrete, spatial, discrete values. It is a shocking and peculiar thing. You may think that, perhaps, we should not believe such things, until your mind is more advanced and ready to accept this kind of an idea. Actually, your mind will never become more advanced—in the sense of being able to accept such a thing easily. There is no way of making it "intelligible" that isn't so subtle our brains just can't seem to get it. It is more complicated than the thing you were trying to understand. The behavior of matter on a small scale, as we have remarked many times, is different from anything that you are used to and is very strange indeed. As we proceed with classical physics, it is a good idea to try to get a growing acquaintance with the behavior of things on a small scale, at first as a kind of experience and not any deep understanding. Understanding of these matters comes very slowly, if at all. Of course, one does not better able to know what's going to happen in a quantum-mechanical situation—if that is what understanding means—but one never gets a comfortable feeling that these quantum-mechanical rules are "natural." Of course they are, but they are not natural to our own experience at an ordinary level. We should explain that the attitude that we are going to take with regard to this is that angular momentum is quite different from many of the other things we were talked about. We are not going to try to "explain" it, but we must at least tell you what happens; it would be dishonest to cover the thermagnetic properties of materials without mentioning the fact that the classical description of magnetism of angular momentum and magnetic moments—is incorrect.

One of the nice, shocking and disturbing features about quantum mechanics is this: if you take the angular momentum along any particular axis you find that it is always an integer or half-integer value \hbar . This is so no matter which axis you take. The subtleties involved in this curious fact—that you can take any other axis and still find the component of it is also limited to the same set of values—we will leave to a later chapter, when you will experience the delight of seeing how this apparent paradox is ultimately resolved.

We will now just repeat the fact that for every atomic system there is a number J , called "the spin of the system"—which must be an integer or a half-integer—and that the component of the angular momentum along any particular axis will always have one of the following values between $+\hbar$ and $-\hbar$:

$$\vec{J}_z = \text{one of } \begin{cases} \hbar & \\ \hbar - 1 & \\ \hbar - 2 & \\ \vdots & \cdots \\ -\hbar + 2 & \\ -\hbar + 1 & \\ -\hbar & \end{cases} \quad (35.1)$$

We have also mentioned that every simple atomic system has a magnetic moment which has the same direction as the angular momentum. This is true not only for atoms and nuclei but also for the fundamental particles. Each fundamental particle has its own characteristic value of J and its magnetic moment

35-1 Quantized magnetic states

35-2 The Stern-Gerlach experiment

35-3 The Rabi molecular-beam method

35-4 The paramagnetism of bulk materials

35-5 Cooling by adiabatic demagnetization

35-6 Nuclear magnetic resonance

Review: Chapter 11, Atomic Discreteness

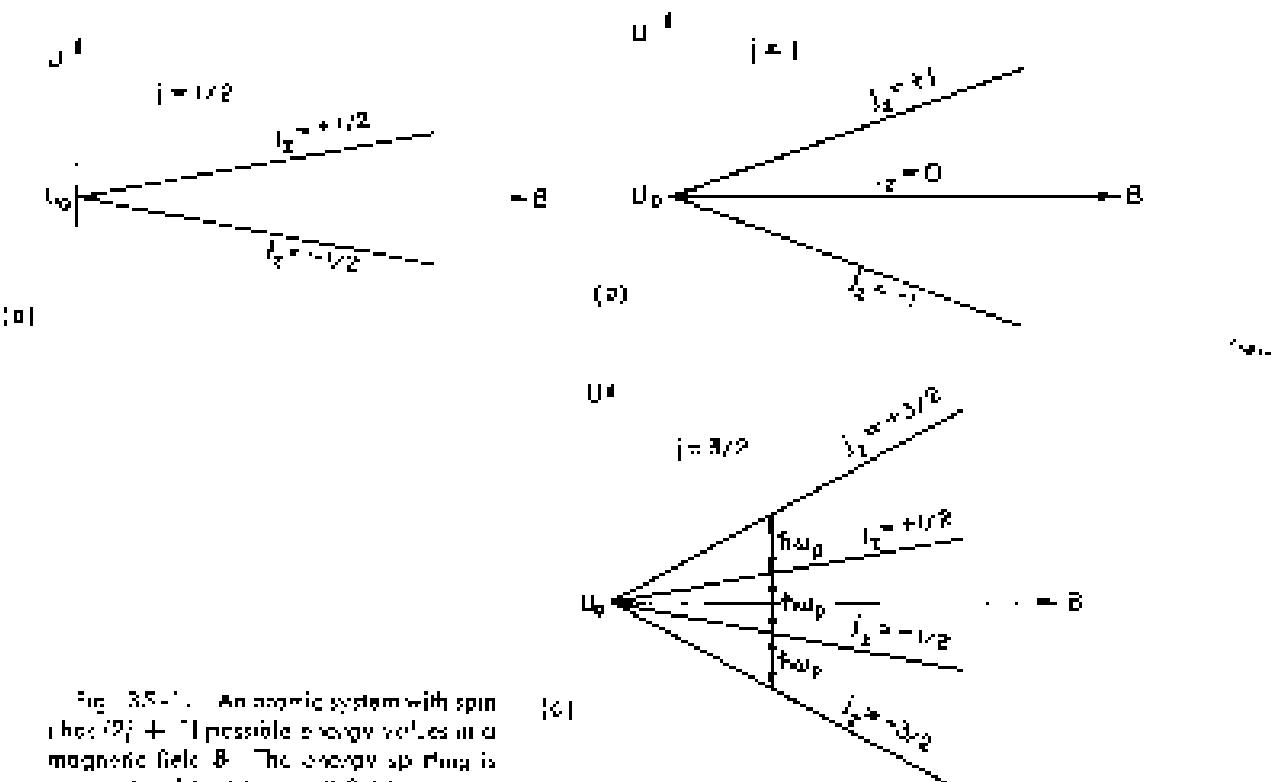


Fig. 35-1. An atomic system with spin $j = 1/2$; 1 ; $3/2$ possible energy values in a magnetic field B . The energy splitting is proportional to B for small fields.

(For some particles, such as zero.) What we mean by "the magnetic moment" in this statement is that the energy of the system in a magnetized field, say in the z -direction, can be written as $-\mu_z B$ for small magnetic fields. We must have the condition that the field should not be too great, otherwise it could destroy the internal motions of the system and the energy would not be a measure of the magnetic moment that was there before the field was turned on. But if the field is sufficiently weak, the field changes the energy by the amount

$$\Delta E = -\mu_z B, \quad (35.1)$$

with the understanding that in this equation μ_z is to be replaced by

$$\mu_z = g \left(\frac{e}{2m} \right) J_z, \quad (35.2)$$

where J_z and v are of the value given in Eq. (35.1).

Suppose we take a system with a spin $j = 1/2$. Without the magnetic field, the system has two different possible states corresponding to the different values of J_z , all of which have exactly the same energy. But the moment we can in the z -direction is not, there is an additional energy of interaction which separates these states into two slightly different energy levels. The energies of these levels are given by a certain energy proportionality factor, called by aches of $1/2$, i.e., $-1/2$, but not $-1/2$, the values of μ_z . The resulting set of energy levels for an atom system with spins of $1/2$, 1 , and $3/2$ are shown in the diagrams of Fig. 35-1. (Remember, that for any magnetic system of electrons the magnetic moment is always directed opposite to the angular momentum.)

It is evident from the diagrams that the "center of gravity" of the energy levels is the same with and without a magnetic field. A noticeable fact the spacings between the levels in a magnet are always equal for a given particle in a given magnetic field. We are going to study the energy spacings, for a given magnetic field B , in Sec. 35-2, which is just a definition of ω_0 . Using Eqs. (35.2) and (35.1), we have

$$\begin{aligned} \Delta E_0 &= B \sum_{j=1}^J J_z B \\ \omega_0 &= g \frac{e}{2m} B. \end{aligned} \quad (35.3)$$

The quantity $\mu g / (2m)$ is just the ratio of the magnetic moment to the angular momentum — it is a property of the particle. Equation (35-4) is the same formula that we got in Chapter 24 for the angular velocity of precessed in a magnetic field, for a gyroscope whose angular momentum is J and whose magnetic moment is μ .

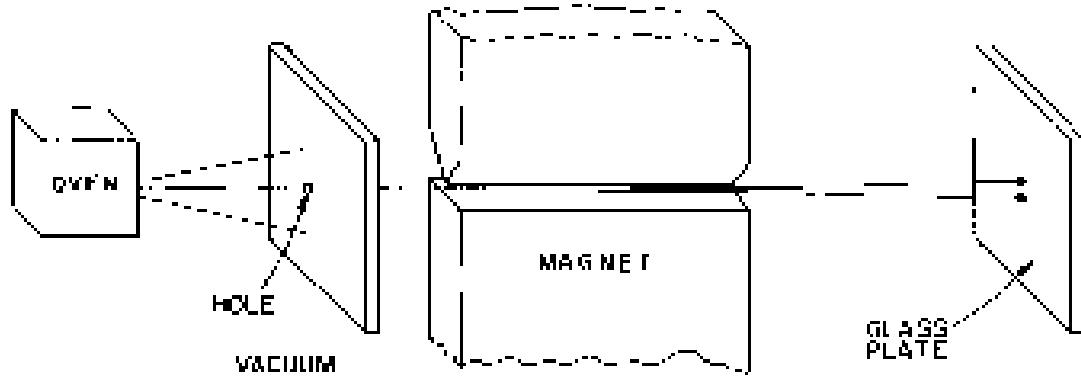


Fig. 35-2. The experiment of Stern and Gerlach.

35-2 The Stern-Gerlach experiment

The fact that the angular momentum is quantized is such a surprising thing that we will talk a little bit about it historically. It was a shock from the moment it was discovered (although it was expected theoretically). It was first observed in an experiment done in 1922 by Stern and Gerlach. If you wish, you can re-enact the experiment of Stern-Gerlach as a direct verification for a below in the supplementary exercises at the end of the chapter. Stern and Gerlach devised an experiment for measuring the magnetic moment of individual silver atoms. They prepared a beam of silver atoms by evaporating silver in a furnace and letting some of them escape out through a series of small holes. This beam was directed between the pole tips of a special magnet, as shown in Fig. 35-2. Their idea was the following. If this beam has a magnetic moment μ , then in a magnetic field B it has an energy $-\mu B$, where θ is the direction of the magnetic field. In the classical theory, θ would be equal to the magnetic moment times the cosine of the angle between the magnet and the magnetic field, so the extra energy in the field would be

$$\Delta E = -\mu B \cos \theta. \quad (35.3)$$

If θ gets out of the plane, then the magnetic moments would point in every possible direction, so there would be all values of θ . Now, if the magnetic field was so very rapidly varied, if there is a strong field going out from the magnet, θ would also vary with position, and there would be a force on the magnetic moments whose direction will depend on whether θ is positive or negative. The atoms will be pulled up or down by a force proportional to the derivative of the magnetic energy, from the $\mu B \cos \theta$ potential work,

$$F_x = -\frac{\partial E}{\partial z} = \mu \sin \theta \frac{\partial B}{\partial z}. \quad (35.4)$$

Stern and Gerlach made their magnet with a very sharp edge on one of the pole tips in order to produce a very rapid variation of the magnetic field. The beam of silver atoms was directed right along this sharp edge, so that the atoms would feel a vertical force in the inhomogeneous field. A silver atom with its magnetic moment directed horizontally would have no force on it and would go straight past the magnet. An atom whose magnetic moment was exactly vertical would have a force pulling it up toward the sharp edge of the magnet. An atom whose magnetic moment was pointed downward would feel a downward push. Thus,

as they left the magnet, the atoms would be spread out according to their vertical components of magnetic moment... In the classical theory all angles are possible so that when the silver atoms were reflected by deposition on a glass plate one should expect a series of silver atoms in a curved line. The height of the line would be proportional to the magnitude of the magnetic moment. The observation of classical theory was completely revealed when Stern and Gerlach saw what actually happened. They found on the glass plate two distinct spots. The silver atoms had formed two beams.

They asked me if you are going to appear in my lecture on molecular beam methods. How does the magnetic moment know there is only allowed to have a certain amount in the direction of the magnetic field? Well, that was only the beginning of the discovery of the quantization of angular momentum, for instead of trying to give you a theoretical explanation, we can just say that you are stuck with the result of this experiment just as the physicists of that day had to accept the result when the experiment was done. It is an experimental fact that the energy of an atom in a magnetic field takes on a series of individual values. For each of these values the energy is proportional to the field strength. So in a region where the field varies, the number of levels will tell us that the possible magnetic fields on the atom will have a series of definite values, the force is different for each state, so the beam of atoms is split into a small number of separate beams. From a measurement of the reflection of the beams, one can find the strength of the magnetic moment.

35-3 The Rabi molecular-beam method

We would now like to describe an improved apparatus for the measurement of magnetic moments which was developed by J. S. Rabi and his collaborators. In the Stern-Gerlach experiment the deflection of atoms is very small and the measurement of the magnetic moment is not very precise. Rabi's technique permits a large-scale improvement in the measurement of the magnetic moment. The method is based on the fact that the typical energy of the atoms in a magnetic field is proportional to the number of cycles per second. And the energy of an atom in a magnetic field can have only certain exactly definite values, it is really no more surprising than it is fact that atoms in general occupy only a few discrete energy levels; something we mentioned often in Volume I. Why should the same thing not hold for atoms in a magnetic field? It does. But it is the attempt to correlate this with the ideas of the second quantum theory that brings out some of the strange implications of quantum mechanics.

When an atom has two levels which differ in energy by the amount ΔE , it can make a transition from the upper level to the lower level by emitting a light quantum of frequency ω , where

$$\hbar\omega = \Delta E \quad (35.7)$$

The same thing can happen with atoms in a magnetic field. Only here, the energy difference is so small that the frequency does not correspond to light, but to microwave or radio frequencies. The atom has to make a lower-energy transition to an upper-energy level of an atom can also take place with such an external light or, in the case of atoms in a magnetic field, by the absorption of microwave energy. Thus if we have an atom in a magnetic field, we can cause transitions if we are able tocouple by applying an additional electromagnetic field of the proper frequency. In other words, if we have an atom in a strong magnetic field and we have the atom with a weak varying electromagnetic field, there will be a certain probability of "switching" it to another level if the frequency is near to the one in Eq. (35.7). For an atom in a magnetic field, this frequency ω_{res} at which we have calculated ω , will be given in terms of the magnetic field by Eq. (35.4). If the atom is coupled with the exciting frequency, the chance of causing a transition is very small. This leads to a sharp resonance at ω_{res} ; the probability of causing a transition. By measuring the frequency of this resonance in a known magnetic field B , we can measure the quantity μ_0/μ_B —and hence the ρ -factor—with great precision.

It is interesting that one comes to the same conclusion from a classical point of view. According to the classical picture, when we place a small gyroscope with a magnetic moment μ and an angular momentum J in an external magnetic field B , the gyroscope will precess about an axis parallel to the magnetic field. (See Fig. 35-2.) Suppose we ask: How can we change the angle of the classical gyroscope with respect to the field—namely, with respect to the z -axis? The magnetic field produces a torque around a horizontal axis. Such a torque you would think is going to line up the magnet with the field, but it only causes the precession. If we want to change the angle of the gyroscope with respect to the z -axis, we must exert a torque east along the x -axis. If we apply a torque which goes in the same direction as the precession, the angle of the z -vector will continue to give a smaller component, $\alpha_p \cdot J$ in the z -direction. In Fig. 35-3, the angle between J and the z -axis would increase. If we try to reduce the precession, J moves toward the z -axis.

For an s -processing atom in a uniform magnetic field, how can we apply the kind of force we want? The answer is: with a weak magnetic field from the side. You might at first think that the direction of this magnetic field would have to coincide with the precession of the magnetic moment, so that it was always at right angles to the moment, as indicated by the field B' in Fig. 35-4(a). Such a field works very well, but an alternating horizontal field is almost as good. If we have a constant horizontal field B' , which is always in the x -direction (this is a tuning fork, which oscillates with the frequency ω_0), then on each one-half cycle the torque on the magnetic moment reverses, so that it has a cumulative effect which is almost as effective as a rotating magnetic field. Classically, then, we would expect the component of the magnetic moment along the x -direction to change if we have a very weak oscillating magnetic field at a frequency which is exactly ω_0 . Classically, of course, μ_x would change continuously, but in quantum mechanics the component of the magnetic moment cannot adjust continuously. It must jump suddenly from one value to another. We have made the comparison between the consequences of classical mechanics and quantum mechanics to give you some idea as to what might happen classically and how it is related to what actually happens in quantum mechanics. You will notice, incidentally, that the required resonance frequency is the same in both cases.

One additional remark: From what we have said about quantum mechanics, there is no apparent reason why there couldn't also be transitions at the frequency $2\omega_0$. It happens that there isn't, any analog of this in the classical case, and also it doesn't happen in the quantum theory either—so far as we know. The particle method of inducing the transitions that we have described. With a nonrotating horizontal magnetic field, the probability that a frequency $2\omega_0$ would cause a jump of two steps at once is zero. It is only at the frequency ω_0 that there are other allowed or allowed, and these in pairs.

Now we are ready to describe Rabi's method for measuring magnetic moments. We will consider here only the apparatus for atoms with a spin of $1/2$. A diagram of the apparatus is shown in Fig. 35-5. This is an oven which gives out a stream of s -process atoms which pass through a hole of three rings. Magnet 1

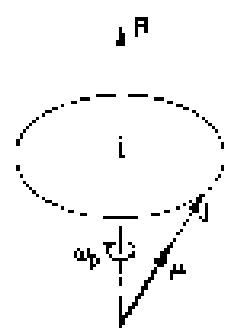


Fig. 35-3. The classical precession of an atom with the magnetic moment μ and the angular momentum J .

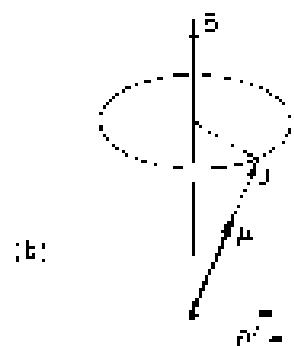
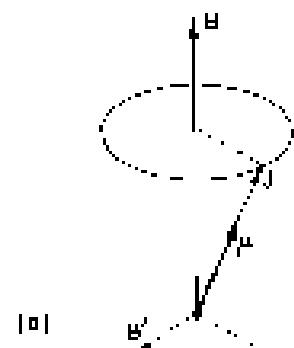


Fig. 35-4. The angular precession of atomic magnets can be changed by a horizontal magnetic field a wave of one angle (a), or by an oscillating (c), as in (b).

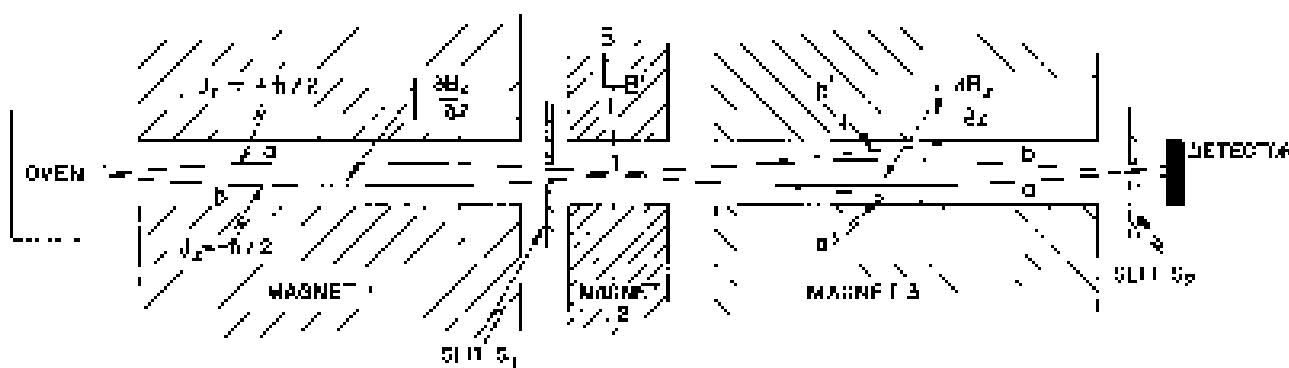


Fig. 35-5. The Rabi molecular-beam apparatus.

is ω , like the one in Fig. 35-2, and the field with a strong field gradient, say, with $\partial B/\partial z$ positive. If the atoms have a magnetic moment, they will be deflected downward if $J_z = -\hbar/2$, or upward if $J_z = \hbar/2$ (since for electrons μ is directly opposite to J). If we consider only those atoms which can get through the slit S_1 , there are two possible trajectories, as shown. Atoms with $J_z = +\hbar/2$ must go along curve a to get through the slit, and those with $J_z = -\hbar/2$ must go along curve b . Atoms which start out from the oven along other paths will not get through the slit.

Magnet 2 has a constant field. There are no forces on the atoms in this region, so they go straight through and enter magnet 3. Magnet 3 is just like magnet 1 but with the field inverted, so that $\partial B/\partial z$ has the opposite sign. The atoms with $J_z = +\hbar/2$ (we say "with spin up"), that felt a downward push in magnet 1, get an upward push in magnet 3; they continue on the path a and go through slit S_2 to a detector. The ones with $J_z = -\hbar/2$ ("with spin down") also have opposite forces in magnet 3 and go along the path b , which also takes them through slit S_2 to the detector.

The reflected ray may be made in various ways, depending on the beam being measured. One way is to let the beam pass through sodium-like sodium, and a detector can be arranged behind the oven to count particles in a stationary current meter. When sodium atoms land on the wire, they undergo ionization at the Na-L shell, leaving an electron behind. There is a current ratio of about 1000 to 1 for the number of sodium atoms arriving per second.

In the gap of magnet 2 there is a set of coils that produces a small oscillating magnetic field B' . The coils are driven with a square wave oscillation at a frequency ω . So between the poles of magnet 2 there is a strong, low $\partial B/\partial z$, vertical field B and a weak, oscillating, horizontal field B' .

Suppose now that the frequency ω of the oscillating field is set at ω_0 , the "precession" frequency of the atoms in the field B . The alternating field ω_0 causes some of the atoms passing by to make transitions from one J_z to the other. An atom whose spin was initially "up" ($J_z = +\hbar/2$) may be flipped "down" ($J_z = -\hbar/2$). Now this atom has the direction of its magnetic moment reversed, so it will feel a downward force in magnet 1 and will move along the path b , shown in Fig. 35-3. It will no longer get through the slit S_2 to the detector. Similarly, some of the atoms whose spins were initially down ($J_z = -\hbar/2$) will have their spins flipped up ($J_z = +\hbar/2$) as they pass through magnet 2. They will then pass through slit S_2 and will not get to the detector.

If the field B' has a frequency appreciably different from ω_0 , it will not cause any spin flips, and all atoms will follow their undisurbed paths to the detector. So you can see that the "precession" frequency ω_0 of the atoms in the field B can be found by varying the frequency ω of the field B' until a decrease in ω increases the current of atoms hitting at the detector. A decrease in ω is called "resonance" when ω_0 is "in resonance" with ω . A plot of the detector current as a function of ω might look like the one shown in Fig. 35-6. Knowing ω_0 , we can obtain the ω_0 value of the atom.

Such microwave-heterodyne, or they are usually called "molecular" beam resonance experiments are a beautiful and delicate way of measuring the magnetic properties of atomic objects. The resonance frequency ω_0 can be determined with great precision—in fact, with a greater precision than we can measure the magnetic field B_0 , which we must know to find ω_0 .

35-4 The paramagnetism of bulk materials

We would like now to describe the phenomenon of the paramagnetism of bulk materials. Suppose we have a substance whose atoms have permanent magnetic moments, for example a crystal like copper sulfide. In the crystal there are copper ions whose inner electron shells are a net angular momentum and a net magnetic moment. So the copper ion is an object who has a permanent magnetic moment. Let's say just a word about which atoms have magnetic moments and which ones don't. Any atom, like sodium for instance, which has an odd number

of electrons will leave a magnetic moment. Sodium has one electron in its un-filled shell. This electron gives the atom a spin and a magnetic moment. Ordinarily, however, when compounds are formed the extra electrons in the valence shell are paired together with other electrons whose spin directions are exactly opposite, so that all the angular momenta and magnetic moments of the valence electrons cancel each other. That's why, in general, molecules do not have a magnetic moment. Of course if you have a gas of separate atoms, there is no such cancellation.⁵ Also, if you have what is called in chemistry a "free radical," an object with an odd number of valence electrons—then the bonds are not completely broken, and there is a net angular momentum.

In most bulk materials there is a net magnetic moment only if there are atoms present whose outer electron shell is not filled. Then there can be a net angular momentum and a magnetic moment. Such atoms are found in the "transition elements" part of the periodic table—Fe, manganese, chromium, cobalt, iron, nickel, cobalt, palladium, and platinum are elements of this kind. Also, all of the rare earth elements have unfilled outer shells and permanent magnetic moments. There are a couple of other strange things that also happen to have magnetic moments, such as liquid oxygen, but we will leave it to the chemistry department to explain the reason.

Now suppose that we have a box full of atoms or molecules with permanent moments, say a gas, or a liquid, or a crystal. We would like to know what happens if we apply an external magnetic field. With no applied field, the moments are kicked around by the thermal motion, and the moments wind up pointing in all directions. But when there is a magnetic field, it tries to line up the little magnets; then there are more moments lying toward the field than away from it. The material is "magnetized."

We define the magnetization M of a material as the net magnetic moment per unit volume, by which we mean the vector sum of all the atomic magnetic moments in a unit volume. If there are N atoms per unit volume and their average moment is $\langle \mathbf{p} \rangle_0$, then M can be written as N times the average atomic moment:

$$M = N\langle \mathbf{p} \rangle_0. \quad (35.8)$$

The definition of M corresponds to the definition of the electric polarization P of Chapter 19.

The classical theory of paramagnetism is just like the theory of the dielectric constant we derived you in Chapter 11. One assumes that each of the atoms has a magnetic moment μ , which always has the same magnitude but which can point in any direction. In a field B , the magnetic torque is $-\mu \cdot B = -\mu B \cos \alpha$, where α is the angle between the moment and the field. From statistical mechanics, the relative probability of having any angle is $e^{-\beta \mu B \cos \alpha}$, so angles near zero are more likely than angles near π . Proceeding exactly as we did in Section 11.3, we find that for small magnetic fields M is directly proportional to B and has the magnitude

$$M = \frac{\chi \mu^2 B}{3kT}. \quad (35.9)$$

[See Eq. (11.30).] This approximate formula is correct only for $\mu B/kT$ much less than one.

We find that the induced magnetization—the magnetic moment per unit volume—is proportional to the magnetic field. This is the phenomenon of paramagnetism. You will see that the effect is stronger at low temperatures and weaker at higher temperatures. When we put a field on a substance, it develops, for small fields, a magnetic moment proportional to the field. The ratio of M to B (for small fields) is called the magnetic susceptibility.

Now we want to look at paramagnetism from the point of view of quantum mechanics. We take first the case of an atom with a spin of $1/2$. In the absence of

⁵ Obviously Na vapor is mostly monatomic, although there are also some molecules of Na_2 .

In a magnetic field the atoms have a certain energy. But in a magnetic field there are two possible energies, one for each value of J_z . For $J_z = +1/2$, the energy is changed by the magnetic field by the amount

$$\Delta E_+ = -g \left(\frac{qA}{2m} \right) \cdot \frac{1}{2} \cdot B \quad (35.10)$$

(The energy shift is positive for an atom because the electron charge is negative.) For $J_z = -1/2$, the energy is changed by the amount

$$\Delta E_- = -g \left(\frac{qA}{2m} \right) \cdot \frac{1}{2} \cdot B \quad (35.11)$$

To save writing, let's set

$$\mu_B = g \left(\frac{qA}{2m} \right) \cdot \frac{1}{2} \cdot B \quad (35.12)$$

Then

$$\Delta E = \pm \mu_B B. \quad (35.13)$$

The meaning of μ_B is clear: μ_B is the component of the magnetic moment in the upper state, and $-\mu_B$ is the component of the magnetic moment in the down-spin case.

New statistical mechanics tells us that the probability that an atom is in one state or another is proportional to

$$e^{-\Delta E / kT} \propto e^{-\Delta E / kT},$$

With no magnetic field the two states have the same energy; so when there is no field, the probabilities are proportional to

$$e^{-\Delta E / kT}. \quad (35.14)$$

The number of atoms per unit volume with spin up is

$$N_{up} = n e^{-\mu_B B / kT}, \quad (35.15)$$

and the number with spin down is

$$N_{down} = n e^{\mu_B B / kT}. \quad (35.16)$$

The constant n is to be determined so that

$$N_{up} + N_{down} = N. \quad (35.17)$$

The total number of atoms per unit volume. So we get that

$$n = \frac{N}{e^{-\mu_B B / kT} + e^{\mu_B B / kT}}. \quad (35.18)$$

What we are interested in is the average magnetic moment along the axis. The atoms with spin up will contribute a moment of $+\mu_B$, and those with spin down will have a moment of $-\mu_B$, so the average moment is

$$\text{moment} = \frac{N_{up}(+\mu_B) + N_{down}(-\mu_B)}{N} \quad (35.19)$$

The magnetic moment per unit volume M is then $M(\mu_B) ...$ Using Eqs. (35.15), (35.16), and (35.17), we get that

$$M = N \mu_B \frac{e^{1.38 \times 10^{-23} T}}{e^{-\mu_B B / kT} + e^{\mu_B B / kT}}. \quad (35.20)$$

This is the quantum-mechanical formula for M for atoms with $m_J = 1/2$. Incidentally, this formula is valid for most elements more complex in terms of the

$$M = N \mu_B \tanh \frac{\omega_0 B}{kT}. \quad (35.21)$$

A plot of M as a function of B is given in Fig. 35-7. When B goes very large, the hyperbolic tangent approaches 1, and M approaches the limiting value $N\mu_B$; $S_z = \pm \frac{1}{2}$ has the magnetization averages. We can see why that is: at high enough fields the moments are all lined up in the same direction. In other words, they are all in the spin-down state, and each atom contributes the moment μ .

In most normal cases—say, for typical materials, room temperatures, and 1 eV fields one can accurately get (like 10,000 esus)—the ratio $\omega_0 B / kT$ is about 0.02. One must go to very low temperatures to see the saturation. For normal temperatures we can usually replace $tanh x$ by x , and write

$$M = \frac{N\mu_B^2 B}{kT}. \quad (35.22)$$

Just as we saw in the classical theory, M is proportional to B . In fact, the formula is almost exactly the same, except that there seems to be a factor of 1/2 missing. But we still need to relate $\omega_0 B$ to our quantum formula to the ω that appears in the classical result, Eq. (35.9).

In the classical formula, what appears is $\omega^2 = \mu \cdot \mu$, the square of the vector magnetic moment... or

$$\mu \cdot \mu = \left(\epsilon \frac{g_e}{2m} \right)^2 J \cdot J. \quad (35.23)$$

We pointed out in the last chapter that you can very likely get the right answer from a classical calculation by replacing $J \cdot J$ by $j(j+1)$. In our particular example, we have $j = 1/2$, so

$$(j(j+1))^2 = \frac{1}{4}.$$

Substituting this for $J \cdot J$ in Eq. (35.23), we get

$$\mu \cdot \mu = \left(\epsilon \frac{g_e}{2m} \right)^2 \frac{1}{4},$$

or in terms of μ_0 , defined in Eq. (35.12), we get

$$\mu \cdot \mu = \frac{3}{4} \mu_0^2.$$

Substituting this for ω^2 in the classical formula, Eq. (35.9), does indeed reproduce the correct quantum formula, Eq. (35.22).

The quantum theory of paramagnetism is easily extended to atoms of spin $\frac{1}{2}$. The 'new-field' magnetization is

$$M = N g_e \mu_0 B + \frac{1}{2} \mu_B B, \quad (35.24)$$

where

$$\mu_B = \frac{e \hbar}{2m} \quad (35.25)$$

is a combination of constants with the dimensions of a magnetic moment. Most atoms have moments of roughly this size. It is called the Bohr magneton. The spin magnetic moment of the electron is almost exactly one Bohr magneton.

35-5 Cooling by adiabatic demagnetization

There is a very interesting second application of paramagnetism. At very low temperatures it is possible to line up the atomic magnets in a strong field. It is then possible to get down to extremely low temperatures by a process called adiabatic demagnetization. We can take a para magnetic salt (for example, one

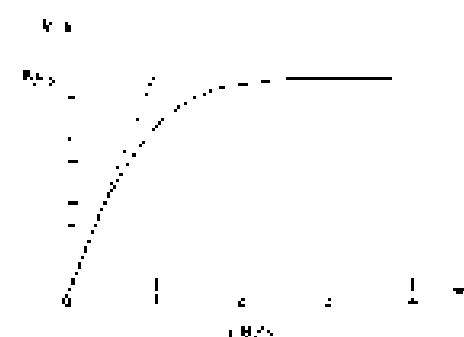


Fig. 35-7. The variation of the paramagnetic magnetization with the magnetic field strength B .

containing a number of rare-earth atoms (like praseodymium or gadolinium), and start by cooling it down with liquid helium to just 4.2 degrees Kelvin in a strong magnetic field. Then the factor $\mu B/kT$ is larger than 1—say more like 10/3. Most of the spins are lined up, and the magnetization is nearly saturated. Let's say, to make it easy, that the field is very powerful and the temperature is very low, so that nearly all the atoms are lined up. Then you rotate the salt. You only pay, by removing the liquid helium and leaving a good vacuum, and turn off the magnetic field. The temperature of the salt goes way down.

Now if you were to turn off the field suddenly, the jiggling and shaking of the atoms in the crystal would gradually knock all the spins out of alignment. Some of them would be up and some down. But if there is no field (and disregarding the interactions between the atomic magnets, which will make only a slight error), it takes no energy to turn over the atomic magnets. They could randomize their spins without any energy change and, therefore, without any temperature change.

Suppose, however, that while the atomic magnets are being flipped over by the thermal motion there is still some magnetic field present. Then it requires some work to flip them over opposite to the field—over *costs* work against the field. This takes energy from the thermal motions and lowers the temperature. So if the strong magnetic field is not removed too rapidly, the temperature of the salt will decrease—it is cooled by the demagnetization. From the quantum-mechanical view, when the field is strong all the atoms are in the lower state, because the odds against any being in the upper state are impossibly big. But as the field is lowered, it gets more and more likely that thermal fluctuations will knock an atom into the upper state. When that happens, the atom absorbs the energy $\Delta E = \mu B\delta$. So if the field is turned off slowly, the magnetic transitions can take energy out of the thermal vibrations of the crystal, cooling it off. It is possible in this way to go from a temperature of a few degrees absolute down to a temperature of a few thousandths of a degree.

Would you like to make something even colder? Let that be James and I. We have already mentioned that there are also paramagnetic moments for the atomic nuclei. One reason for paramagnetism weak, just as well for nuclei, except that the moments of nuclei are correspondingly *tiny* smaller. [They are of the order of magnitudes of μ_B/m_N , where m_N is the proton mass, so they are smaller by the ratio of the masses of the electron and proton.] With such magnetic moments, even at a temperature of 2°K, the factor $\mu B/kT$ is only a few parts in a thousand. But if we use the paramagnetic demagnetization process to get down to a temperature of a few thousandths of a degree, $\mu B/kT$ becomes a number near 1—or these low temperatures we can begin to saturate the nuclear moments. That is good luck, because we can then use the adiabatic demagnetization of the nuclear magnetism to reach still lower temperatures. Thus it is possible to do two stages of magnetic cooling. First we use adiabatic demagnetization of paramagnetic ions to reach a few thousandths of a degree. Then we use the cold paramagnetic salt to cool some material which has a strong nuclear magnetism. Finally, when we remove the magnetic field from this material, its temperature will go down to within a millionth of a degree of absolute zero—if we have done everything very carefully.

35-6 Nuclear magnetic resonance

We have said that atomic paramagnetism is very small and that nuclear magnetism is even a thousand times smaller. Yet it is relatively easy to observe the nuclear magnetism by the phenomenon of "nuclear magnetic resonance." Suppose we take a substance like water, in which all of the electron spins are exactly balanced so that their net magnetic moment is zero. The molecules will still have a very, very tiny magnetic moment due to the nuclear magnetism, of the hydrogen nuclei. Suppose we put a single sample of water in a magnetic field B . Since the protons (of the hydrogen) have a spin of 1/2, they will have two possible energy states. If the water is in thermal equilibrium, there will be slightly more

protons in the lower energy states, with their moments directed parallel to the field. There is a small net magnetic moment per unit volume. Since the proton moment is only about one-thousandth of an atomic moment, the magnetization which goes as μ^2 —using Eq. (25.22)—is only about one-millionth as strong as liquid atomic paramagnetism. (That's why we have to pick a material with the same magnetism.) If you work it out, the difference between the number of protons with spin up and with spin down is only one part in 20%, so the effect is indeed very small! It can still be observed, however, in the following way.

Suppose we surround the water sample with a small coil that produces a small, horizontal oscillating magnetic field. If this field oscillates at the frequency ω_0 , i.e., will induce transitions between the two energy states just as we described for the Rabi experiment in Section 35-3. When a proton flips from an upper energy state to a lower one, it will give up the energy $\mu_0 B$ which, as we have seen, is equal to E_{sp} . That flip from the lower energy state to the upper one, it will absorb the energy $\mu_0 B$, leave the coil. Since there are slightly more protons in the lower state than in the upper one, there will be a net absorption of energy from the coil. Although the effect is very small, the signal energy absorbed can be even well measured by the right kind of filter.

Just as in the Rabi molecular-beam experiment, the energy absorption will be zero only when the oscillating field is in resonance, that is, when

$$\omega = \omega_0 = 2 \left(\frac{\gamma}{g \mu_0} \right) B$$

It is often more convenient to search for the resonance by varying B while keeping ω_0 fixed. The energy absorption will evidently appear when

$$B = \frac{g \mu_0 \omega}{0.6}$$

A typical nuclear magnetic resonance apparatus is shown in Fig. 35-8. A high-frequency oscillator drives a small coil placed just above the mass of a large electromagnet. Two small iron bars (yokes) around the pole tips are driven with a 60-cycle current so that the magnet has a "resonator" winding which carries about as a very short loop circuit. As a result, say that the main current of the magnet is set to give a field of 5000 gauss, and the auxiliary coils make a current of +1 gauss about this value. If the oscillator is set at 21.2 megacycles per second, it will tune to the proton resonance each time the field sweeps through 5000 gauss [using Eq. (34.13) with $g = 2.0$ for the proton].

The circuit of the oscillator is arranged to give an additional output signal proportional to any change in the power being absorbed from the oscillator. This signal is fed to the vertical deflection amplifier of an oscilloscope. The horizontal sweep of the oscilloscope is triggered once during each cycle of the field sweeping frequency. (More usually, the horizontal deflection is made to follow in proportion to the sweeping field.)

Before the water sample is placed inside the high-frequency coil, the power drawn from the oscillator is some value. (It doesn't change with the magnetic field.) When a small bottle of water is placed in the coil, however, a signal appears on the oscilloscope, as shown in the figure. We see a picture of the power being absorbed by the flipping over of the protons.

In practice, it is difficult to know how to set the main magnet to exactly 5000 gauss. What one does is to adjust the main magnet current until the resonance signal appears on the oscilloscope. It turns out that this is now the most convenient way to make an accurate measurement of the strength of a magnetic field. Of course, at some time someone had to measure accurately the magnetic field and therefore to determine the g -value of the proton. But now that this has been done, a coil-in-resonance apparatus has that of the figure can be used as a "precision magnetometer."

We should say a word about the shape of the signal. If we were to sweep the magnetic field very slowly, we would expect to see a normal resonance curve. The energy absorption would reach a maximum when ω_0 arrived exactly at the

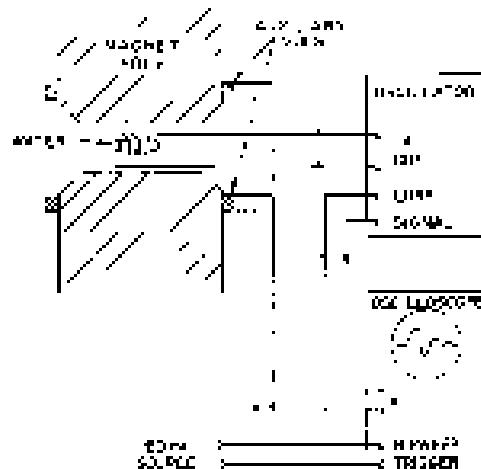


Fig. 35-8. A nuclear magnetic resonance apparatus.

oscillator frequency. There would be some absorption at nearby frequencies because all the protons are not in exactly the same field—and different fields mean slightly different resonant frequencies.

One might wonder, incidentally, whether at the resonance frequency we should see any signal at all. Shouldn't we expect the high-frequency field to equilibrate the populations of the two states—so that there should be no signal except, when the water is first put in?² No, exactly, because although we are trying to equilibrate the two populations, the thermal motions of their part are trying to keep the proton ratios the same temperature T . If we sit at the resonance, the power being absorbed by the nuclei is just what is being lost to the thermal motions. This is, however, relatively little “torsional motion” between the motion magnetic moments and proton positions. The protons are relatively localized down in the center of the electron distributions. So, in pure water, the resonance signal is, in fact, usually too small to be seen. To increase the absorption, it is necessary to increase the “precessing magnet.” This is usually done by adding salt to the water. The ions provide like small magnets; as they precess around in their local fields, they make tiny precessing magnetic fields of their own. These varying fields “couple” the proton magnets to the salt so strongly and long that equilibrium does not establish itself. It is through this “coupling,” that protons in the higher energy states can lose their charge so that they are again capable of absorbing energy from the oscillator.

To practice the共振 signal of a nuclear resonance apparatus does not look like a normal resonance curve. It is usually a more complicated signal with oscillations like the one drawn in the figure. Such signal shapes appear because of the changing fields. The explanation should be given in terms of quantum mechanics, but it can be shown that in such experiments the classical ideas of precessing momenta always give the correct answer. Classically, we would say that when we arrive at resonance we start driving a lot of the precessing nuclear magnets synchronously. In so doing, we make them precess together. These nuclear magnets, all rotating together, will set up an induced emf in the oscillator coil at the frequency ω_0 . But because the magnetic field is increasing with time, the precession frequency is increasing also, and the induced voltage is soon at a frequency a little higher than the oscillator frequency. As the induced emf goes alternately in phase and out of phase with the oscillator, the “absorbed” power goes alternately positive and negative. So on the oscilloscope we see the beat note between the precess frequency and the oscillator frequency. Because the proton frequencies are not all identical (different protons are in slightly different fields) and also possibly because of the disturbance of ions that come into the water, the steady precession increases from jet out of phase, and the beat signal disappears.

These phenomena of nuclear resonance have been put to use in many ways as does the living cell does. Little about matter—especially in electricity and nuclear physics. It goes without saying that the numerical values of the magnetic moments of nuclei tell us something about their structure. In particular, much has been learned from the studies of the shifts of the resonances. Because of the static fields produced by nearby nuclei, the exact position of a nuclear resonance is shifted somewhat, depending on the environment in which any particular nucleus finds itself. Measuring these shifts helps determine which atoms are near which other ones and helps to elucidate the details of the structure of molecules. Equally important is the electron spin resonance of free radicals. Although not present to any very large extent in equilibrium, such radicals are often intermediate states of chemical reactions. A measurement of an electron spin resonance is a delicate test for the presence of free radicals and is often the key to understanding the mechanism of certain chemical reactions.

Paramagnetism

36-1 Magnetization currents

In this chapter we will discuss some materials in which the net magnetization, \mathbf{M} , is much greater than in the case of paramagnetism or diamagnetism. The phenomenon is called *ferromagnetism*. In paramagnetic and diamagnetic materials the induced magnetic moments are usually so weak that we don't have to worry about the additional fields produced by the magnetic moments. For ferromagnetic materials, however, the magnetic moments induced by applied magnetic fields are quite enormous and have a great effect on the fields themselves. In fact, the induced moments are so strong that they are often the dominant effect on producing the observed fields. So one of the things we will have to worry about is the mathematical theory of large induced magnetic moments. That is, of course, just a technical question. The real problem is, why are the magnetic moments so strong—how does it all work? We will come to that question in a little while.

Induce the magnetic fields of ferromagnetic materials is something like the problem of finding the electrostatic field in the presence of dipoles. You will remember that we first described the internal properties of a dielectric in terms of a vector field, \mathbf{P} , the dipole moment per unit volume. Then we figured out that the effects of this polarization are equivalent to a charge density ρ_{pol} given by the expression of P :

$$\rho_{pol} = -\nabla \cdot \mathbf{P}. \quad (36.1)$$

The total charge in any situation can be written as the sum of this polarization charge plus all other charges, whose density we write⁴ ρ_{total} . Then the Maxwell equation which relates the divergence of E to the charge density becomes

$$\nabla \cdot E = \frac{\rho}{\epsilon_0} = \frac{\rho_{total} + \rho_{pol}}{\epsilon_0},$$

or

$$\nabla \cdot E = -\frac{\nabla \cdot P}{\epsilon_0} + \frac{\rho_{total}}{\epsilon_0}.$$

We can then pull out the polarization part of the charge and put it on the other side of the equation, to get the new law

$$\nabla \cdot (\epsilon_0 E + \mathbf{P}) = \rho_{total}. \quad (36.2)$$

The new law says the divergence of the quantity $(\epsilon_0 E + \mathbf{P})$ is equal to the density of the other charges.

Moving E and P together as in Eq. (36.2), of course, is useful only if we know some relation between them. We have seen that the theory which relates the induced electric dipole moment to the field was a relatively complicated business and can easily only be applied to certain simple situations, and even then as an approximation. We would like to use in your of one of the approximations we used. To find the induced dipole moment of an atom inside a dielectric, it is necessary to know the electric field that acts on an individual atom. We made the approximation—which is true for bar in many cases—that the field on the atom

36-2 Magnetization currents

36-2 The field H

36-3 The magnetization curve

36-4 Iron-core inductances

36-5 Electromagnets

36-6 Spontaneous magnetization

Review: Chapter 10, Dielectrics
Chapter 11, The law of the
surfaces

⁴ If all of the "other" charges were on conductors, ρ_{total} would be the same as our ρ of Chapter 10.

so the source of a would be at the center of the small hole which would be left if we took out the atom (keeping the dipole moments of all the neighboring atoms the same). You will also remember that the electric field in a hole in a polarized crystal depends on the shape of the hole. We summarize our earlier results in Fig. 36.1. For a thin, disc-shaped hole perpendicular to the polarization, the electric field in the hole is given by

$$E_{\text{hole}} = E_{\text{exterior}} + \frac{P}{\epsilon_0},$$

which we showed by using Gauss' law. On the other hand, in a needle-shaped slot parallel to the polarization, we showed—by using the fact that the end of \mathbf{A} is zero—that the electric fields inside and outside of the slot are the same. Finally, we found that for a spherical hole the electric field was one-third of the way between the field of the slot and the field of the cavity.

$$E_{\text{hole}} = E_{\text{exterior}} + \frac{1}{3} \frac{P}{\epsilon_0} \text{ (spherical hole)} \quad (36.3)$$

This was the field we used in thinking about what happens in an atom inside a polarized dielectric.

Now we have to discuss the analog of all this for the case of magnetism. The simplest, easiest way of doing this is to say the M , the magnetic moment per unit volume, is just like P , the electric dipole moment per unit volume, and that, therefore, the negative of the divergence of \mathbf{M} is equivalent to a "magnetic charge density" ρ_m , whatever that may mean. The trouble is, of course, that there is no such thing as a "magnetic charge" in the physical world. As we know, the divergence of \mathbf{B} is always zero. Far that does not stop us from making an artificial entity and writing

$$\nabla \cdot \mathbf{M} = -\rho_m \quad (36.4)$$

where it is to be understood that ρ_m is purely mathematical. Then we could make a complete analogy with the electrostatic case and use all our old equations from electrodynamics. People have often done something like that. In fact, historically, people once believed that the analogy was right. They believed that the quantity ρ_m represented the density of "magnetic poles." These days, however, we know that the magnetization of materials comes from circulating currents in the atoms—either from the spinning electrons or from the motion of the electrons in the atom. J is therefore not even a physical point of view to describe things realistically in terms of the atomic currents rather than in terms of a density of some artificial "magnetic poles." Historically, these currents are sometimes called "Amperian" currents, because Ampere first suggested that the magnetism of matter came from circulating atomic currents.

The actual microscopic current density in an insulating material is, of course, very complicated. Its value depends on where you look in the atom—it's large at some places and small in others; it goes one way in one part of the atom and the opposite way in another part (just as the microscopic electric field varies enormously inside a dielectric). In many practical problems, however, we are interested only in the fields outside of the material or in the average magnetic field inside of the material. When we measure energy taken over time, many atoms. It is only for such macroscopic problems that it is convenient to drop the magnetic state of the matter in terms of M , the average dipole moment per unit volume. What we want to show now is that the atomic currents of magnetized matter can give rise to certain large-scale currents which are related to M .

What we are going to do, then, is to separate the current density j —which is the net source of the magnetic field—into various parts: one part to describe the individual currents of the atomic magnets, and the other parts to describe what other currents there may be. It is extremely convenient to separate the currents into three parts. In Chapter 32 we made a distinction between two concepts which few freely on conductors and the ones which lead one to the back side for the others

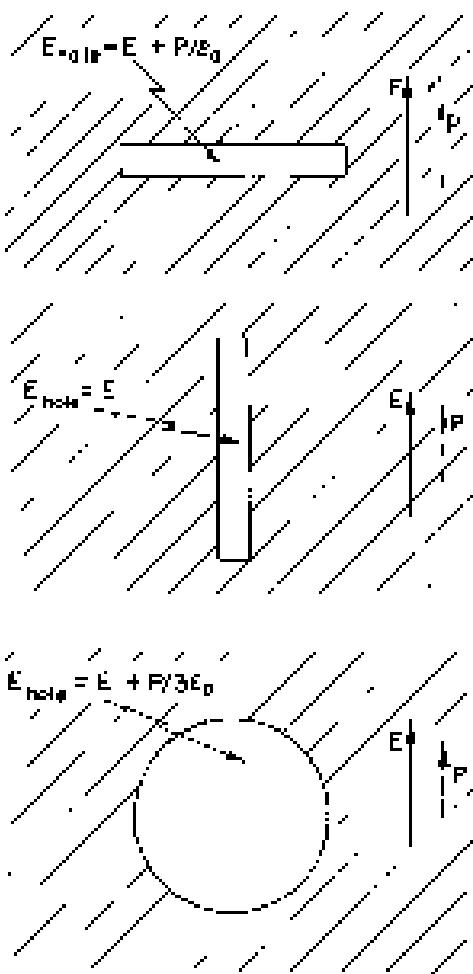


Fig. 36.1. The electric field in a cavity in a dielectric depends on the shape of the cavity.

of the bound charges in dielectrics. In Section 32-2 we wrote

$$\mathbf{j} = \mathbf{j}_{\text{ext}} - \mathbf{j}_{\text{bound}}$$

where \mathbf{j}_{ext} represented the currents from the motion of the bound charges in the electric field and $\mathbf{j}_{\text{bound}}$ took care of all other currents. Now we want to go further. We want to separate $\mathbf{j}_{\text{bound}}$ into one part, \mathbf{j}_{mag} , which describes the average currents inside of magnetized materials, and an additional term which we can call \mathbf{j}_{rest} for whatever is left over. The last term will generally refer to currents in conductors, but it may also include other currents. For example, the currents from charges moving freely through empty space. So we will write for the total current density:

$$\mathbf{j} = \mathbf{j}_{\text{ext}} - \mathbf{j}_{\text{mag}} - \mathbf{j}_{\text{rest}}. \quad (36.5)$$

Of course it is this total current which belongs in the Maxwell equation for the curl of \mathbf{B} :

$$\nabla^2 \mathbf{A} \times \mathbf{B} = \frac{\mathbf{j}}{\sigma} + \frac{\partial \mathbf{E}}{\partial t}. \quad (36.6)$$

Now we have to relate the current \mathbf{j}_{mag} to the magnetization vector \mathbf{M} . So that you can see where we are going, we will tell you that the result is going to be that

$$\mathbf{j}_{\text{mag}} = \nabla \times \mathbf{M}. \quad (36.7)$$

If we are given the magnetization vector \mathbf{M} everywhere in a magnetic material, the circulation current density is given by the curl of \mathbf{M} . Let's see if we can understand why this is so.

First, let's take the case of a cylindrical rod which has a uniform magnetization parallel to its axis. Physically, we know that such a uniform magnetization really means a uniform density of atomic circulating currents everywhere inside the material. Suppose we try to imagine what the actual currents would look like in a cross section of the material. We would expect to see currents something like those shown in Fig. 36-2. Each atomic current goes around and around in a little circle, with all the circulating currents going around in the same direction. Now what is the effective current of such a thing? Well, in most of the bar there is no effect at all, because right next to each current there is another current going in the opposite direction. If we imagine a small surface—but one still quite a bit larger than a single atom—such as is indicated in Fig. 36-2 by the line \overline{AB} , the net current through such a surface is zero. There is no net current anywhere inside the material. Now, however, just at the surface of the ends of the bar there are atomic currents which are not canceled by neighboring currents going the other way. At the surfaces there is a net current always going in the same direction around the rod. Now you see why we say earlier that a uniformly magnetized rod requires a net current along its entire length.

How does this view fit with Eq. (36.7)? First, recall the material the magnetization \mathbf{M} is constant, so all derivatives are zero. It is represented in the geometric picture. At the surfaces, however, \mathbf{M} is not really constant—it is constant up to the edge and then suddenly collapses to zero. So, right at the surface there are terrific gradients which, according to (36.7), will give a big current density. Suppose we look at some barbers near the point C in Fig. 36-2. Taking the x - and y -directions as in the figure, the magnetization \mathbf{M} is in the z -direction. Writing out the components of Eq. (36.7), we have

$$\begin{aligned} \frac{\partial M_x}{\partial y} &= (j_{\text{mag}})_x \\ -\frac{\partial M_z}{\partial y} &= (j_{\text{mag}})_y. \end{aligned} \quad (36.8)$$

At the point C, the derivative $\partial M_x / \partial y$ is zero, but $\partial M_z / \partial y$ is large and positive. Equation (36.8) says that there is a large current density in the x -direction. This agrees with our picture of a surface current going around the bar.

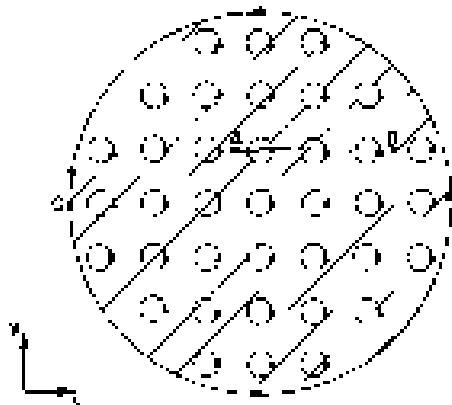


Fig. 36-2. Schematic diagram of the circulating atomic currents as seen in a cross section of an iron rod magnetized in the z -direction.

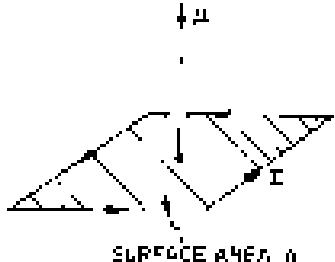


Fig. 36-3. The dipole moment μ of a current loop is IA .

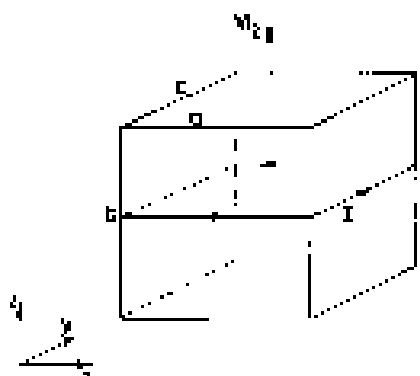


Fig. 36-4. A small magnetized block is equivalent to a circulating surface current.

Fig. 36-5. If the magnetization of two neighbouring blocks is not the same, there is a net surface current between them.

Now we want to find the current density for a more complicated case in which the magnetization varies from point to point in a material. It is easy to see qualitatively that if the magnetization is different in two neighbouring regions, there will not be a perfect cancellation of the circulating currents so that there will be a net current in the volume of the material. It is this effect that we want to work out quantitatively.

First, we need to recall the results of Section 14-5 that a circulating current I has a magnetic moment given by

$$\mu = IA, \quad (36-9)$$

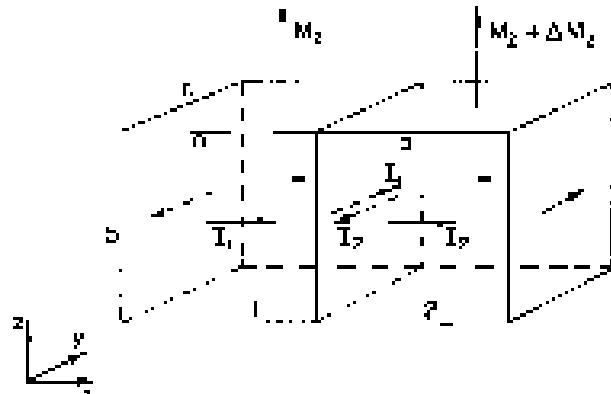
where A is the area of the current loop (see Eq. 36-5). Now let's consider a small rectangular block made of the magnetized material as sketched in Fig. 36-4. We take the block as small that we can consider that the magnetization is uniform inside it. If this block has a magnetization M_2 , in the z-direction, the net effect will be the same as a uniform vertical going around the vertical faces, as shown. We can find the net current off these surfaces from Eq. (36-9). The total magnetic moment of the block is equal to the magnetization times the volume:

$$\mu = M_2 b d b,$$

from which we get (remembering that the area of the loop is bd)

$$I = M_2 b.$$

In other words, the current per unit length (vertically) on each of the vertical surfaces is equal to M_2 .



Now suppose that we insert two such little slabs of size $a \times b \times d$, as shown in Fig. 36-5. Because block 2 is slightly displaced from block 1, i.e., will have a slightly different vertical component of magnetization, which we call $\Delta M_2 = \partial M_2 / \partial x$. Now on the surface between the two blocks there will be two contributions to the total current. Block 1 will produce a current I_1 flowing in the positive y -direction, and block 2 will produce a current I_2 flowing in the negative y -direction. The total surface current on the positive y -direction is the sum:

$$I = I_1 - I_2 = M_1 b - (M_2 + \Delta M_2)b \\ = -\Delta M_2 b.$$

We can write ΔM_2 as the derivative of M_2 in the x -direction to get the displacement from block 1 to block 2, which is, say, σ ,

$$\Delta M_2 = \frac{\partial M_2}{\partial x} \sigma.$$

The current flowing between the two blocks is then

$$I = -\frac{\partial M_2}{\partial x} b \sigma.$$

To relate the current J to our average volume current density \bar{J} , we must realize that the current J is really spread over a certain cross-sectional area. If we imagine the whole volume of the material to be filled with such little blocks, one such side face (perpendicular to the x -axis) can be associated with each block.⁷ Then we see that the area to be associated with the current J is just the area a_2 of one of the front faces. We get the result:

$$\bar{J}_x = \frac{J}{a_2} = -\frac{\partial M_x}{\partial x}.$$

We have at least the beginning of the curl of M .

There should be another term in \bar{J}_y from the variation of the x -component of the magnetization M_x . This contribution to \bar{J}_y will come from the surface between two little blocks stacked one on top of the other, as shown in Fig. 36-6. Using the same arguments we have just made, you can show that this surface will contribute to \bar{J}_y the amount $\partial M_y/\partial x$. These are the only surfaces which contribute to the y -component of the current so we have that the total current density in the y -direction is

$$\bar{J}_y = \frac{\partial M_x}{\partial z} = \frac{\partial M_z}{\partial x}.$$

Working out the currents on the remaining faces of a cube, or using the fact that our z -direction is completely arbitrary, we can conclude that the vector current density is indeed given by the equation

$$\bar{j} = \nabla \times M.$$

So if we choose to describe the magnetic situation in matter in terms of the average magnetic moment per unit volume M , we find that the circulating electric currents are equivalent to an average current density in matter given by Eq. (36.10). If the material is also a dielectric, there may be, in addition, a polarization current $j_{pol} = \partial P/\partial t$. And if the material is also a conductor, we may have a conductive current j_{cond} as well. We can write the total current as

$$\bar{j} = j_{pol} + \nabla \times M + \frac{\partial P}{\partial t}. \quad (36.10)$$

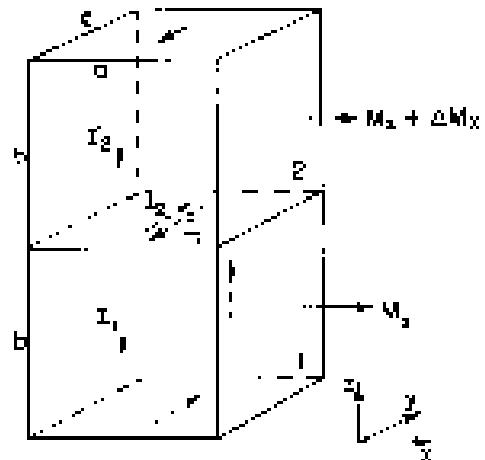


Fig. 36-6. Two blocks, one above the other, may also contribute to \bar{j}_y .

36-2 The field H

Next, we want to insert the current as written in Eq. (36.10) into Maxwell's equations. We get

$$c^2 \nabla \times B = \frac{i}{\epsilon_0} + \frac{\partial E}{\partial t} - \frac{1}{\epsilon_0} \left(j_{pol} + \nabla \times M + \frac{\partial P}{\partial t} \right) - \frac{\partial E}{\partial z}.$$

We can move the term in ΔM to the left-hand side:

$$c^2 \nabla \times \left(B - \frac{M}{\epsilon_0 c^2} \right) = j_{pol} - \frac{i}{\epsilon_0} - \frac{1}{\epsilon_0} \left(E + \frac{P}{c^2} \right). \quad (36.11)$$

As we remarked in Chapter 33, many people like to write $(E - P/c^2)$ as a new vector field B/c^2 . Similarly, it is often convenient to write $(B - M/\epsilon_0 c^2)$ as a single vector field. We choose to define a new vector field H by

$$H = B - \frac{M}{\epsilon_0 c^2}. \quad (36.12)$$

Then Eq. (36.11) becomes

$$c^2 \nabla \times H = j_{pol} - \frac{i}{\epsilon_0}. \quad (36.13)$$

It looks simpler, but all the complexity is just hidden in the letters B and H .

⁷ Or, if you prefer, the current J in each face should be split 50-50 with the blocks on the two sides.

Now we have to give you a warning. Most people who use the mks units have chosen to use a different definition of H . Calling their field H' (Oe , mT , etc.), they still call it H without the prime. It is defined by

$$H' = \mu_0 H = M. \quad (36.14)$$

(Also, they usually write μ_0^{-1} as a pure number 1/4π; then they have one more constant to keep track of!) With this definition, Eq. (36.13) looks even simpler:

$$\nabla \times H' = J_{\text{ext}} + \frac{\partial M}{\partial t}. \quad (36.15)$$

But the difficulties with this definition of H' are, first, that it doesn't agree with the convention of people who don't use the mks system, and second, that it makes H' and B have different units. We think it is more appropriate if H and B have the same units as \mathbf{B} —namely, the units of M , as H' does. But I guess it's going to be an ingrained idea for a week or two. (A dozen? transistors, resistors, and such, yes?) We have to watch out. You will find many books which use the H definition of Eq. (36.1)—rather than our definition of Eq. (36.13), and many other books—especially handbooks of magnetic materials—that use B and H the way we have done. You'll have to be careful to figure out which convention they are using.

One way to tell is by the unit of \mathbf{B} . Remember that in the mks system, B has the same dimensions as the unit of flux: one weber per square meter, equal to 10^8 gauss. In the oersted system, a magnetic moment (a current times an area) is called the "ampere-meter". The unit magnetization M , then, has the units coulombs per meter. But H' (in oersts) has the same units as \mathbf{B} . You can see that this also agrees with Eq. (36.13), since $\nabla \times \mathbf{B}$ (which is of one over length) has the same units as H' (in oersts). People who use the H with B definition often complain about the "unit of H " (with the H definition) being "webers/meter per meter"—thinking of the unit of flux density as a weber/meter. But a "web" is really a dimensionless number, so that doesn't mean much to you. Since our H is equal to H' , erg/cm^2 , if you are using the oersted system, H (in webers/meter²) is equivalent to $4\pi \times 10^{-4}$ times H' (in ampères per meter). It is perhaps more convenient to remember that H (in gauss) = 10^4 times H' (in oersts).

There is one more horrible thing. Many people who use our definition of H have decided to call the ratio of H and B by different names! Even though they have the same dimensions, they call the ratio of H to B "gauss" and the ratio of H to B "oersted" (after Gauss and Oersted, of course). So, in many books you will find graphs with H plotted in gauss and B in oersts. They are really the same unit— 10^{-4} of the mks unit! We have summarized the conversions about magnetic units in Table 36-1.

36-1 The magnetization curve

Now we will look at some simple situations in which the magnetic field is constant, or in which the light charge slowly enough that we can neglect $\partial B/\partial t$ in comparison with j_{ext} . Then the fields obey the equations

$$\nabla \cdot \mathbf{B} = 0, \quad (36.16)$$

$$\nabla \times \mathbf{H} = J_{\text{ext}}/\mu_0 \epsilon_0, \quad (36.17)$$

$$\mathbf{H} \cdot \mathbf{B} = M/\mu_0 \epsilon_0. \quad (36.18)$$

Suppose we have a torus (a donut) of iron wrapped with a coil of copper wire, as shown in Fig. 36-7(a). A current I flows in the wire. What is the magnetic field? The magnetic field will be mainly inside the iron; there, the law of \mathbf{B} will be valid, as shown in Fig. 36-7(b). Since the law of \mathbf{B} is continuous, its divergence is zero, and Eq. (36.16) is satisfied. Next, we write Eq. (36.17) in another form by 16-8

integrating around the closed loop Γ drawn in Fig. 36-7(b). From Stokes's theorem, we have thus

$$\oint \mathbf{B} \cdot d\mathbf{s} = \frac{1}{\epsilon_0 c^2} \int_S J_{\text{wind}} \cdot \mathbf{n} d\mathbf{a}, \quad (36-29)$$

where the integral of J is to be carried out over any surface S bounded by Γ . This surface is cut once by each turn of the winding. Each turn contributes the current I to the integral, and, if there are N turns in all, the integral is NI . From the symmetry of our problem, B is the same all around the curve Γ ; if we assumed that the magnetization were uniform, the field B is also constant along Γ . Eq. (36-19) becomes

$$NI = \frac{NI}{\mu_0 c^2} \cdot$$

where l is the length of the curve Γ . So,

$$B = \frac{NI}{\epsilon_0 c^2 l}. \quad (36-30)$$

This is because B is directly proportional to the magnetizing current, at least like this one that B is sometimes called the *magnetizing field*.

Now all we need is an equation which relates NI to B . But there isn't any such equation! There is, of course, Eq. (36-18), but it is no help because there is no direct relation between M and B for a ferromagnetic material like iron. The magnetization M depends on the whole past history of the iron, and not only on what B is at the moment.

All is not lost, though. We can get solutions in certain simple cases. If we start you with unity-gauge iron—let's say well iron that has been annealed at high temperature. Then in the simple geometry of the torus, all the iron will have the same magnetic history. Then we can say something about M , and therefore about the relation between B and H , from experiments measurements. The field B in the torus is, from Eq. (36-29), given as a constant times the current I in the winding. The field B can be measured by integrating over time the emf in the coil (or in an extra coil wound over the magnetizing coil shown in the figure). This emf is equal to the rate of change of the flux of B , so the integral of the emf with time is equal to B times the cross-sectional area of the torus.

Figure 36-8 shows the relation between B and H , observer with a torus of soft iron. When the current is first turned on, B increases with increasing H along the curve a . Note the different scales on B and H : initially, it takes only a relatively small H to make a large B . Why is B so much larger with the iron than it would be with air? Because there is a large magnetization M which is equivalent to a large surface current on the iron—the field B comes from the sum of this current and the conduction current in the winding. Why M should be so large, we will discuss later.

At higher values of H , the magnetization curve levels off. We say that the iron saturates. With the scales of our figure, the curve appears to become horizontal. Actually, it continues to rise slightly. For large fields, B becomes proportional to H , and with a unit slope. There is no further increase of M . Incidentally, we should point out that if the torus were made of some diamagnetic material, M would be zero and B would equal H for all fields.

The first thing we notice is that curve a on Fig. 36-8—which is the so-called *magnetization curve*—is highly nonlinear. But it's worse than that. When reaching saturation, we decrease the current in the coil to bring H back to zero, the magnetic field B falls along curve b . When H is zero, there is still some B left. Even with no magnetizing current there is a magnetic field in the iron. It has become permanently magnetized. If we now turn on a negative current in the coil, the B - H curve continues along b until the coil is saturated in the negative direction. If we then bring the current back to zero again, B goes along curve c . If we alternate the current between large positive and negative values, the B - H curve goes back and forth along very nearly the curves b and c . If we vary H in some arbitrary

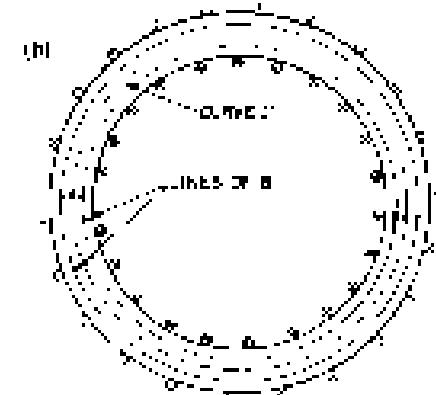
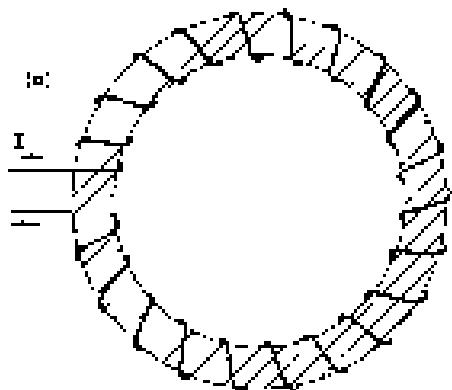


Fig. 36-7. (a) A torus of iron wound with a coil of insulated wire. (b) Cross section of torus showing field lines.

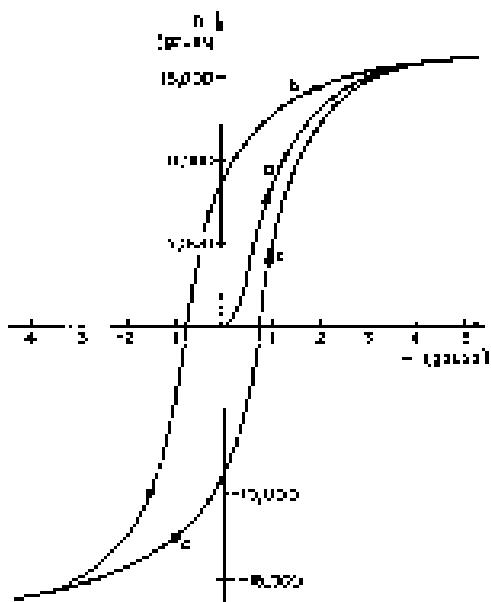


Fig. 36-8. Typical magnetization and hysteresis curve for soft iron.

way. However, we can get more complicated curves which will, in general, lie somewhere between the curves b and c . The loop made by repeated oscillation of the field B is called a hysteresis loop or hysteresis.

We see from this that we cannot write a functional relationship like $B = f(H)$, because the value of B at any instant depends not only on what H is at that time, but on its whole past history. Naturally, the magnetization and hysteresis curves are different for different substances. The shape of the curves depends critically on the elemental composition of the material, and also on the details of its preparation and subsequent physical treatments. We will discuss some of the physical implications of these complications in the next chapter.

36-4 Iron-core inductances

One of the most important applications of magnetic materials is in electrical circuits—for example, in transformers, electric motors, and so on. The reason is that with iron we can control where the magnetic fields go, and also get much larger fields for a given electric current. For example, the typical “toroidal” inductance is made very much like the object shown in Fig. 36-7—but a given inductance, it can be much smaller in volume and use much less copper if it has an equivalent “air-core” inductance. But a point of interest: we get a much greater resistance in the winding, so the inductor is usually very “leaky”—particularly for low frequencies. It is very easy to understand qualitatively how such an inductance works. If I is the current in the winding, then the field B which is produced in the iron is proportional to I , as given by Eq. (36.20). The voltage V across the terminals is related to the magnetic field B . Neglecting the resistance of the winding, the voltage V is proportional to dV/dB . The inductance L , which is the ratio of V to dI/dt (see Section 17-2), thus involves the relation between B and I in the iron. Since the B is so much bigger than the I , we get a large factor in the inductance. Physically, what happens is that a small current in the coil, which would normally produce a small magnetic field, causes the little “Weiss” magnets in the iron to line up and produce a tremendously greater “magnetic” current than the external current in the winding. It is as if we had a lot more current going through the coil than we really have. When we reverse the current, all the little magnets flip over—all those internal currents reverse—and we get a much bigger induced emf than we would get without the iron. If we want to calculate the inductance, we can do so through the energy—as described in Section 17-8. The rate s , which energy is delivered from the current source is I^2L . The voltage V is the cross-sectional area A of the core times N times dV/dB . From Eq. (36.20), $L = \mu_0 A^2/NH$. So we have

$$\frac{dU}{dt} = V^2 = (\mu_0 A^2/N) H \frac{dH}{dt}.$$

Integrating over time, we have

$$U = (\mu_0 A^2/N) \int B dH. \quad (36.21)$$

Notice that $\int B dH$ is the volume of the iron; we will soon show that the energy density $\omega = U/VN$ in a magnetic material is given by

$$\omega = \mu_0 A^2 \int B dH. \quad (36.22)$$

An interesting feature is noticed here. When we use alternating currents, the iron is driven around a hysteresis loop. Since B is not a single-valued function of H , the integral of $\int B dH$ around one complete cycle is not equal to zero. It is the same audience inside the hysteresis curve. Thus, the driving source delivers a certain net energy each cycle—an energy proportional to the area inside the hysteresis loop. And that energy is “lost.” It is lost from the electromagnetic energy or, put this up to Law of Conservation, it is called the hysteresis loss. To keep such energy losses small, we would like the hysteresis loop to be as narrow as

possible. One way to decrease the area of the loop is to reduce the maximum field that is reached during each cycle. For smaller maximum fields, we get a hysteresis curve like the one shown in Fig. 36-9. Also, special materials are designed to have a very narrow loop. The so-called "transformer iron"—which are iron alloys with a small amount of silicon—have been developed to have this property.

When an inductor is run over a small hysteresis loop, the relationships between B and H can be approximated by a linear equation. People usually write

$$B = \mu_0 H \quad (36.23)$$

The constant μ is *not* the magnetic moment we have used before. It is called the *permeability* of the iron. (It is also sometimes called the "relative permeability.") The permeability of ordinary iron is typically several thousand. Transformer iron always has a "superconducting" which has permeabilities as high as a million.

If we use the approximation that $B = \mu_0 H$ in Eq. (36.21), we can write the energy in A toroidal inductors as

$$U = \left(\epsilon_0 c^2 M\right) \mu \int B dH = (\epsilon_0 c^2 / 4) \frac{\mu H^2}{2}. \quad (36.24)$$

So the energy density is approximately

$$\epsilon = \frac{\epsilon_0 c^2}{2} \mu H^2.$$

We can now set the energy of Eq. (36.24) equal to the energy $I^2/2$ of an inductance, and solve for I . We get

$$I = \left(\epsilon_0 c^2 / 4\right) \mu \left(\frac{H}{J}\right)^2$$

Using H/J from Eq. (36.20), we have

$$I = \frac{\mu M^2 A}{\epsilon_0 c^2 J}. \quad (36.25)$$

The inductance is proportional to μ . If you want inductances for such things as audio amplifiers, you will try to operate them on a hysteresis loop where the B - H relationship is as linear as possible. (You will remember that we spoke in Chapter 30, Vol. I, about the generation of harmonics in nonlinear systems.) For such purposes, Eq. (36.24) is a useful approximation. On the other hand, if you want to generate harmonics, you may use an inductance which is intentionally operated in a highly nonlinear way. Then you will have to use the complete B - H curves, and analyze wave shapes by graphical or numerical methods.

A "transformer" is often made by putting two coils on the same core—or core—of a magnetic material. (For the larger "transformers, the core is made with laminations for convenience.) Then a varying current in the "primary" winding causes the magnetic field in the core to change, which induces an emf in the "secondary" winding. Since the flux through each turn of both windings is the same, the emfs in the two windings are in the same ratio as the number of turns on each. A voltage applied to the primary is transformed to a different voltage at the secondary. Since a certain net current around the core is needed to produce the required change in the magnetic field, the algebraic sum of the currents in the two windings will be fixed and equal to the required "magnetizing" current. If the current drawn from the secondary increases, the primary current must increase in proportion—there is a "transformation" of currents as well as voltage.

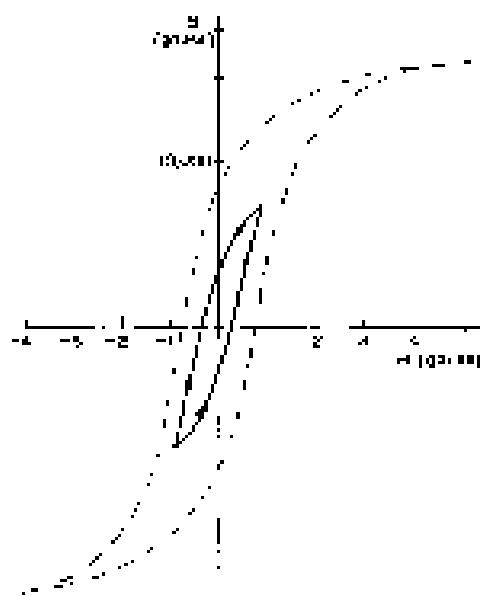


Fig. 36-9. A hysteresis loop that doesn't reach saturation.

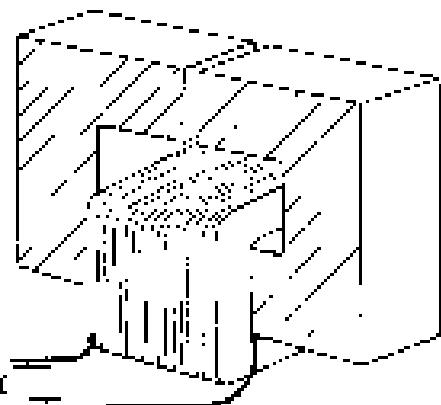


Fig. 36-10. An electromagnet.

36-3 Electromagnets

Now let's discuss a practical situation which is a little more complicated. Suppose we have an electromagnet of the rather standard form shown in Fig. 36-10—there is a "C-shaped" yoke of iron, with a slot or "air gap" in the center. What is the magnetic field B in the gap?

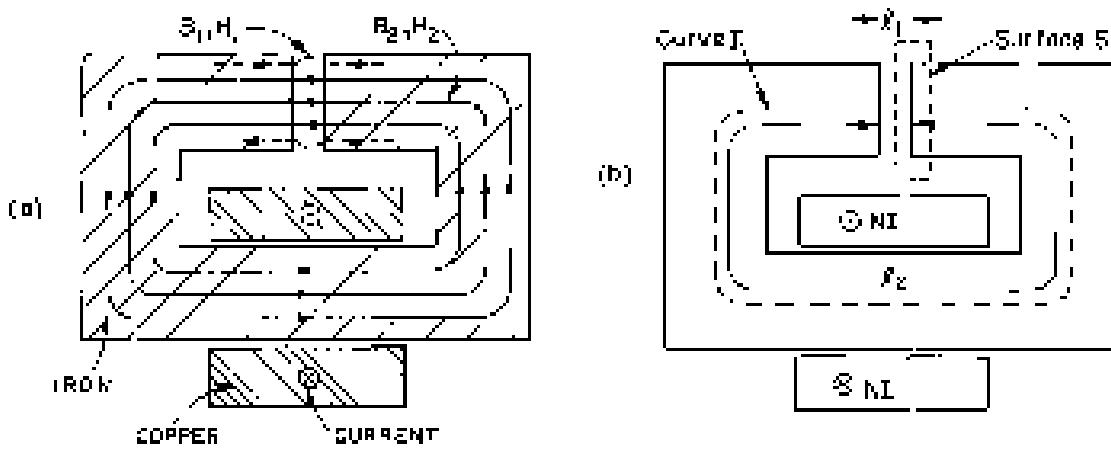


Fig. 36-11. Cross section of an electromagnet.

If the gap thickness is small compared with all the other dimensions, we can, as a first approximation, assume that the lines of \mathbf{B} will go around straight the loop, just as they did in the torus. They will look more or less as shown, in Fig. 36-11(a). They tend to spread out, somewhat in the gap, but if the gap is narrow, this will be a small effect. It is a fair approximation to assume that the flux of \mathbf{B} through any cross section of the yoke is a constant. If the yoke has a uniform cross-sectional area—and if we neglect any edge effects at the yoke ends—the current —we can, say. Let B be uniform across the yoke.

Also, B will have the same value in the gap. This follows from Eq. (36.6), because the closed surface S , shown in Fig. 36-11(b), which lies one face in the gap and the other in the iron. The total flux of \mathbf{B} out of this surface must be zero. Calling B_1 the field in the gap and B_2 the field in the iron, we have that

$$B_1 A_1 - B_2 A_2 = 0.$$

Since $A_1 = A_2$ (in our approximation), it follows that $B_1 = B_2$.

Now let's look at M . We can again use Eq. (36.19), taking the line integral around the curve J , in Fig. 36-11(b). As before, the right-hand side is MI , the number of turns times the current. Now, however, H will be different in the iron and in the air. Call the H_2 the field in the iron and H_1 the path length around the yoke, this part of the curve will contribute the amount $H_2 l_2$ to the integral. Calling H_1 the field in the gap and t the gap thickness, we get the contribution $H_1 l_1$ from the gap. We have that

$$H_1 l_1 + H_2 l_2 = \frac{\mu I}{\epsilon_0 A}. \quad (36.26)$$

Now we know something else: that in the air gap, the magnetization is negligible, so that $B = H$. Since $B_1 = B_2$, Eq. (36.26) becomes

$$B_2 l_2 + H_2 l_2 = \frac{M}{\epsilon_0 A}. \quad (36.27)$$

We still have two unknowns. To find B_2 and H_2 , we need another relationship between the one which relates B to H in the iron.

If we can make the approximation that $B_2 = \mu H_2$, we can solve this equation algebraically. However, let's do the graphical test, in which the magnetization curve of the iron is one like that shown in Fig. 36-3. What we want is the simultaneous solution of this functional relationship together with Eq. (36.27). We can find it by plotting a graph of Eq. (36.27) on the same graph with the magnetization curve, as is done in Fig. 36-12. Where the two curves intersect, we have our solution.

For a given current I , the function (36.27) is the straight line marked $f > 0$ in Fig. 36-12. This line intersects the H -axis ($B_2 = 0$) at $H_2 = M/\mu_0 \epsilon_0 A_2$, and the slope is $-\mu_0/\mu_1$. Different currents just shift the line horizontally. From Fig. 36-10

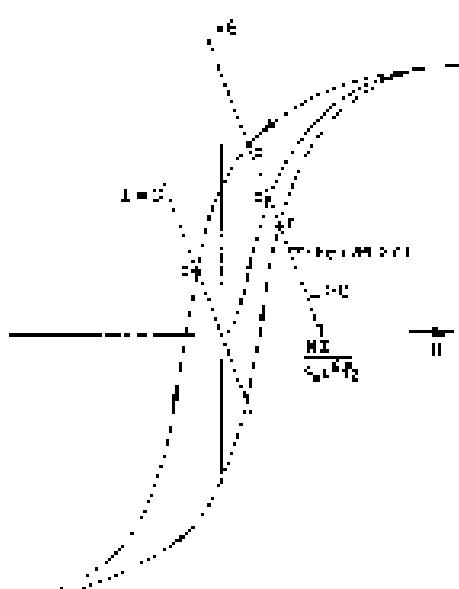


Fig. 36-12. Solving for the field in an electromagnet.

36-12, we see that for a given applied there are several different solutions, depending on how you put it in. If you have just made the magnet and turned the current up to I_1 , the field B_2 (which is also B_1) will have the value given by point *a*. If you increase the current to some very high value and come down to I_1 , the field will be given by point *b*. Or, if you have just had a high negative current in the magnet, and then come up to I_1 , the field is the one at point *c*. The field in the gap will depend on what you have done in the past.

When the current in the magnet is zero, the relation between M_1 and M_2 in Eq. (36.27) is shown by the blue marker ($I = 0$) in the figure. There are still various possible solutions. If you have first saturated the iron, there may be a considerable residual field in the magnet as given by point *d*. You can take the coil off, and you have a permanent magnet. You can see that for a good permanent magnet, you simply want a material with a wide hysteresis loop. Special alloys, such as Alnico V, give very wide loops.

36-6 Spontaneous magnetization

We now turn to the question of why it is that in ferromagnetic materials a small magnetic field produces such a large magnetization. The magnetization of ferromagnetic materials like iron and nickel comes from the magnetic moment of the electrons in the inner shell of the atom. Each electron has a magnetic moment μ equal to $e/2m$ times its v -factor, times its angular momentum J . For a single electron with no net orbital motion, $v = 2$, and the component of J in any direction—say the z -direction—is $-1/2$, so the component of μ along the z -axis is

$$\mu_z = \frac{e\hbar}{2m} = 0.923 \times 10^{-23} \text{ amp}\cdot\text{m}^2. \quad (36.28)$$

In an iron atom, there are actually two electrons that contribute to the ferromagnetism, so to keep the discussion simpler we will talk about nickel, which is ferromagnetic like iron but which has only one electron in the inner shell. It is easy to extend the arguments to iron.

Now the point is that, in the presence of an external field B , the atomic magnets tend to line up with the field, but are knocked about by thermal motions just as we described for paramagnetic materials. In the last chapter we found out that the balance between a magnetic field trying to line up the atomic magnets and the thermal motions tends to dominate their produce the result that the magnetization per unit volume will end up as

$$M = M_0 \tanh \frac{\mu B}{kT}. \quad (36.29)$$

By B_0 we mean the field acting at the atom, and kT is the Boltzmann energy. In the theory of paramagnetism we used for B_0 just B itself, neglecting the part of the field at any given atom contributed by the atoms nearby. In the ferromagnetic case, there is a complication. We shouldn't use the average field in the iron for the B_0 acting on an individual atom. Instead, we must do as we did in the case of dielectrics—we have to find the local field acting on a single atom. For an exact calculation we should add up the fields of the atoms in question contributed by all of the other atoms in the crystal lattice. But as we did for dielectrics, we will make the approximation that the field at a atom is the same as we would find in a small spherical hole in the material—assuming that the moments of the atoms in the neighborhood are not changed by the presence of the hole.

Following the arguments we made in Chapter 11, we might think that we could write

$$B_{\text{atom}} = M + \frac{1}{3} \frac{M}{r_{\text{sat}}}, \quad (\text{wrong}).$$

But that is not right. We can, however, make use of the results of Chapter 11 if we make a careful comparison of the equations of Chapter 11 with the April 30

for ferromagnetism in this chapter. Let's put together the corresponding equations. For regions where there are no conductive currents or charges we have:

Electrostatics Static ferromagnetism

$$\nabla \cdot \left(\mathbf{E} + \frac{\mathbf{P}}{\epsilon_0 c^2} \right) = 0 \quad \nabla \cdot \mathbf{B} = 0 \quad (36.10)$$

$$\nabla \times \mathbf{E} = 0 \quad \nabla \times \left(\mathbf{B} - \frac{\mathbf{M}}{\mu_0 c^2} \right) = 0$$

These two sets of equations can be thought of as analogies if we make the following purely *mathematical* correspondence:

$$\mathbf{E} \rightarrow \mathbf{B} - \frac{\mathbf{M}}{\mu_0 c^2}, \quad \mathbf{E} + \frac{\mathbf{P}}{\epsilon_0 c^2} \rightarrow \mathbf{B}.$$

This is the same as making the analogy

$$\mathbf{E} \rightarrow \mathbf{H}, \quad \mathbf{P} \rightarrow M/c^2. \quad (36.11)$$

In other words, if we write the equations of ferromagnetism as

$$\begin{aligned} \nabla \cdot \left(\mathbf{H} + \frac{\mathbf{M}}{\epsilon_0 c^2} \right) &= 0, \\ \nabla \times \mathbf{H} &= 0, \end{aligned} \quad (36.12)$$

they look like the equations of electrostatics.

This purely algebraic correspondence has led to some confusion in the past. Poynting tended to think that \mathbf{H} was "the magnetic field." But, as we have seen, \mathbf{B} and \mathbf{E} are physically the fundamental fields, and \mathbf{H} is a derived field. Although the equations are analogous, the physics is not analogous. However, that doesn't mean we stop us from using the principle that the same equations have the same solutions.

We can use our earlier results for the electric field inside of holes of various shapes in dielectrics—summarized in Fig. 36-1—to find the field \mathbf{H} inside of corresponding holes. Knowing \mathbf{H} , we can determine \mathbf{B} . For instance (using the results we summarized in Section 1), the field \mathbf{H} in a needle-shaped hole parallel to \mathbf{M} is the same as the \mathbf{E} in the material.

$$H_{\text{hole}} = H_{\text{material}}$$

But since M in the hole is zero, we have

$$H_{\text{hole}} = B_{\text{material}} = \frac{M}{\mu_0 c^2}. \quad (36.13)$$

On the other hand, for a disc-shaped hole, perpendicular to \mathbf{M} , we have

$$E_{\text{hole}} = E_{\text{material}} + \frac{P}{c},$$

which translates into

$$H_{\text{hole}} = B_{\text{material}} = \frac{M}{\mu_0 c^2}.$$

Or, in terms of B ,

$$B_{\text{hole}} = B_{\text{material}}. \quad (36.14)$$

Finally, for a spherical hole, by making our analogy with Eq. (36.3) we would have

$$H_{\text{hole}} = H_{\text{material}} + \frac{M}{\mu_0 c^2}$$

or

$$B_{\text{hole}} = B_{\text{material}} + \frac{2}{3} \frac{M}{\mu_0 c^2}. \quad (36.15)$$

This result is quite different from what we got for E .

It's, of course, possible to get these results in a more physical way, by using the Maxwell equations directly. For example, Eq. (36.34) follows directly from $\nabla \cdot \mathbf{B} = 0$. (You use a gaussian surface that is half in the material and half out.) Similarly, you can get Eq. (36.35) by using a line integral along a curve that goes up inside the hole and returns through the 'material'. Physically, the field in the hole is reduced because of the surface current, which are given by $\nabla \times \mathbf{M}$. We will leave it for you to show that Eq. (36.36) can also be obtained by considering the effects of the surface currents on the boundary of the spherical cavity.

In finding the equilibrium magnetization from Eq. (36.29), it turns out to be most convenient to deal with M , so write

$$B_0 = H + \lambda \frac{M}{kT}, \quad (36.37)$$

In the spherical hole approximation, we would have $\lambda = \beta$, but, as you will see, we will want later to use some other value, so we leave it as an adjustable parameter. Also, we will take all the fields in the same direction so that we won't need to worry about the vector directions. If we were now to substitute Eq. (36.36) into Eq. (36.29), we would have one equation that relates the magnetization M to the magnetizing field H :

$$\alpha x = \tanh \left(\frac{H - \lambda M_{sat}}{kT} \right).$$

It is, however, an equation that cannot be solved explicitly, so we will do it graphically.

Let's put the problem in a generalized form by writing Eq. (36.29) as

$$\frac{M}{M_{sat}} = \tanh x, \quad (36.38)$$

where M_{sat} is a constant value of the magnetization, namely, M_0 , and x represents $(H - \lambda M_{sat})/kT$. The dependence of M/H_{ext} on x is shown by curves in Fig. 36.13. We can also write x as a function of M using Eq. (36.36) for B_0 , as

$$x = \frac{\alpha B_0}{kT} = \frac{\alpha H}{kT} + \left(\frac{\alpha \lambda M_{sat}}{kT^2} \right) \frac{M}{M_{sat}}. \quad (36.39)$$

For any given value of H , this is a straight-line relationship between M/M_{sat} and x . The x -intercept is at $x = \alpha H/kT$, and the slope is $\alpha \lambda^2 T^2 / (\alpha \lambda kT) M_{sat}$. For any particular H , we would have a line like the one marked b in Fig. 36.13. The intersection of curves a and b gives us the solution for M/M_{sat} . We have solved the problem.

Let's look at how the solutions will go for various temperatures. We start with $H = 0$. There are two possible situations, x over or the lines b_1 and b_2 in Fig. 36.14. You will notice from Eq. (36.48) that the slope of the line is proportional to the absolute temperature T . So, at high temperatures we would have x below b_1 . The solution is $M/M_{sat} = 0$. When the magnetizing field H is zero, the magnetization is also zero. But, at low temperatures, we would have x like b_2 , and there are two solutions for M/M_{sat} , one with $M/M_{sat} = 0$ and one with $M/M_{sat} > 0$. It turns out that only the upper solution is stable, as you can see by doing a bit of calculations about these solutions.

According to these ideas, then, a magnetic material should magnetize itself spontaneously at sufficiently low temperatures. In short, when the thermal motions are small enough, the coupling between the atomic magnetic causes them all to line up parallel to each other, we have a permanently magnetized material analogous to the ferroelectrics we discussed in Chapter 11.

If we start at high temperatures and come down, there is a critical temperature, called the Curie temperature T_c , where the ferromagnetic behavior suddenly stops. This temperature corresponds to the line b_3 of Fig. 36.14, which is tangent to the curve a , and has, therefore, a slope of 1. The Curie temperature is given by

$$\frac{e\alpha kT_c}{\mu M_{sat}} = 1. \quad (36.40)$$

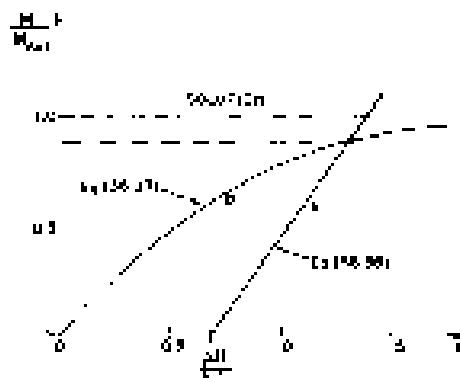


Fig. 36.12. A graphic solution of Eqs. (36.37) and (36.38).

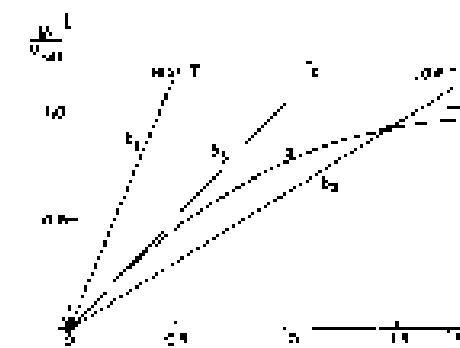


Fig. 36.14. Finding the magnetization when $H = 0$.

We can, if we wish, write Eq. (36.38) more simply in terms of T , as

$$\sigma \approx \frac{\mu H}{kT} + \frac{T_c}{T} \left(\frac{M}{M_{\text{sat}}} \right) \quad (36.40)$$

Now we want to see what happens for small magnetizing fields H . We can see from Fig. 36-4 how things will go if we shift our straight lines a little to the right. For the low-temperature case, the intersection point will move out a little since the low-slope part increases, and M will change relatively little. For the high-temperature case, however, the intersection point runs up the steep part of curve σ , and M will change relatively rapidly. In fact, we can approximate this part of curve σ by a straight line of unit slope, and write:

$$\frac{d\sigma}{dT} = 1 - \frac{\mu H}{kT} - \frac{T_c}{T} \left(\frac{M}{M_{\text{sat}}} \right)$$

Now we can solve for M/M_{sat} :

$$\frac{M}{M_{\text{sat}}} = \frac{kT}{k(T - T_c)} \cdot \frac{\mu H}{1 - \frac{\mu H}{kT}} \quad (36.41)$$

We have a law that is something like the one we had for paramagnetism. For unmagneation, we find

$$\frac{M}{M_{\text{sat}}} = \frac{\mu H}{kT} \quad (36.42)$$

One difference now is that we have the magnetization in terms of H , which includes some of the effects of the saturation of the atomic magnets, but the main difference is that the magnetization is inversely proportional to the difference between T and T_c , instead of to the absolute temperature T , above. Neglecting the interactions between neighboring atoms corresponds to taking $\lambda = 0$, which from Eq. (36.39) means taking $T_c = 0$. Then the results are just what we had in Chapter 35.

We can check our theoretical picture with the experimental data for nickel. It is observed empirically that the Curie-Weiss law above of M vs T disappears when its temperature is raised above 601°K. We can compare this with T_c calculated from Eq. (36.39). Remembering that $A_{\text{sat}} = \mu N$, we have

$$T_c = \frac{3}{N} \frac{A_{\text{sat}}^2}{k\mu_0^2}$$

From the density and atomic weight of nickel, we get

$$N = 9.3 \times 10^{28} \text{ m}^{-3}$$

Taking μ from Eq. (36.28), and setting $\lambda = \frac{1}{3}$, we get

$$T_c = 0.24^\circ\text{K}$$

There is a discrepancy of a factor of about 2600! Our theory of ferromagnetism fails completely.

We can try to "patch up" the theory as Weiss did, by saying that for some unknown reason λ is not one-third, but 2600 times it, or about 2000. It turns out that one gets similar values for other ferromagnetic materials like iron. To see what this means, let's go back to Eq. (36.36). We see that a large λ means that B_s , the local field on the atom, appears to be much, much larger than we would think. In fact, writing $H = B - M/\mu_0 r^2$, we have

$$B_s = B - \frac{(2 - 1)M}{\mu_0 r^2}$$

According to our original idea—with $\lambda = \frac{1}{3}$ —the local magnetization M reduces the effective field B_s by the amount $-4M/\mu_0 r^2$. Even if our model of a spherical hole were not very good, we would still expect some reduction. Instead, to explain

the phenomenon of ferromagnetism, we have to imagine that the magnetization of the field enhances the local field by some large factor—like one thousand or more. There doesn't seem to be any reasonable way to manufacture such enormous fields at an atom—or even likely of the proper sign! Clearly, our “magnetic” theory of ferromagnetism is a dream vision. We must conclude then, that ferromagnetism has to do with some antiferromagnetic interaction between the spinning electrons in neighboring atoms. This interaction must generate a strong tendency for all of the nearby spins to line up in one direction. We will see later that it has to do with quantum mechanics and the Pauli exclusion principle.

Finally, we look at what happens at low temperatures—see $T \ll T_c$. We have seen that there will then be a spontaneous magnetization—even with $H = 0$ —given by the intersection of the curves σ and b_2 of Fig. 36-14. If we solve for M for various temperatures—by varying the slope of the line b_2 —we get the theoretical curve shown in Fig. 36-15. This curve should be the same in all ferromagnetic materials for which the quantum moment consists of a single system. The curves for other materials are only slightly different.

In the limit as T goes to absolute zero, M goes to M_{sat} . As the temperature is increased, the magnetization decreases, falling to zero at the Curie temperature. The points in Fig. 36-16 are the experimental observations for nickel. They fit the theoretical curve quite well. Even though we don't understand the basic mechanism, the general features of the theory seem to be correct.

Finally, there is one more interesting discrepancy in our attempt to understand ferromagnetism. We have found that above some temperature the material should behave like a paramagnetic substance with a magnetization M proportional to H (or B), and that below that temperature it should become spontaneously magnetized. But that's not what we find when we measure the magnetization curves for iron, i.e., only because permanently magnetized after we had “magnetized” it. According to the ideas just discussed, it would demagnetize itself! What is wrong? Well, it turns out that if you look at a small enough piece of iron or nickel, i.e., a piece completely magnetized! But at large pieces of iron, there are many small regions or “domains” that are magnetized in different directions, so that on a large scale the average magnetization appears to be zero. In each small domain, however, the iron has a locked-in magnetization with M equal to M_{sat} . The consequences of this domain structure are that most properties of large pieces of material are quite different from the microscopic properties that we have really been treating. We will take up in the next lecture the story of the practical behavior of bulk magnetic materials.

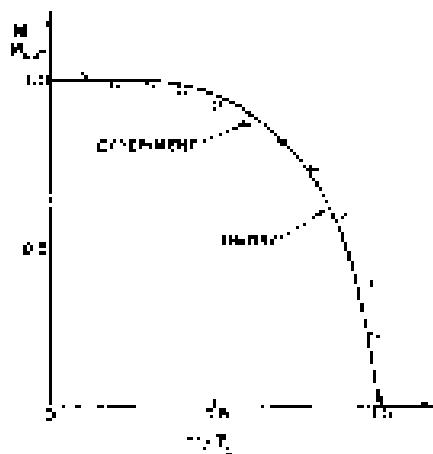


Fig. 36-15. Spontaneous magnetization as a function of temperature for nickel.

Magnetic Materials

37-1 Understanding Ferromagnetism

In this chapter we will discuss the behavior and peculiarities of ferromagnetic materials and of other strong magnetic materials. Before proceeding to study magnetic materials, however, we will review very quickly some of the things about the general theory of magnets that we learned in the last chapter.

First, we imagine the atomic currents inside the material that are responsible for the magnetism, and then describe them in terms of a volume current density $\mathbf{j}_{\text{vol}} = \mathbf{v} \times \mathbf{M}$. We emphasize that this is not supposed to represent the actual currents. When the magnetization is uniform the currents do not really cancel out precisely; that is, the whirling currents of one electron in one atom and the whirling currents of an electron in another atom do not overlap in such a way that the sum is exactly zero. Even within a single atom the distribution of magnetism is not smooth. For instance, in an iron atom the magnetization is distributed in a more or less spherical shell, not too close to the nucleus and not too far away. Thus, magnetism in matter is quite a complicated thing in its details; it is very irregular. However, we are obliged now to ignore this detailed complexity and discuss phenomena from a gross, average point of view. Then it is true that the average current in the interior region over any little area that is big compared with an atom is zero when $\mathbf{M} = 0$. So, what we mean by magnetism is not volume and long range, so on, at the tree, we are now concerning us an average over regions that are large enough just with the spins occupied by a single atom.

In the last chapter, we also discovered that a ferromagnetic material has the following interesting property: above a certain temperature it is not strongly magnetic, whereas below this temperature it becomes magnetic. This fact is easily demonstrated. A piece of nickel wire at room temperature is attracted by a magnet. However, if we heat it above its Curie temperature with a gas flame, it becomes non-magnetic; it is attracted toward the magnet—even when brought quite close to the magnet. If we cool the wire the magnet while it exceeds T_C , the instant its temperature falls below the critical temperature it is suddenly attracted again by the magnet!

The general theory of ferromagnetism [as we will see] supposes that the spin of the electron is responsible for the magnetization. The electron has spin (one-half the charge) and Bohr magneton ($1/2$ a quantum of magnetism) $\mu_B = e\tau = q\hbar/2m_e$. The electron spin can be pointed either "up" or "down." Because the electron has a negative charge, when its spin is "up" it has a negative moment, and when its spin is "down" it has a positive moment. With our usual conventions, the moment \mathbf{m} of the electron is opposite its spin. We have found that the energy of orientation of a magnetodipole in a given applied field \mathbf{B} is $-\mu \cdot \mathbf{B}$, and the energy of the spinning electron depends on the neighboring spin alignments as well. In iron, if the moment of a nearby atom is "up," there is a very strong tendency that the moment of the one next to it will also be "up." That is what makes iron, cobalt, and nickel so strongly magnetic—the moments all want to be parallel. The first question we have to discuss is why.

Soon after the development of quantum mechanics, it was noticed that there is a very strong quantum force—not a magnetic force or any other kind of normal force, but only an apparent force—trying to line the spins of nearby electrons opposite to one another. These forces are closely related to chemical valence forces. There is a principle in quantum mechanics called the exchange principle that

37-1 Understanding Ferromagnetism

37-2 Thermodynamic properties

37-3 The hysteresis curve

37-4 Ferrimagnetic materials

37-5 Extraordinary magnetic materials

References: Becker, R. M., "Magnetism," *Encyclopedia Britannica*, Vol. 14, 1957, pp. 645-647.

Kittel, C., *An Introduction to Solid State Physics*, John Wiley and Sons, Inc., New York, 2nd ed., 1956.

Two electrons cannot occupy exactly the same state, and they cannot be in exactly the same position as to location and spin orientation.² For example, if they are at the same point, the only alternative is to have their spins opposite. So, if there is a gap of space between atoms where electrons like to congregate (as in a chemical bond) and we want to put another electron on top of one already there, the only way to do it is to have the spin of the second one parallel opposite to the spin of the first one. To have the spins parallel is against the law, unless the electrons stay away from each other. This has the effect that a pair of parallel-spin electrons try to go in different directions away from a pair of oppo. in-spin electrons; the net effect is as though there were a force trying to turn the spin layer. Sometimes this spin-torsion force is called the exchange force, but that only makes it more mysterious, it is not a very good term. It is just, because of the exclusion principle, that electrons have a tendency to have their spins opposite. In fact, that is the explanation of the kind of magnetism in most solid substances! The spins of the free electrons and the moments of the nuclei have tremendous tendency to balance in opposite directions. The problem is to explain why for materials like iron, cobalt, etc., the forces of what we should expect.

We can summarize the proposed magnetic effect by adding a suitable term to the energy equation, by saying that if the electrons happen to be neighboring and have a given magnetization M , then the moment of an electron has a strong tendency to have the same orientation as the average magnetization of the atoms in the neighborhood. Thus, we may write for the two possible spin orientations:

$$\begin{aligned} \text{Spin "up" energy} &= E_0 \left(H + \frac{\lambda M}{\epsilon_0 c^2} \right), \\ \text{Spin "down" energy} &= -E_0 \left(H + \frac{\lambda M}{\epsilon_0 c^2} \right). \end{aligned} \quad (3.11)$$

When it was clear that quantum mechanics could supply a reasonable spin-orbit coupling, even if, apparently, of the wrong sign, it was suggested that ferromagnetism might arise in similar fashion, but due to the complexities of iron and the large number of electrons involved, the size of the coupling energy would come out the other way around. Since the time this was written off in 1937, some quantum mechanics was finally understood, many people have been working very seriously and successfully, trying to get the correct prediction for λ . The best calculations of the energy between the two electrons spins in iron, assuming that the interaction is a direct one between the two electrons in neighboring atoms, will give the wrong sign. The present understanding of this is again to assume that the complexity of the situation is somehow responsible and to hope that the perturbation which takes us into a more complicated situation will give the right answer!

It is believed that the energy of one valence electron in the neighborhood, which is making the magnetism come to pass, is completely electric in character, and the outside have the opposite spin. One might expect this to happen because the conduction electrons come into the same region as the "magnetic" electrons. Since they move around, they can carry their magnetism by being upside down over to the next atom; thus, the "magnetic" electron tries to force the conduction electrons to be opposite, and the conduction electron then makes the next "magnetic" electron opposite to it. The double interaction is a very similar to the interaction which tries to line up the two "magnetic" electrons. In other words, the tendency to make parallel spins is the result of no necessity that needs to cause either to be opposite to both. This mechanism does not require that the conduction electron be completely "upside down." They can do just about a slight angle to be enough, just enough to "tip" the "magnetic" atom the other way. Thus, if the magnetism is big

² See Chapter 13.

³ We write these equations with $H = B - \lambda M/c^2$, so that $\lambda M/c^2$ is added to the work of the last chapter. You might prefer to write $H = -\mu B - \mu(\lambda B + \lambda M/c^2)/c^2$, where $\lambda = 1 - 1/c^2$ for very things.

the people who have calculated such things now believe is responsible for ferromagnetism. But we must emphasize that to this day nobody can calculate the magnitude of α simply by knowing that the material is number 26 in the periodic table. In short, we don't thoroughly understand it.

Now let us continue with the theory, and then come back later to discuss a certain error involved in the way we have set it up. If the magnetic moment of a certain electron is "up," energy comes both from the external field and also from the tendency of the spins to be parallel. Since the energy is lower when the spins are parallel, the effect is sometimes described as due to an "effective internal field." But remember, i.e., is not due to a true magnetic force, it is an interaction that is more complicated. In any case, we use Eqs. (37.1) as the formulas for the energies of the two spin states of a "magnetic" electron. At a temperature T , the relative probability of these two states is proportional to $e^{-E/E}$, which we can write as $e^{-\beta E}$, with $\beta = \mu(H + 2M/k_B T)/k_B^2$. Then, if we continue the model so far all the magnetic moments, we find that in the last chapter that is

$$M = N_\text{A} \mu \tanh \beta. \quad (37.2)$$

Now we would like to calculate the internal energy of the material. We note that the energy of an electron is exactly proportional to the magnetic moment, so that the calculation of the mean moment and the calculation of the mean energy are the same—except that in place of μ in Eq. (37.2) we would write $-M\beta$, which is $-\mu(H + 2M/k_B T)^2/k_B^2$. The mean energy is then

$$\langle U \rangle_{\text{av}} = -N_\text{A} \left(H + \frac{M\beta}{k_B T} \right) \tanh \beta.$$

Now this is very queer! The term $M\beta/k_B T$ represents interactions of all possible pairs of atoms, and we must remember to count each pair only once. (When we calculate the energy of one electron in the field of the rest and then the energy of a second electron in the field of the rest, we have counted part of the Coulomb energy twice.) Thus, we must divide the *internal interaction term* by two, and our formula for the energy then turns out to be

$$\langle U \rangle_{\text{av}} = -N_\text{A} \left(H + \frac{M\beta}{2k_B T} \right) \tanh \beta. \quad (37.3)$$

In the last chapter we discovered an interesting thing—that below a certain temperature the material finds a solution to the equations in which the magnetic moment A is zero, even with no external magnetizing field. When we set $H = 0$ in Eq. (37.3), we found that

$$\frac{M}{M_{\text{sat}}} = \tanh \left(\frac{T_c}{T} \frac{M}{M_{\text{sat}}} \right). \quad (37.4)$$

where $M_{\text{sat}} = N_\text{A} \mu$ and $T_c = (2M_{\text{sat}}/\mu k_B)^{1/2}$. When we solve this equation (graphically or otherwise), we find that the ratio M/M_{sat} , as a function of T/T_c , is a curve like that labeled "quantum theory" in Fig. 37.1. The dashed curves marked "copper, nickel" show the experimental results for crystals of these elements. The theory and experiment are in moderately good agreement. The figure also shows the result of the classical theory in which the calculation is carried out assuming that the atoms happen to have all possible orientations in space. You can see that this assumption gives a prediction that is not even close to the experimental facts.

Even the quantum theory deviates from the observed behavior at both high and low temperatures. The reason for the deviations is that we have made a rather simple approximation in the theory: We have assumed that the energy of a given electron comes from magnetization of its neighboring atoms. In other words, for each one that is "up" in the neighborhood of a given atom, there will be a contribution of energy due to this quantum mechanical alignment effect. But how many *are* there pointed "up"? On the average, that is measured by the

Fig. 37-1. The spontaneous magnetization $M = M_0$ of ferromagnetic crystals as a function of temperature. [Permission from Encyclopaedia Britannica.]

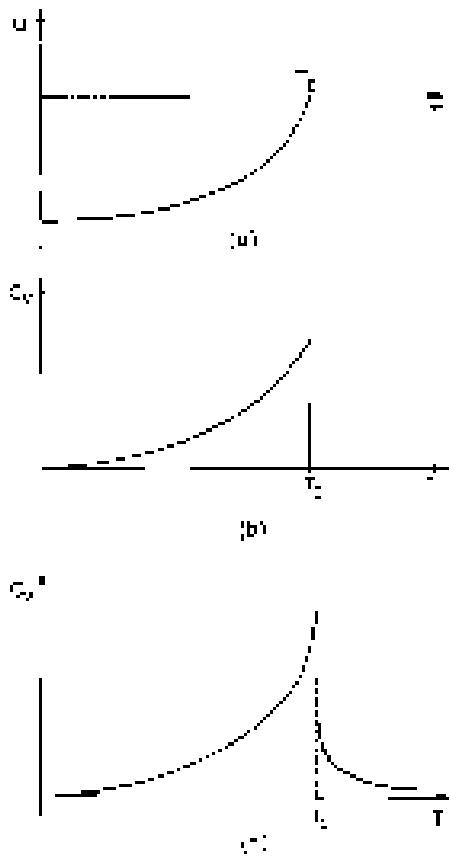
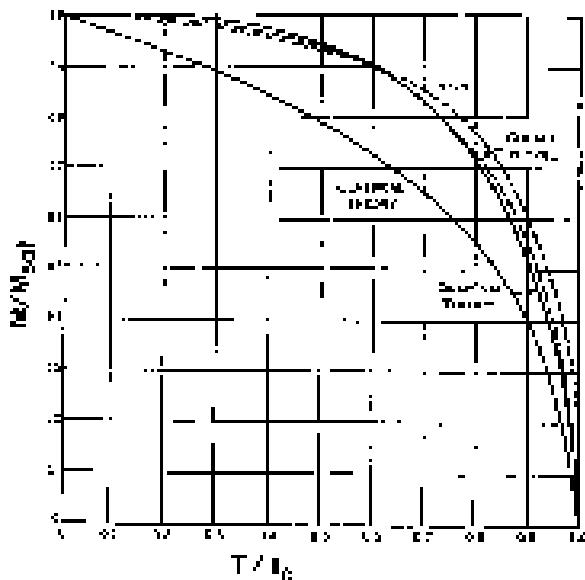


Fig. 37-2. The energy per unit volume and specific heat of a ferromagnetic crystal.

magnetization M —but only on the average. A particular atom somewhere might find all its neighbors "up." Then its energy will be larger than the average. Another one might find some up and some down, perhaps averaging to zero, and it would have no energy from that term, and so on. What we ought to do is to use a somewhat complicated kind of average, because the atoms in different places have different environments, and the numbers up and down are different for different cells. Instead of just taking one atom, supposed to be the average individual, we should take each one in its actual situation, compute its energy, and find the average average. But how do we find out how many are "up" and how many are "down" in the neighborhood? That is, of course, just what we are trying to calculate—the numbers "up" and "down"—so we have a very complicated autocorrelation problem of correlations, a problem which has never been solved. It is an intriguing and tantalizing one which has cast a fog over much of the related research in physics for several years, but even they have not completely solved it.

I think that at low temperatures, + or - almost all the atomic magnets are "up" and only a few are "below," it is easy to solve, and at high temperatures, far above the Curie temperature T_C , when they are almost all random, it is again easy. It is also easy to calculate small temperatures from some simple idealized situation, so I am fairly well understood why there are deviations from the simple theory at low temperature. It is less understood physically that for physical reasons the magnetization might deviate at high temperatures. But the exact behavior near the Curie point has never been thoroughly figured out. This is an interesting problem to work out someday if you want a problem that has never been solved.

37-2 Thermodynamic properties

In the last chapter we laid the groundwork necessary for calculating the thermodynamic properties of ferromagnetic materials. These are, naturally, related to the internal energy of the crystal, which includes interactions of the various spins, given by Eq. (37.1). For the energy of the spontaneous magnetization below the Curie point, we can set $M = 0$ in Eq. (37.1), and noticing that $m = M/M_0$, we find a mean energy proportional to M^2 :

$$(37.5) \quad \langle U \rangle_s = -\frac{8\pi\mu_0 M^2}{350e^2 M_0^2}.$$

If we now plot the energy due to the magnetism as a function of temperature, we get a curve which is the negative of the curve of the curve of Fig. 37-1, as shown in Fig. 37-2(a). If we were to measure the specific heat of such a magnet we would obtain a curve which is the derivative of 37-2(a). It is shown in Fig. 37-4.

37-3g). It rises slowly with increasing temperature, but falls according to zero at $T = 0$. The sharp dip is due to the change in sign of the magnetic energy and is reached right at the Curie point. So without any magnetic measurement at all we could have discovered that something was going on inside of iron or nickel by measuring this thermodynamic property. However, both experiment and improved theory (with fluctuations included) suggest that this simple curve is wrong and that the true situation is really more complicated. The curve goes higher at the peak and falls to zero much more slowly. Even if the temperature is high enough to randomize the spins on the average, there are still local regions where there is a certain amount of polarization, and in these regions the spins will have a little extra energy of interaction which only dies out slowly as things get more and more random with further increases in temperature. So the actual curve looks like Fig. 37-2(c). One of the challenges of theoretical physics today is to find an exact theoretical description of the character of the specific heat near the Curie transition—an intriguing problem which has not yet been solved. Naturally, this problem is very closely related to the shape of the magnetization curve in the same region.

Now we want to describe some experiments, other than thermodynamic ones, which show that there is something right about our interpretation of magnetism. When the material is magnetized to saturation at low enough temperatures, M is very nearly equal to M_s , - nearly all the spins are parallel, as well as their magnetic moments. We can check this by an experiment. Suppose we suspend a bar magnet by a thin fiber and then surround it by a coil so that we can reverse the magnetic field without touching the magnet or putting any torque on it. This is a very difficult experiment, because the magnetic forces are so enormous that any inaccuracy, any deposit of dust, or any lack of perfection in the coil will produce accelerated decays. However, the experiment has been done under careful conditions in which such accelerations are minimized. By means of the magnetic field from a coil that surrounds the bar, we turn all the atomic magnets over at first. When we do this we also change the angular momenta of all the spins from "up" to "down" (see Fig. 37-3). If angular momentum is to be conserved when the spins all turn over, the axis of the bar must have an opposite change in angular momentum. The whole magnet will start to spin. And sure enough, when we do the experiment, we find a slight turning of the magnet. We can measure the total angular momentum given to the whole magnet, and this is simply N times \hbar , the change in the angular momentum of each spin. The ratio of angular momentum to angular momentum is $1/N$. This why you can't be within about 10 percent of what we calculate. Actually, one says that the atoms' magnetic moments are due, mostly to the electrons. But there is, in addition, some orbital motion also in these materials. The orbital motion is not completely free of the lattice and does not contribute much more than a few percent to the magnetism. As a matter of fact, the saturation magnetic field that one gets taking $M_s \approx M_p$ and using the density of iron of 7.9 and the mass m_e of the spinning electron is about 20,000 gauss. This is a typical magnitude of error - 5 or 10 percent - due to neglecting the contributions of the orbital motions that have not been included in solving the analysis. Thus, a slight discrepancy with the gyromagnetic measurements is quite understandable.

37-3 The hysteresis curve

We have concluded from our theoretical analysis that a ferromagnetic material should spontaneously become magnetized below a certain temperature so that all the magnetic would be in the same direction. But we knew that this is not true for an ordinary piece of unannealed iron. Why isn't all iron magnetized? We can explain it with the help of Fig. 37-4. Suppose the iron were all a big single crystal of the shape shown in Fig. 37-4(a) and spontaneously magnetized all in one direction. Then there would be a considerable external magnetic field, which would have a lot of energy. We can reduce that field energy if we arrange that one side is

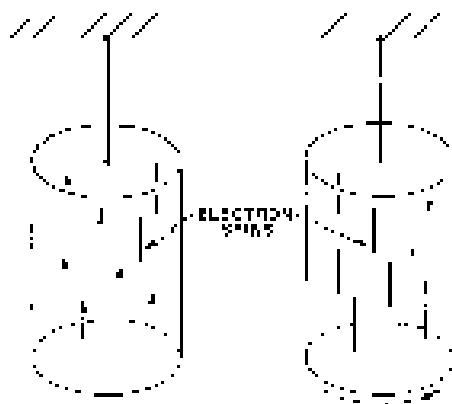
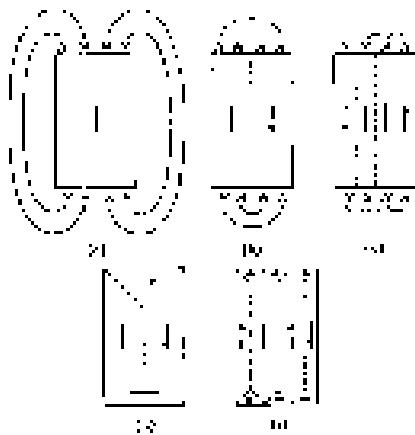


Fig. 37-3. When the magnetization of a bar of iron is reversed, the bar is given some angular velocity.



Let's say it is magnetized "up" and the other side magnetized "down," as in Fig. 37-4(b). Then, since the two take up less volume, so there would be less energy there.

But, but wait! In the layer between the two regions we have up-spinsing electrons adjacent to down-spinsing electrons. But ferromagnetism appears only in those materials for which the energy is reduced if the electrons are parallel rather than opposite. So, we have added some extra energy along the dotted line in Fig. 37-4(b); this energy is sometimes called *wall energy*. A region having only one direction of magnetization is called a domain. At the interface—the "wall"—between two domains, where we have up-spins on opposite sides which are spinning in different directions, there is an energy per unit area of the wall. We have described it as though two adjacent atoms were spinning exactly opposite, but it turns out that nature adjusts things so that the transition is more gradual. But we don't need to worry about such fine details at this point.

Now the question is: When is it better or worse to make a wall? The answer is that it depends on the size of the domains. Suppose that we were to scale up a block so that the whole thing was twice as big. The volume in the space outside filled with a given magnetic field strength would be eight times bigger, and the energy in the magnetic field, which is proportional to the volume, would also be eight times greater. But the surface area between two domains, which will give the wall energy, would be only four times as big. Therefore, if the piece of iron is big enough, it will pay to split it into many domains. This is why very big crystals can have but a single domain. Any large object—say, more than about a μm long² of a millimeter in size—will have at least one domain wall, and the larger the "magnetic" object will be split into many domains, as shown in the figure. Spinning out a domain gives you just the energy needed to get it moving; the wall is as large as the energy difference in the magnetic field outside the crystal.

Actually nature has devised a still another way to lower its energy. It is not necessary to have that and previously described, a little bit of the domain magnetized sideways, as in Fig. 37-4(a). Then with the arrangement of Fig. 37-4(c) we see that there is no external field, but instead only a little more domain wall.

But this introduces a new kind of problem. It turns out that when a single crystal of iron is magnetized, it changes its length in the direction of magnetization, so an "ideal" cube with its magnetization, say, "up," is no longer a perfect cube. The "vertical" dimension will be different from the "horizontal" dimension. This effect is called *magnetostriction*. Because of such geometric changes, the little triangular pieces of Fig. 37-1(d) do not, so to speak, "fit" into the available space anymore—the crystal has got too long one way and too short the other way. Of course, it does fit, really, but only by being squashed in; and this involves some mechanical stresses. So, this arrangement also introduces an extra energy. It is the balance of all these various energies which determines how the domains finally arrange themselves in their complicated fashion in a piece of unmagnetized iron.

Now, what happens when we put on an external magnetic field? To take a simple case, consider a crystal whose domains are as shown in Fig. 37-4(c). If we apply a horizontal magnetic field in the upward direction, in what manner does the crystal become magnetized? First, the middle domain wall can move over sideways (to the right) and reduce the energy. It turns out so that the region—which is "up" becomes bigger than the region which is "down". There are more elementary magnets lined up with the field, and it is given a lower energy. So, for a piece of iron in weak fields—at the very beginning of magnetization—the domain walls begin to move and ear into the regions which are magnetized opposite to the field. As the field continues to increase, a whole crystal shifts gradually into a single

² You may be wondering how come that here the words "up" or "down" can also be "sideways"! That's a good question, but we won't worry about it right now. Well, sort of... In the classical point of view, thinking of the atomic magnets as classical dipoles which can be polarized sideways. Quantum mechanics requires somewhat experience to understand how things can be quantized but "up-down-left-right" and "right-and-left" all at the same time.

large domain which the external field helps to keep lined up. In a strong field, the crystal "lives" to bend one way just because its energy in the applied field is reduced. It is no longer merely the crystal's own external field which matters.

What if the geometry is not so simple? What if the axes of the crystal and its spontaneous magnetization are in one direction, but we apply the magnetic field in some other direction—say at 15° ? We might think that domains would return themselves with their magnetization parallel to the field, and then as before, they could all grow into one domain. But this is not easy for the iron to do, for the energy needed to magnetize a crystal depends on the direction of magnetization relative to the crystal axis. It is relatively easy to magnetize iron in a direction parallel to the crystal axes, but it takes more energy to magnetize it in some other direction—like 15° with respect to one of the axes. Therefore, if we apply a magnetic field in such a direction, what happens first is that the domains which point along one of the preferred directions which is near to the applied field grow until the magnetization is all along one of those directions. Then, with much stronger fields, the magnetization is gradually pulled around, parallel to the field, as sketched in Fig. 37-5.

In Fig. 37-6 are shown some observations of the magnetization curves of single crystals of iron. To understand them, we must first explain something about the notation that is used in describing directions in a crystal. There are many ways in which a crystal can be sliced so as to produce a face which is a plane of atoms. Everyone who has driven past an orchard or surveyed lawns knows this. It is fascinating to watch. If you look carefully, you see lines of trees; if you look another way, you see different lines of trees, and so on. In a similar way, a crystal has definite "planes" of planes that hold many atoms, and the planes have this important characteristic (we consider a cubic crystal to make it easier): If we observe where the planes intersect the three Cartesian axes, we find that the reciprocals of the three coordinates from the origin are in the ratio of simple whole numbers. These three whole numbers are taken as the definition of the planes. For example, in Fig. 37-7(a), a plane parallel to the yz plane is shown. This is called a [100] plane; the reciprocals of its intersection of the x - and z -axes are both zero. The direction perpendicular to such a plane (in a cubic crystal) is given the same set of numbers. It is easy to understand the idea in a cubic crystal, for then the indices [100] mean a vector which has a unit component in the x -direction and none in the y - or z directions. The [1,0,0] direction is in a direction 15° from the x - and y -axes, as in Fig. 37-7(a); and the [111] direction is in the direction of the cube diagonal, as in Fig. 37-7(c).

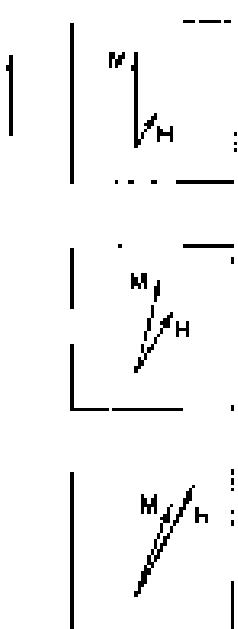


Fig. 37-5. A magnetizing field H at an angle with respect to the crystal axis will gradually change the direction of the magnetization without changing its magnitude.

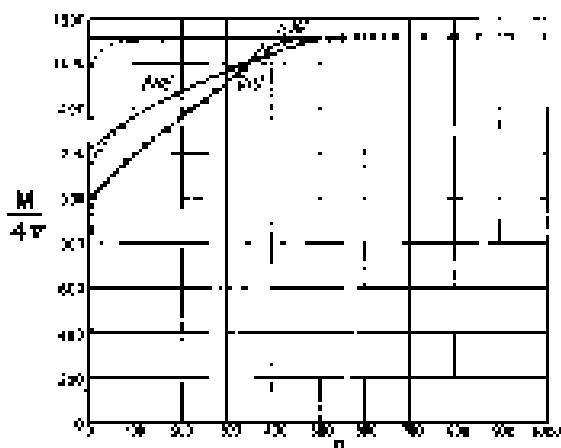


Fig. 37-6. The component of M parallel to H , for different directions of H (with respect to the crystal axes). [From E. Biller, *Introduction to Ferromagnetism*, McGraw-Hill Book Co., Inc., 1937.]

Returning now to Fig. 37-6, we see the magnetization curves of a single crystal of iron for various directions. First, note that for very tiny fields—so weak that it is hard to see them on the scale at all—the magnetization increases extremely rapidly to quite large values. If the field is in the [100] direction—namely along one of those nice, easy directions of magnetization—the curve goes up to a high value, curves around a little, and then is saturated. What happened is that the

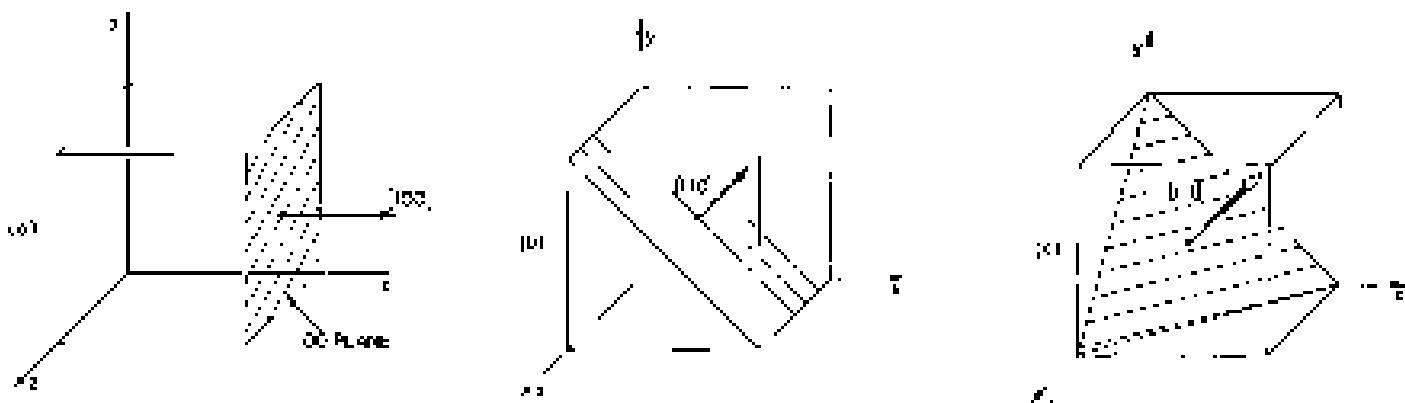


Fig. 37-7. The way the crystal planes are labeled.

domains which were already there are very easily reviewed. Only a small field is required to make the domain walls move and set up all of the "two-way" domains. Single crystals of iron are enormously permeable (magnetic sense), much more so than ordinary polycrystalline iron. A perfect crystal magnetizes extremely easily. Why is it curved at all? Why does it just go right up to saturation? We are not sure. You might study that some day. We do understand why it is flat for high fields. When the whole block is a single domain, the extra magnetic field cannot make any increase in the magnetization—it is already at M_s , with all the electrons lined up.

Now, if we try to do the same thing in the [110] direction—which is at 45° to the crystal axes—what will happen? We turn on a little bit of field and the magnetization loops up as the domains grow. Then as we increase the field some more, we find that it takes quite a lot of field to get up to saturation, because now the magnetization is turning easier from an "easy" direction. If this explanation is correct, the point at which the [110] curve extrapolates back to the vertical axis should be at $1/\sqrt{2}$ of the saturation value. It turns out, in fact, to be very, very close to $1/\sqrt{2}$. Similarly, in the [111] direction—which is along the cube diagonal—we find, as we would expect, that the curve extrapolates back to nearly $1/\sqrt{3}$ of saturation.

Figure 37-8 shows the corresponding curves for two other elements, nickel and cobalt. Nickel is different from iron. In nickel, it turns out that the [111] direction is the easy direction of magnetization. Cobalt has a hexagonal crystal form, and people have figured out the system of magnetization for this case. They want to have three axes on the bottom of the octahedron and one perpendicular to these, so they have used four indices. The [1000] direction is the direction of the easy axis of the hexagon, and [0100] is perpendicular to that axis. We see that crystals of different metals behave in different ways.

Now we must start a polycrystalline material, such as in your many pieces of iron. Inside such materials there are many, many little crystals with their crystalline axes pointing every which way. These are not as easy to analyze. Remember that the domains were all part of a single crystal, but in a piece of iron there are

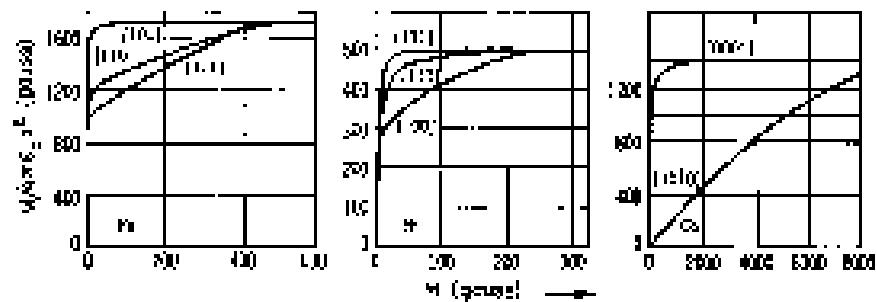


Fig. 37-8. Magnetization curves for single crystals of iron, nickel, and cobalt. (From Charles Kittel, *Introduction to Solid State Physics*, John Wiley and Sons, Inc., New York, 2nd ed., 1956.)

many different crystals with axes at different orientations, as shown in Fig. 37-9. Within each of these crystals, there will also generally be some domains. When we apply a small magnetic field to a piece of polycrystalline material, what happens is that the domain walls begin to move, and the domains which have a favorable direction of easy magnetization grow larger. This process is reversible so long as the field stays very small. If we turn the field off, the magnetization will return to zero. The part of the magnetization curve is marked in Fig. 37-10.

For larger fields—in the region B of the magnetization curve shown—things get more complicated. In every small crystal of the material, there are strains and dislocations; there are impurities, shift, and imperfections. And at all but the smallest fields, the domain wall, in moving, gets stuck or loose. There is an interaction of energy between the domain wall and a dislocation, or a grain boundary, or an impurity. So when the wall tries to run off, it gets stuck. It stays there at a certain field. But then if the field is raised a little more, the wall suddenly snaps past. So the motion of the domain wall is not smooth the way it is in a perfect crystal—it gets hung up every once in a while and moves in jerks. If we were to look at the magnetization on a microscopic scale, we would see something like the insert of Fig. 37-10.

Now the important thing is that these jerks in the magnetization can cause an energy loss. In the first place, when a boundary finally slips past an impediment, it moves very quickly to the next one, since the field is already above what would be required for the unimpeded motion. The rapid motion means that there are rapidly changing magnetic fields which produce eddy currents in the crystal. These currents lose energy in heating the metal. A second effect is that when a domain suddenly changes, part of the crystal changes its dimensions from the magnetostriction. Each sudden shift of a domain wall sets up a little sound wave that carries away energy. Because of such effects, the second part of magnetization curve is irreversible, and there is energy loss here. This is the origin of the hysteresis effect, because to move a boundary wall forward—snap—and then to move it back—snap—produces a different result. It's like "jerky" friction, and it takes energy.

Eventually, for large enough fields, when we have moved all the domain walls and magnetized each crystal in its best direction, there are still some crystallites which happen to have their easy directions of magnetization not in the direction of our external magnetic field. Then it takes a lot of extra field to turn these magnetic moments around. So the magnetization increases slowly, but smoothly, for high fields—namely in the region marked C in the figure. The magnetization does not snap sharply to its saturation value, because not to say, just all the easy directions align are having in the strong field. So we see why the magnetization curve of an ordinary polycrystal is hysteresis, such as the one shown in Fig. 37-10. Notice that the last and reversible part has three steps by, and then curves over steadily. Of course, there is no sharp break-point between the three regions—they blend smoothly one into the other.

It is not hard to show that the magnetization process in the middle part of the magnetization curve is jerky—that the domain walls jerk and snap as they shift. All you need is a coil of wire with many thousands of turns, connected to an amplifier and a loudspeaker, as shown in Fig. 37-11. If you put a few thin iron sheets (of the type used in transformers) at the center of the coil and bring a bar magnet slowly near the steel, the sudden changes in magnetization will produce imulses of electricity in the coil, which are heard as distinct clicks in the loudspeaker. As you move the magnet nearer to the iron you will hear a whole rush of clicks that sound something like the noise of sand grains falling over each other as a can of sand is tilted. The domain walls are jumping, stopping, and jiggling as the field is increased. This phenomenon is called the Barkhausen effect.

As you move the magnet even closer to the iron sheets, the noise grows louder and louder for a while but then there is relatively little noise when the contact gets very close. Why? Because nearly all the domain walls have moved as far as they can go. Any greater field is merely turning the magnetization in each crystal, which is a smooth process.



Fig. 37-9. The microscopic structure of an unmagnetized ferromagnetic material. Each crystal grain has an easy direction of magnetization and is broken up into domains which are spontaneously magnetized usually parallel to this direction.

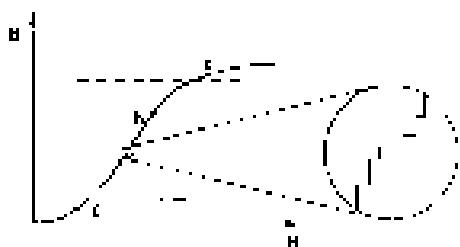


Fig. 37-10. The magnetization curve for polycrystalline iron.

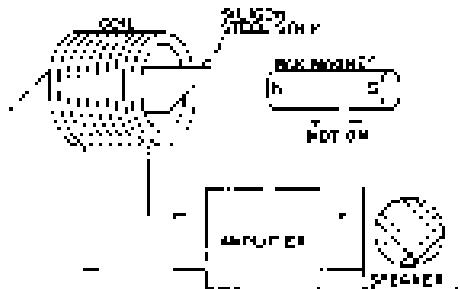


Fig. 37-11. The sudden changes in the magnetization of the steel strip are heard as clicks in the loudspeaker.

If you play with your magnet, as we have done, you can see the downward branch of a hysteresis loop. You might try to get back to zero energy again, and you find a series of "backward-going peaks." You can appreciate that if you bring the magnet too far down, it becomes unstable, because then, the magnetic force is very little indeed. It is unique, like iron, to have a "soft" metal, one the plumb softest piece, in all measurements of the magnetic field of the magnet, because the same current in the magnetic field can't produce the next tiny hysteresis loop, say in the "up" part.

37-4 Ferromagnetic materials

Now we would like to talk about the various kinds of magnetic materials that there are in the technical world and to consider some of the problems involved in designing magnetic materials for different purposes. First, the term "the magnetic properties of iron," which one often hears, is a misnomer. There is no such thing. "Iron" is not a well-defined material—the properties of iron depend critically on the amount of impurities and also on how the iron is formed. You can appreciate that the magnetic properties will depend on how easily the domain walls move and that this is a gross property, not a property of the individual atoms. So practical ferromagnetism is not really a property of an iron atom—it is a property of what you do to a certain atom. For example, iron can take on two different crystalline forms. The common form has a body-centered cubic lattice, but it can also have a face-centered cubic lattice, which is, however, stable only at temperatures above 1200°C. Of course, at that temperature the body-centered cubic structure is already past the Curie point. However, by alloying chromium and nickel with the iron, two possible amounts of 18 percent chromium and 5 percent nickel, we can get what is called stainless steel, which, although it is mainly iron, retains the body-centered cubic even at low temperatures. Because its crystal structure is different, it has completely different magnetic properties. Most kinds of stainless steel are not magnetic to any appreciable degree, although there are some kinds which are somewhat magnetic—it depends on the composition of the alloy. Even when such an alloy is magnetized, it is *not* ferromagnetic like ordinary iron—even though it is mostly just iron.

We want to look now to close to a few of the special materials which have been developed for their particular magnetic properties. First, if we want to make a permanent magnet, we want the material with an extremely wide hysteresis loop so that, when we turn the current off and come down to zero magnetizing field, the magnetization will remain large. For such materials the domain boundaries should be "frozen" in place as much as possible. One such material is the remarkable alloy "Alnico V" (50% Fe, 20% Al, 12% Ni, 8% Cu, 2% Ti, 2% Cr). The rather complex composition of this alloy is because of the fact that it contains a bit that has gone into making good magnets. What happens is to take the two things together and heat them until you find the most ideal situation. When aluminum solidifies, there is a "second phase" which precipitates out, making many tiny cycles and very high internal strains. In this material, the domain boundaries have a hard time moving at all. In addition to having a precise composition, Alnico is microscopically "worked" in a way that causes the crystals appear in the form of long grains along the direction in which the magnetization is going to be. Then the magnetization will have a natural tendency to be lined up in these directions and will be fixed there from the annealing effects. Furthermore, the material is even cooled in an external magnetic field when it is manufactured, so that the grains will grow with the right crystal orientation. The hysteresis loop of Alnico V is shown in Fig. 37-12. You see that it's about 300 times wider than the hysteresis curve for iron, and that we showed in the last chapter in Fig. 36-8.

Let's turn now to a different kind of material. For inducting transformers and motors, we want a material which is magnetically "soft"—one in which the magnetism is easily changed so that an enormous amount of magnetization results from a very small applied field. For example, this we need pure, well-crucible materials which will have very low dislocations and impurities so that the domain

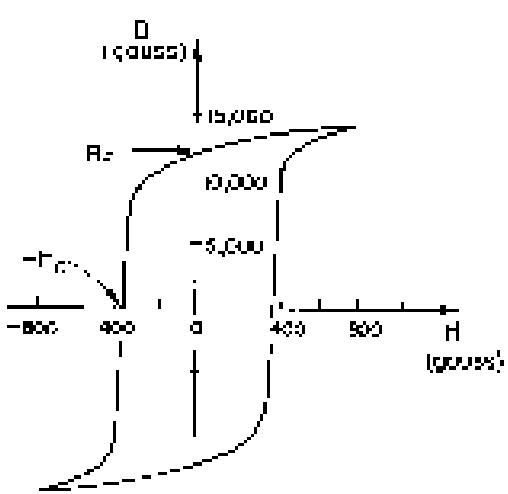


Fig. 37-12. The hysteresis curve of Alnico V.

walk can move easily. It would also be nice if we could make the anisotropy small. Then, even if a grain of the material sits at the wrong angle with respect to the field, it will still magnetize easily. Now we have said that iron prefers to magnetize along the [001] direction, whereas nickel prefers the [111] direction; so, if we mix iron and nickel in various proportions, we might hope to find that with just the right proportions the alloy wouldn't prefer any direction. The [100] and [111] directions would be equivalent. It turns out that this happens with a mixture of 70 percent nickel and 30 percent iron. In addition—possibly by luck or maybe because of some physical relationship between the anisotropy and the magnetostatic effects—it turns out that the magnetostatic of iron and nickel has the opposite sign. And in an alloy of the two metals, this property goes through zero at about 30 percent nickel. So somewhere between 30 and 50 percent nickel we get very "soft" magnetic materials—alloys that are very easy to magnetize. They are called the *permalloys*. Permalloys are useful for high-quality transformers (a few signal levels), but they would be no good at all for induction, magnets. Permalloys are very easily made and handled. The magnetic properties of a piece of permalloy are essentially unchanged if stressed beyond the elastic limit—*i.e.*, until it's bent. This shape stability is reduced because of the dislocations, slip bands, and so on, which are produced by the mechanical deformations. The domain boundaries are no longer very precise. The high permeability can, however, be destroyed by passing it at high temperatures.

It is often useful to have some numbers to characterize the various magnetic materials. One useful one is the intercept of the hysteresis loop with the B - and H -axes, as indicated in Fig. 37-12. These intercepts are called *coercive forces* (*field* B_c and *current* H_c). In Table 37-1 we list these numbers for a few magnetic materials.

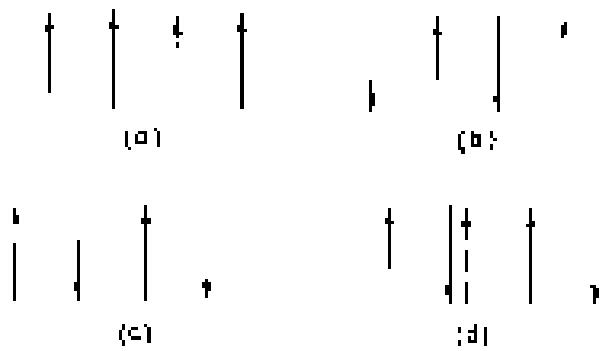


Fig. 37-12. Relative orientation of electron spins in various materials. (a) ferromagnetic, (b) anti-ferromagnetic, (c) ferrimagnetic, (d) chromium-alloy. (Remember always show direction of total angular momentum, including orbital motion.)

37-6. Extraordinary magnetic materials

We would now like to discuss some of the more exotic magnetic materials. There are many elements in the periodic table which have negative mass-electron shells and hence have negative magnetic moments. For instance, right next to Sc (Scandium) you have Cr, Mn, Co, Ni, Cu, and Fe. You will find chromium and manganese. Why aren't they ferromagnetic? The answer is that the term in Eq. (37.1) has the opposite sign for these elements. In the chromium lattice, for example, the spins of the chromium atoms alternate above and below, as shown in Fig. 37-13(d). So chromium is magnetized from the point of view, but it is not technically interesting because there are no *anomalous magnetic effects*. Chromium, then, is an example of a material in which quantum mechanical effects make it a spin insulator. Such a material is called *antiferromagnetic*. The "anomalous" in a ferromagnetic material is also temperature dependent. Below a critical temperature, all the spins are lined up in the alternating array, but when the material is heated above a certain temperature—which is again called the Curie temperature—the spins suddenly become random. There is, interestingly, a surface transition. This behavior can be seen in the specific heat curve. Also, it shows up in some spin-orbit effects. For instance, the existence of the alternating spins can be seen, for instance, in the neutron Bragg scattering of chromium. Because a neutron "isn't" has a spin,

Table 37-1

Properties of some ferromagnetic materials

Material	B_c Gauss (approx.)	H_c Oerst. (approx.)
Supermagnet	> 9000	0.03
Stainless steel (permeability)	12000	1.0
Alnico V	2400	0.6
Alnico V	12000	0.03

(of a magnetic moment), it has different amplitude to be generated, depending on whether its spin is parallel or opposite to the spin of the neighbor. Thus, we get a different influence per unit μ on the spins in a crystal according to how we are when they have σ or τ along a given direction.

There is another kind of interaction which quantum mechanical effects make the electron spins interact, but which is not ferromagnetic—that is, the crystal has a permanent net magnetization. The idea behind such materials is shown in Fig. 37-1. The figure shows the crystal structure of spinel, a magnesium-aluminum oxide which—as it is shown—is nonmagnetic. The oxide has two kinds of metal atoms: magnesium and aluminum. Now if we replace the nonmagnetic ones by two magnetic elements like iron and zinc, or by zinc and manganese—in other words, two paramagnetic atoms instead of the nonmagnetic ones—an interesting thing happens. Let's call one kind of metal atom a and the other kind of metal atom b ; then the following combination of forces is in the crystal. There is an a - a interaction which tries to make the a 's align, and the a 's have opposite spins because quantum mechanics always gives the opposite sign (except for the mysterious crystals of iron, nickel, and cobalt). Then there is a direct a - b interaction which tries to make the a 's opposite, and also a b - b interaction which tries to make the b 's opposite. Now, of course we cannot have everything opposite; everything else— a opposite b , a opposite a , and b opposite b . Presumably because of the distances between the a 's and the presence of the oxygen (which we really don't know well), it turns out that the a - a interaction is stronger than the a - b or the b - b . So the solution that nature uses in this case is to make all the a 's parallel to each other, and all the b 's parallel to each other, but the two systems opposite. That gives the lowest energy by use of the stronger a - a interaction. The result is all the a 's are spinning in one all-use θ^+ and spinning over—at very small, of course. But in the nonmagnetic case of the regular spinel and the v -type spinel are no θ^+ , we can get the situation shown in Fig. 37-1(a), and there will be a net magnetization in the material. The materials will then be ferromagnetic—although somewhat weak. Such materials are called ferrites. They do not have as high a saturation magnetization as iron, for example, because \rightarrow they are only useful for smaller fields. But they have a very important feature—they are insulators, the b 's are not ferromagnetic conductors. In large magnetic fields, they will have very small eddy currents and so can be used, for example, in microwave systems. The microwave fields will be able to get into such insulating material, whereas they would be kept out by the eddy currents in a conductor like iron.

There is another class of magnetic materials which has only recently been discovered—members of the family of the orthosilicates of yttrium. They are again crystals in which the lattice contains two kinds of metallic atoms, and we have again a situation in which two kinds of atoms can be substituted almost at will. Among the many compounds of interest there is one which is completely ferromagnetic. It has yttrium and iron in the garnet structure, and the reason it is ferromagnetic is very curious—it's again quantum mechanics: making the neighboring spins opposite so that there is a locked-in system of spins with the electron spins of the iron one way and the electron spins of the yttrium the opposite way. But the yttrium atom is complicated. It's a rare earth element and gets a large contribution to its magnetic moment from orbital motion of the electrons. For yttrium, the orbital motion contribution is opposite that of the spin and also is bigger. Thus, although quantum mechanics, working through the exclusion principle, makes the spins of the yttrium opposite those of the iron, it makes the total magnetic moment of the yttrium atom *parallel* to the iron because of the orbital effect—as sketched in Fig. 37-2(d). The compounds therefore a regular ferrimagnet.

Another interesting example of ferrimagnetism occurs in some of the rare-earth elements. It has to do with a still more peculiar arrangement of the spins. The material is not ferromagnetic in the sense that the spins are all parallel, nor is it antiferromagnetic in the sense that every site is opposite. In these crystals all of the spins of one layer are parallel and lie in a plane of the layer. In the next

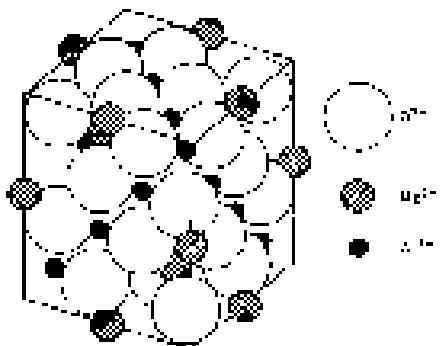


Fig. 37-1. Crystal structure of the mineral spinel ($MgAl_2O_4$), the Mg^{2+} ions occupy tetrahedral sites, and are surrounded by four oxygen ions; the Al^{3+} ions occupy octahedral sites, each surrounded by six oxygen ions. (From Charles Kittel, Introduction to Solid State Physics, John Wiley and Sons, Inc., New York, 2nd ed., 1956.)

layer all spins are again parallel to each other, but point in a somewhat different direction. In the following layer they are in still another direction, and so on. The result is that the local magnetization vector varies in the form of a spiral—the magnetic moments of the successive layers rotate as we proceed along a line perpendicular to the layers. It is interesting to try to analyze what happens when a field is applied to such a spiral—all the twisting and turning that must go on in all these various magnets. (Some people like to amuse themselves with the theory of these things.) Not only are there cases of "flat" spirals, but there are also cases in which the directions the magnetic moments of successive layers map out a cone, so that it has a spiral component and also a uniform ferromagnetic component at every location!

The magnetic properties of materials worked out on a more advanced level than we have been able to do here, have fascinated physicists of all kinds. In the first place, there are those practical people who love to work out ways of making things in a better way—they love to design better and more interesting magnetic materials. The discovery of things like ferrites, or their application, immediately delights people who like to see clever new ways of doing things. Besides this, there are those who find a fascination in the terrible complexity that nature can produce using a few basic laws. Starting with one and the same general idea, nature goes from the ferromagnetism of iron and its alloys, to the antiferromagnetic state of chromium, to the magnetism of ferrites and garnets, to the spiral structure of the rare earth elements, and on, and on. It is fascinating to discover experimentally a lot of strange things that go on in these special substances. Then, to the theoretical physicists, this complex situation presents a number of very interesting, unusual, and beautiful challenges. One challenge is to understand why it exists at all. Another is to predict the statistics of the interacting spins in a solid lattice. Even neglecting any possible exchanges or spin-orbits, the problem has, as far, defied full understanding. The reason that it is so interesting is that it is such an easily solved problem: Given a lot of electron spins in a regular lattice, interacting with such-and-such a law, what do they do? It is so simple stated, but it has defied complete analysis for years. Although it has been analyzed rather carefully for temperatures not too close to the Curie point, the theory of the sudden transition to the Curie point still needs to be completed.

Finally, the whole subject of the system of spinning atomic magnets—in ferromagnetic, or in paramagnetic materials and in nuclear magnetic, has also been a fascinating thing to advanced students in physics. The system of spins can be pushed on and pulled on with external magnetic fields, so one can do things like with resonances, with rotation effects, with spin-orbits, and with other effects. It serves as a prototype of many complicated thermodynamic systems. But in paramagnetic materials the situation is often fairly simple, and people have been delighted both in the experiments and in explaining the phenomena theoretically.

We now close our study of electricity and magnetism. In the first chapter, we spoke of the great strides that have been made since the early Greek observation of the strange behaviors of amber and of lodestone. Yet in all our long and involved discussion we have never explained why it is that when we rub a piece of amber we get a charge on it nor have we explained why a lodestone is magnetic at all. You may say, "Oh, we just didn't get the right sign." No, it is worse than that. Even if we did get the right sign, we would still have the question: Why is the piece of lodestone in the ground magnetized? There is the earth's magnetic field, of course, but where does the earth's field come from? Nobody really knows—there have only been some good guesses. So you see, the physics of ours is a lot of liberty—we start out with the phenomena of lodestone and amber, and we end up not understanding either of them very well. But we have learned a tremendous amount of very exciting and very practical information in the process!

Kinematics

38-1 Hooke's law

The subject of elasticity deals with the behavior of those substances which have the property of recovering their size and shape when the forces producing deformations are removed. We find the elastic properties to some extent in all solid bodies. If we had the time to deal with them at length, we would want to look into many things: the behavior of materials, the general laws of elasticity, the general theory of elasticity, the atomic mechanisms that determine the elastic properties, and finally the limitations of elastic laws when the forces become so great that plastic flow and fracture occur. It would take more time than we have to cover all these subjects in detail, so we will have to leave out some things. For example, we will not discuss plasticity or the limitations of the elastic laws. (We touched on these subjects briefly when we were talking about dislocations in metals.) Also, we will not be able to discuss the internal mechanisms of elasticity—so our treatment will not have the completeness we have tried to achieve in the earlier chapters. Our aim is mainly to give you an acquaintance with some of the ways of dealing with such practical problems as the bending of beams.

When you pull on a piece of material, if "never," the material is unextended. If the force is small enough, the slight expansion rate of the various points in the material are proportional to the force—we say the behavior is *elastic*. We will discuss only the elastic behavior. First, we will write down the mathematical law of elasticity, and then we will apply it to a number of different situations.

Suppose we have a rectangular block of material of length l , width w , and height h , as shown in Fig. 38-1. If we pull on the ends with a force F , i.e., the length increases by an amount Δl , we will suppose in all cases that the change in length is a small fraction of the original length. As a matter of fact, for materials like wood and steel, the material will break if the change in length is more than a few percent of the original length. For a large number of materials, especially as shown in Fig. 38-1, for sufficiently small extensions the force is proportional to the extension:

$$F \propto \Delta l \quad (38.1)$$

This is often known as Hooke's law.

The length change Δl of the bar will also depend on its length. We can figure out how by the following argument. If we hang two identical blocks together end to end, the stress increases in each block, and will stretch by Δl . Thus, the stretch of a block of length $2l$ will be twice as big as a block of the same cross section, $w \times h$, of length l . In order to get a more complete comprehension of the problem, unless of course you do the steps, we can associate with the Δl of (38.1) the length to the original length. This ratio is proportional to the force but independent of it:

$$\frac{\Delta l}{l} \propto F. \quad (38.2)$$

The force F will also depend on the area of the block. Suppose that we put two blocks side by side. Then for a given stretch Δl we would have the force F on each block, or twice as much on the combination of the two blocks. The force, for a given amount of stretch, must be proportional to the cross-sectional area A of the block. To obtain a law in which the coefficient of proportionality is independent of the dimensions of the body, we write Hooke's law for a rectangular

38-1 Hooke's law

38-2 Uniform stretching

38-3 The tension half-shear waves

38-4 The heat waves

38-5 Buckling

Review: Chapter 47, Vol. 1, showed:
the Hooke's law.

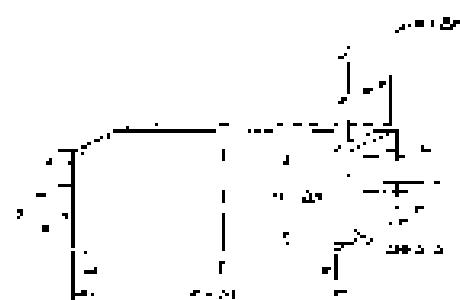


Fig. 38-1. The stretching of a horizontal deformation.

block in the form

$$\sigma = E \frac{\delta l}{l} \quad (38.3)$$

The constant E is a property only of the nature of the material; it is known as Young's modulus. (Usually you will see Young's modulus called E . But we've used E for electric fields, energy, and emf's, so we prefer to use a different letter.)

The force per unit area is called the stress, and the stretch per unit length—i.e., fractional stretch—is called the strain. Equation (38.3) can therefore be rewritten in the following way:

$$\frac{F}{A} = \sigma \times \frac{\delta l}{l} \quad (38.4)$$

$$\text{Stress} = (\text{Young's modulus}) \times (\text{Strain})$$

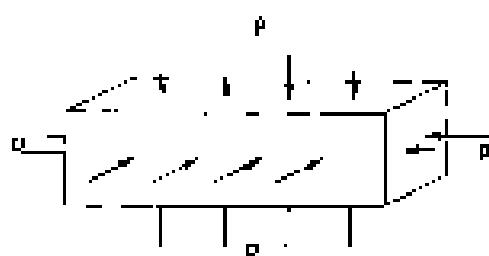


Fig. 38-2 A block under $w/4$ atm hydrostatic pressure.

There is another part to Hooke's law: When you stretch a block of material in one direction it contracts at right angles to the stretch. The contraction in width is proportional to the width w and also to $\delta l/l$. The sideways contraction is in the same proportion for both width and height, and is usually written

$$\frac{\Delta w}{w} = \frac{\Delta h}{h} = -\alpha \frac{\delta l}{l} \quad (38.5)$$

where the constant α is another property of the material called Poisson's ratio. It is always positive in sign and is a number less than 1/2. (It is "reasonable" that α should be generally positive, but it is not quite clear what is meant by that.)

The two constants E and α specify completely the elastic properties of a homogeneous isotropic (that is, noncrystalline) material. In crystal-like materials the stretches and contractions can be different in different directions, so there can be many more elastic constants. We'll return our discussion temporarily to homogeneous/anisotropic materials whose properties can be described by E and α . As usual there are different ways of describing things; some people like to describe the elastic properties of materials by different constants. It always takes two, and they can be related to E and α .

The last general law we need is the principle of superposition. Since the two laws (38.1) and (38.3) are linear in the forces and in the displacements, superposition will work. If you have one set of forces and get some displacements, and then you add a new set of forces and get some additional displacements, the resulting displacements will be the sum of the ones you would get with the two sets of forces acting independently.

Now we have all the general principles—the superposition principle and Eqs. (38.1) and (38.3)—and that's all there is to elasticity. But that is like saying that once you have Newton's laws that's all there is to mechanics. Or, given Maxwell's equations, that's all there is to electricity. It is, of course, true that with these principles you have a great deal, because with your present mathematical ability you could go a long way. We will, however, work out a few special applications.

38-2 Uniform strains

As our first example let's find out what happens to a rectangular block under uniform hydrostatic pressure. Let's put a block under water in a pressure tank. Then there will be a force acting inward on every face of the block proportional to the area (see Fig. 38-2). Since the hydrostatic pressure is uniform, the stress (force per unit area) on each face of the block is the same. We will work out first the change in the length. The change in length of the block can be imagined as the sum of changes in length that would occur in the three independent problems which are sketched in Fig. 38-4.



Fig. 38-3 Hydrostatic pressure is the superposition of three independent compressions.

Problem 1. If we push on the ends of the block with a pressure p , the compressional strain is $\frac{\Delta l}{l} = -\frac{p}{Y}$ and it is negative,

$$\frac{\Delta l}{l} = -\frac{p}{Y}.$$

Problem 2. If we push on the two sides of the block with pressure p , the compressional strain is again $\frac{\Delta l}{l} = -\frac{p}{Y}$, but now we want the lengthwise strain. We can get that from the sideways strain multiplied by $-w$. The sideways strain is

$$\frac{\Delta w}{w} = -\frac{p}{E};$$

so

$$\frac{\Delta l}{l} = -w \frac{p}{E}.$$

Problem 3. If we push out the top of the block, the compressional strain is once more $\frac{\Delta l}{l} = -\frac{p}{Y}$, and the corresponding strain in the sideways direction is again $\frac{\Delta w}{w} = -\frac{p}{E}$. We get

$$\frac{\Delta l}{l} = -w \frac{p}{E}$$

Combining the results of the three problems—that is, taking $\Delta l = \Delta l_1 + \Delta l_2 + \Delta l_3$ —we get

$$\frac{\Delta l}{l} = \frac{p}{Y}(1 - 2w). \quad (36.6)$$

The problem is, of course, symmetric in all three directions; it follows that

$$\frac{\Delta v}{w} = \frac{\Delta h}{h} = -\frac{p}{Y}(1 - 2w). \quad (36.7)$$

The change in the volume under hydrostatic pressure is also of some interest. Since $V = lwh$, we can write, for small displacements,

$$\frac{\Delta V}{V} = \frac{\Delta l}{l} + \frac{\Delta w}{w} + \frac{\Delta h}{h}.$$

Using (36.6) and (36.7), we have

$$\frac{\Delta V}{V} = -2 \frac{p}{Y}(1 - 2w). \quad (36.8)$$

People like to call $\Delta V/V$ the volume strain and write

$$\nu = -K \frac{\Delta V}{V}.$$

The volume strain ν is proportional to the volume strain—Hooke's law once more. The effective K is called the bulk modulus; it is related to the other constants by

$$K = \frac{Y}{(1 + 2w)}. \quad (36.9)$$

Since K is of some practical interest, many textbooks give Y and K instead of Y and ν . If you want ν you can always get it from Eq. (36.9). We can also see from Eq. (36.9) that Poisson's ratio, w , must be less than one-half. If w were not, the bulk modulus K would be negative, and the material would expand under increasing pressure. That would allow us to get mechanical energy out of any old block—it would mean that the block was in unstable equilibrium. If it started to expand it would continue by itself until it release of energy.

Now we want to consider what happens when you put a "shear" stress on something. By shear strain we mean the kind of distortion shown in Fig. 36-4. As is preliminary to this, let's look at the strains in a cube of material subjected to the forces shown in Fig. 36-5. Again we can break it up into two problems: the vertical

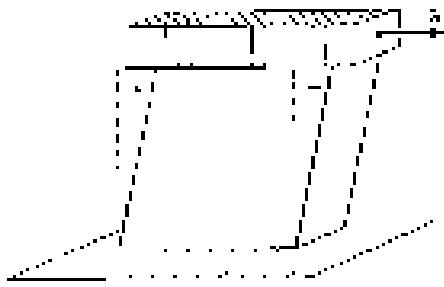


Fig. 36-4. A cube in uniform shear.

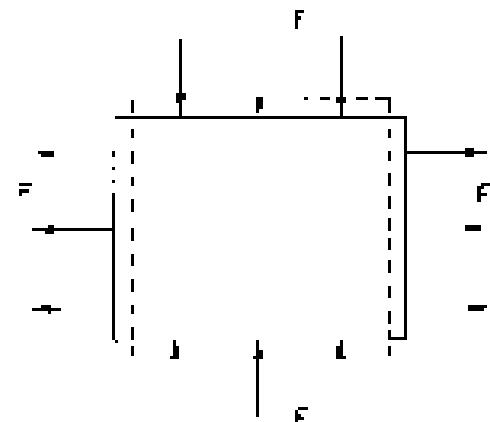


Fig. 36-5. A cube with compressing forces on top and bottom and equating-stressing forces on two sides.

pulses, and the horizontal pull. Calling A the area of the cube face, we have for the change in horizontal length

$$\frac{\Delta l}{l} = \frac{1}{Y} F - \sigma \frac{1}{Y} F = \frac{1 + \sigma}{Y} \frac{F}{A}. \quad (38.10)$$

The change in the vertical height is just the negative of this.

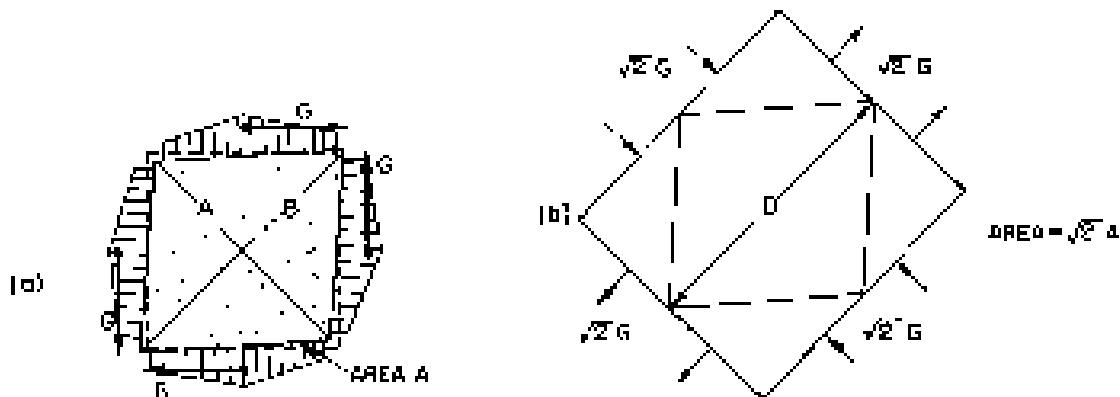


Fig. 38-5. The two pairs of shear forces in (a) produce the same stress as the compressing and stretching forces of (b).

Now suppose we take the same cube and subject it to the shearing forces shown in Fig. 38-6(a). Note that all the forces have to be equal if there are to be no net torques and the cube is to be in equilibrium. (Similar forces must also exist in Fig. 38-4, since the block is in equilibrium.) They are passed through the "glue" that holds the block to the table. (The cube is then said to be in a state of pure shear.) But note that if we cut the cube by a plane at 45°, say along the diagonal d in the figure, the total force acting across the plane is normal to plane d , and is equal to $\sqrt{2}G$. The area over which this force acts is $\sqrt{2}/4$, therefore, the tensile stress normal to this plane is simply $G/\sqrt{2}$. Similarly, if we examine a plane at an angle of 45° the other way—the diagonal d' in the figure—we see that there is a compressional stress normal to this plane of $-G/\sqrt{2}$. From this, we see that the stress in a "pure shear" is equivalent to a combination of tension and compression stresses of equal strength and at right angles to each other, and at 45° to the original faces of the cube. The internal stresses and strains are the same as we would find in the larger block of material with the forces shown in Fig. 38-6(b). (This is the problem we have already solved. The change in length of the diagonal is given by Eq. (38.10).)

$$\frac{\Delta l}{l} = \frac{1 + \sigma}{Y} \frac{F}{A}. \quad (38.11)$$

(One diagonal is elongated, the other is elongated.)

It is often convenient to express a shear strain in terms of the angle by which the cube is twisted—the angle θ in Fig. 38-7. From the geometry of the figure you can see that the longitudinal shift, δ , of the top edge is equal to $\sqrt{1 + \theta^2} - 1$. So

$$\theta = \frac{\delta}{l} = \frac{\sqrt{1 + \theta^2} - 1}{l} = \frac{\sqrt{2}\Delta D}{l} = \frac{2\Delta D}{l}. \quad (38.12)$$

The shear stress σ is defined as the tangential force on one face divided by the area, $\sigma = F/A$. Using Eq. (38.10) in (38.12), we get

$$\theta = k + \frac{\sigma}{Y} l. \quad (38.13)$$

Fig. 38-7. The shear strain θ is $\frac{2\Delta D}{l}$.

(It is useful to remember "shear = constant times strain.")

$$\sigma = \mu k. \quad (38.13)$$

The proportionality coefficient ν is called the shear modulus (or, sometimes, the coefficient of rigidity). It is given in terms of E and σ by

$$\nu = \frac{E}{2(1+\nu)}. \quad (38.14)$$

Obviously, the shear modulus must be positive—otherwise you could get work out of a self-stressing block. From Eq. (38.14), ν must be greater than -1 . We know, then, that ν must be between -1 and $-\frac{1}{2}$; in practice, however, it is always greater than zero.

As a first example of the type of situation where the stresses are uniform through the material, let's consider the problem of a block which is stretched, while it is at the same time constrained so that no lateral contraction can take place. (Technically, it's a little easier to reciprocate it while keeping the sides from bulging out—but it's the same problem.) What happens? Well, there must be sideways forces which keep it from changing its thickness—forces we don't know off-hand but we have to calculate. It's the same kind of problem we have already done, only with a little different algebra. We compute forces on all three sides, as shown in Fig. 38-8; we calculate the changes in dimensions, and we choose the necessary forces to make the width and height equal again. Following the usual arguments, we get for the three strains:

$$\frac{\Delta L_x}{L} = \frac{1}{Y} \frac{F_x}{A_x} - \frac{\sigma}{Y} \frac{F_y}{A_y} - \frac{\sigma}{Y} \frac{F_z}{A_z} = \frac{1}{Y} \left[\frac{F_x}{A_x} - \sigma \left(\frac{F_y}{A_y} + \frac{F_z}{A_z} \right) \right], \quad (38.15)$$

$$\frac{\Delta L_y}{L} = \frac{1}{Y} \left[\frac{F_y}{A_y} - \sigma \left(\frac{F_x}{A_x} + \frac{F_z}{A_z} \right) \right]. \quad (38.16)$$

$$\frac{\Delta L_z}{L} = \frac{1}{Y} \left[\frac{F_z}{A_z} - \sigma \left(\frac{F_x}{A_x} + \frac{F_y}{A_y} \right) \right]. \quad (38.17)$$

Now since ΔL_x and ΔL_z are supposed to be zero, Eqs. (38.16) and (38.17) give two equations relating F_y and F_z to F_x . Solving them together, we get that

$$\frac{F_y}{A_y} = \frac{F_z}{A_z} = \frac{\sigma}{1-\sigma} \frac{F_x}{A_x}. \quad (38.18)$$

Substituting in (38.15), we have

$$\frac{\Delta L_x}{L} = \frac{1}{Y} \left(1 - \frac{\sigma^2}{1-\sigma} \right) \frac{F_x}{A_x} = \frac{1}{Y} \left(\frac{1-\sigma}{1-\sigma^2} \right) \frac{F_x}{A_x}. \quad (38.19)$$

(Once you will see this turned around, and with the quantities in a factored out, it is then written

$$\frac{F_x}{A_x} = \frac{1-\sigma}{(1-\sigma)^2 - 2\sigma} Y \frac{\Delta L}{L}. \quad (38.20)$$

When we constrain the sides, Young's modulus gets multiplied by a complicated function of σ . As you can now easily see from Eq. (38.19), the factor in front of Y is always greater than 1. It is harder to stretch the block when the sides are held, which also means that a block is stronger when the sides are held than when they are not.

38-3 The torsion bar; shear waves

Let's now turn our attention to an example which is more complicated because different parts of the material are stressed by different amounts. We consider a twisted spring, as you would find in a drive shaft of some machinery, or in a spiral telephone suspension used in a telephone instrument. As you probably know from experiments with the torsion pendulum, the rotation of a twisted rod is proportional to the torque. The constant of proportionality obviously depends upon the length L , the radius r , the radius of the rod, and on the properties of the material. The question is "In what way?" We are now in a position to answer this question: it's just a matter of working out some geometry.

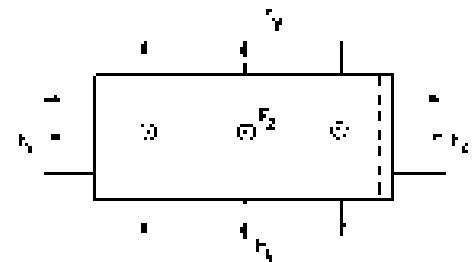


Fig. 38-8. Stretching without lateral contraction.

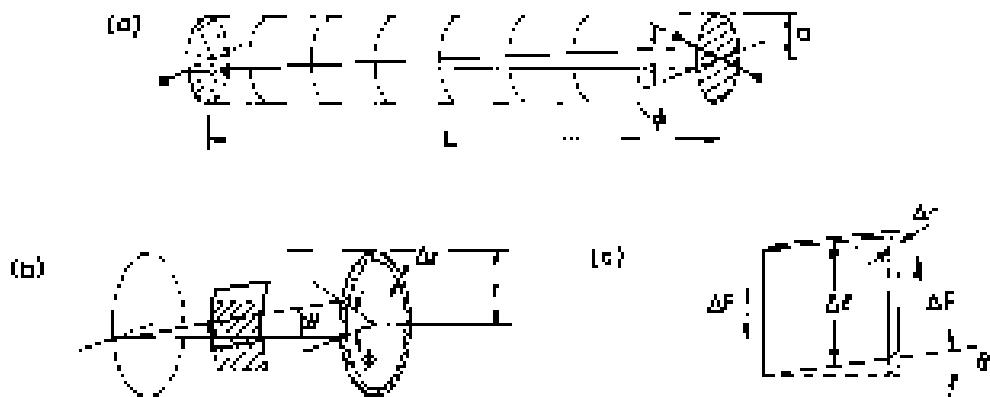


Fig. 19-9. (a) A cylindrical bar in torsion. (b) A cylindrical shell in torsion.
(c) Each small piece of the shell is *r* sheared.

Fig. 19-9(c) shows a cylindrical segment of length $\Delta\ell$ and radius r , with one end twisted by the angle $\Delta\theta$ with respect to the other. If we want to relate the stresses to what we already know, we can think of the rod as being made up of many cylindrical shells and we know separately what happens to each shell. We start by looking at a thin, short cylinder of radius r less than $\Delta\ell$ around the axis of rotation, as shown in Fig. 19-9(b). Now if we look at a piece of this cylinder that was originally a small square, we see that it has been distorted into a parallelogram. Each such element of the cylinder is a *shear*, and the shear angle is

$$\delta = \frac{\Delta\theta}{\ell}.$$

The shear stress τ in the material is, therefore (from Eq. (38.10)),

$$\tau = \mu\delta = \mu \frac{\Delta\theta}{\ell}. \quad (38.21)$$

The shear stress is the tangential force ΔF on the end of the square divided by the area ΔA of the end [see Fig. 19-9(c)].

$$\tau = \frac{\Delta F}{\Delta A}.$$

The force ΔF on the end of such a square contributes a torque $\Delta\tau$ around the axis of the rod equal to

$$\Delta\tau = r \Delta F = r \tau \Delta A. \quad (38.22)$$

The total torque τ is the sum of such torques around the entire diameter of the cylinder. By putting together enough slices as in the disk adding in Fig. 2.6, we find that the total torque for a hollow tube is

$$(\pi r^2 \tau) \Delta r. \quad (38.23)$$

Or, using (38.21),

$$\tau = 2\pi r \frac{r^2 \Delta\theta}{L}. \quad (38.24)$$

We saw that the rotational surface, r/ϕ , of a hollow tube is proportional to the cube of the radius r and to the thickness Δr , and inversely proportional to the length L .

We can now imagine a solid rod to be made up of a series of concentric tubes, each twisted by the same angle $\Delta\theta$ although the internal stresses are different for each tube. The total torque is the sum of the torques required to rotate each solid; for the solid rod

$$\tau = 2\pi r \frac{r^2}{L} \int r^2 dr.$$

where the integral goes from $r = 0$ to $r = a$, the radius of the rod. Integrating, we have

$$\tau = \rho \frac{\partial \phi}{\partial L} \cdot \phi. \quad (38.25)$$

For a rod in torsion, the torque is proportional to the angle and is proportional to the fourth power of the diameter—a rod twice as thick is sixteen times as stiff for torsion.

Before leaving the subject of torsion, let us apply what we have just learned to an interesting problem: torsional waves. If you take a long rod and suddenly twist one end, a wave of twist works its way along the rod, as sketched in Fig. 38-13(a). That's a little more exciting than a steady twist—let's see what happens.

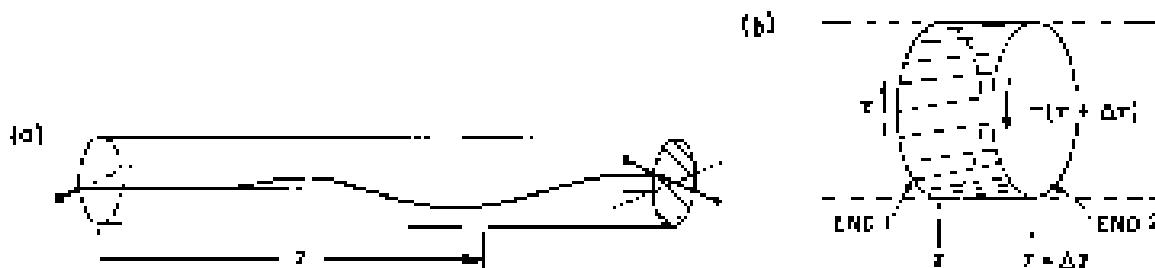


Fig. 38-13. (a) A torsional wave on a rod. (b) A volume element of the rod.

Let z be the distance to some point along the rod. For a static torsion, the torque is the same everywhere along the rod, and is proportional to $\rho \tau / L$, the total torsion angle over the rod length. What matters to the material is the local torsional strain, which is, you will appreciate, $\Delta\phi/L$. When the torsion along the rod is not uniform, we should replace Eq. (38.25) by

$$\tau(z) = \rho \frac{\pi a^4}{2} \frac{\partial \phi}{\partial z}. \quad (38.26)$$

Now let's look at what happens to an element of length Δz shown magnified in Fig. 38-13(b). There is a torque $\tau(z)$ at end 1 of the little hunk of rod, and a different torque $\tau(z + \Delta z)$ at end 2. If Δz is small enough, we can use a Taylor expansion and write

$$\tau(z + \Delta z) = \tau(z) + \left(\frac{\partial \tau}{\partial z} \right)_{z=z} \Delta z. \quad (38.27)$$

The net torque $\Delta\tau$ acting on the little piece of rod between z and $z + \Delta z$ is clearly the difference between $\tau(z)$ and $\tau(z + \Delta z)$, or $\Delta\tau = (\partial\tau/\partial z) \Delta z$. Differentiating Eq. (38.26), we get

$$\Delta\tau = \rho \frac{\pi a^4}{2} \frac{\partial^2 \phi}{\partial z^2} \Delta z. \quad (38.28)$$

The effect of this net torque is to give an angular acceleration to the little slice of the rod. The mass of the slice is

$$\Delta M = (\rho a^2 \Delta z) a,$$

where a is the density of the material. We worked out in Chapter 10, Vol. 1, that the moment of inertia of a circular cylinder is $\pi a^4/4$; calling the moment of inertia of our piece ΔI , we have

$$\Delta I = \frac{\pi}{4} \rho a^4 \Delta z. \quad (38.29)$$

Newton's law says the torque is equal to the moment of inertia times the angular acceleration, or

$$\Delta\tau = \Delta I \frac{\partial^2 \phi}{\partial z^2}. \quad (38.30)$$

Pulling everything together, we get

$$\begin{aligned} \frac{\pi a^2 \rho^2 c^2}{2} \frac{\partial^2 \phi}{\partial r^2} \Delta r &= \frac{\pi}{2} \rho a^4 \Delta r \frac{\partial^2 \phi}{\partial t^2}, \\ \text{or} \quad \frac{\partial^2 \phi}{\partial r^2} - \frac{\rho a^2 \partial^2 \phi}{\mu c^2 \partial t^2} &= 0 \end{aligned} \quad (38.31)$$

You will recognize this as the one-dimensional wave equation. We have found that waves of motion will propagate down the rod with the speed

$$C_{\text{shear}} = \sqrt{\frac{E}{\rho}}. \quad (38.32)$$

The *stiffer* the rod—*for the same stiffness*—the *faster* the waves; and the *stiffer* the rod, the quicker the waves work their way down. The speed does not depend upon the diameter of the rod.

Torsional waves are a special example of shear waves. In general, shear waves are those in which the strains do not change the volume of any part of the material. In torsional waves, we have a permanent distribution of such shear stresses—namely, distributed on a circle. But for any arrangement of shear stresses, waves will propagate with the same speed—the one given in Eq. (38.32). (For example, the seismologists find such shear waves travelling in the interior of the earth.)

We can have another kind of a wave in the elastic world inside a solid material. If you push something, you can start "longitudinal" waves—also called "compressional" waves. They are like the sound waves in air or in water—the displacements are in the same direction as the wave propagation. (At the surfaces of an elastic body there can also be other types of waves, called "Rayleigh waves" or "Love waves." In them, the strains are neither purely longitudinal nor purely transverse. We will not have time to study them.)

While we're on the subject of waves, what is the velocity of the pure compression waves in a dog, in a bird, like the earth? We say "large" because the speed of sound in a thick body is different from what it is, for instance, along a thin rod. By a "thick" body we mean one in which the transverse dimensions are much larger than the wavelength of the sound. Then, when we push on the object, it cannot expand sideways—it can only compress in one dimension. Fortunately, we have already worked out the special case of the compression of a constrained elastic material. We have also worked out in Chapter 17, Vol. I, the speed of sound waves in a gas. Following the same argument, you can see that the speed of sound in a solid is equal to $\sqrt{Y/\rho}$, where Y is the "longitudinal modulus"—or pressure divided by the relative change in length. (For the constrained case, this is just the ratio of E/l to ρ/l ; we got it in Eq. (38.20).) So the speed of the longitudinal waves is given by

$$C_{\text{long}} = \sqrt{\frac{Y}{\rho}} = \sqrt{\frac{1-\sigma}{(1+\sigma)(1+2\sigma)}} \frac{E}{\rho}. \quad (38.33)$$

So long as σ is between zero and $1/2$, the shear modulus ρ is less than Young's modulus Y , and also, Y is greater than ρ , so:

$$2 < Y < 12.$$

This means that longitudinal waves travel faster than shear waves. One of the most precise ways of measuring the elastic constants of a substance is by measuring the density ρ of the material and the speeds of the two kinds of waves. From this information we can get both Y and ρ . It is, incidentally, by measuring the difference in the arrival times of the two kinds of waves from an explosion that a seismologist can estimate—even from the signal at only one station—the distance to the quake.

38-4 The bent beam

We want now to look at another practical situation—the bending of a rod or a beam. What are the forces when we bend a bar of some arbitrary cross section? We will work it out thinking of a bar with a circular cross section, but our answer will be good for any shape. To see this, however, we will cut some corners, so our theory will work out only approximately. Our results will be correct, only when the radius of the bend is much larger than the thickness of the beam.

Suppose you grab the two ends of a straight bar and bend it into some curve like the one shown in Fig. 38-11. What goes on inside the bar? Well, it is curved, that means that the material on the inside of the curve is compressed and the material along the outside is stretched. There is some surface which goes along more or less parallel to the axis of the bar that is neither stretched nor compressed. This is called the *neutral surface*. You would expect this surface to be near the "middle" of the cross section. It can be shown (but we won't do it here) that, for small bending of a simple beam, the neutral surface goes through the "center of gravity" of the cross section. This is true only for "plane" bending—if you are not stretching or compressing the beam at the same time.

For pure bending, then, a thin transverse slice of the bar is distorted as shown in Fig. 38-12(a). The material below the neutral surface has a compressional strain which is proportional to the distance from the neutral surface; and the material above is stretched, as its position is farther from the neutral surface. So the longitudinal stretch is proportional to the height y . The constant of proportionality is just $1/R$, the radius of curvature of the bar (see Fig. 38-12):

$$\frac{dy}{y} = \frac{\beta}{R}$$

So the force per unit area—the stress—in a small strip at y is also proportional to the distance from the neutral surface:

$$\frac{\Delta F}{\Delta A} = \frac{F}{R}, \quad (38.34)$$

Now let's look at the forces that would produce such a strain. The forces acting on the little segment drawn in Fig. 38-12 are shown in the figure. If we think of any transverse cut, the forces acting across it are one way above the neutral surface and the other way below. They come in pairs to make a "bending moment" M —by which we mean the torque about the neutral line. We can compute the total moment by integrating the force times the distance from the neutral surface for one of the strips of the segment of Fig. 38-12:

$$M = \int_{-R/2}^{R/2} y dF. \quad (38.35)$$

From Eq. (38.34), $dF = YI/R dA$, so

$$M = \frac{YI}{R} \int_{-R/2}^{R/2} y^2 dA.$$

The integral of $y^2 dA$ is what we shall call the "moment of inertia" of the geometric cross section about a horizontal axis through its "center of gravity," we will call it I :

$$I = \frac{YI}{R}. \quad (38.36)$$

$$I = \int r^2 dA. \quad (38.37)$$

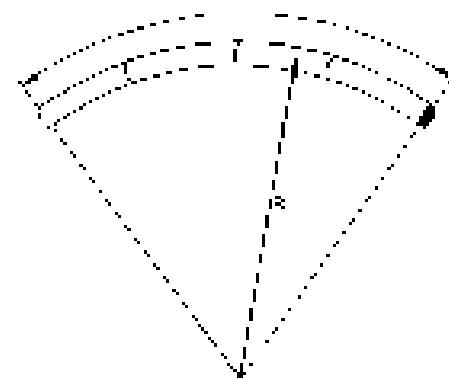


Fig. 38-11. A bent beam.

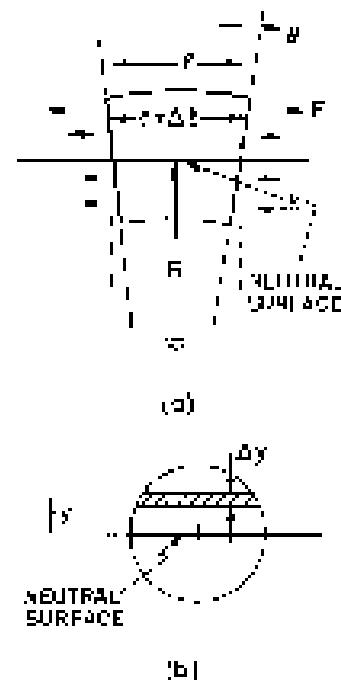


Fig. 38-12. (a) Small segment of a bent beam. (b) Cross section of the beam.

* It is, of course, really the moment of inertia of a beam with width Δx , mass $\rho \Delta x$, and a radius of gyration R .

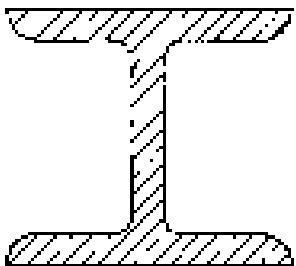


Fig. 38-13. An "I" beam.

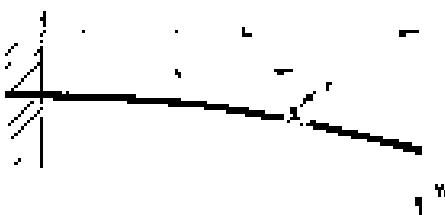


Fig. 38-14. A cantilevered beam with a weight at one end.

Equation (38.36), then, gives us the relation between the bending moment M and the curvature $1/R$ of the beam. The "stiffness" of the beam is proportional to E and to the moment of inertia I . In other words, if you want the stiffer possible beam with a given amount of, say, aluminum, you want to put as much of it as possible as far as you can from the neutral surface, to make a larger moment of inertia. You can't carry this to an extreme, however, because then the thing will not curve as we have supposed—it will buckle or twist and become weaker again. But now you see why structural beams are made in the form of an I or an H, as shown in Fig. 38-13.

As an example of the use of our beam equation (38.36), let's work out the deflection of a concentrated beam with a concentrated force W acting at the free end, as sketched in Fig. 38-14. (By "cantilevered" we simply mean that the beam is supported in such a way that both the position and the slope are fixed at one end—let's call it the center, wall.) What is the shape of the beam? Let's call the deflection at the distance x from the fixed end ζ ; we want to know $\zeta(x)$. We'll work it out only for small deflections. We will also assume that the beam is long in comparison with its cross section. Now, as you know from your mathematics courses, the curvature $1/R$ of any curve $\zeta(x)$ is given by

$$\frac{1}{R} = \frac{\zeta''(x)/dx^2}{1 - (\zeta'(x)/dx)^2} \quad (38.35)$$

Since we are interested only in small slopes (this is usually the case in engineering structures), we neglect $(\zeta'(x)/dx)^2$ in comparison with 1, and take

$$\frac{1}{R} = \frac{\zeta''(x)}{\zeta'(x)} \quad (38.36)$$

We also need to know the bending moment M . It is a function of x because it is equal to the torque about the neutral axis of any cross section. Let's neglect the weight of the beam and take only the downward force W at the end of the beam. (You can put in the beam weight yourself if you want.) Then the bending moment at x is

$$M(x) = W(L - x),$$

because that is the torque about the point x exerted by the weight W —the torque which the beam resists, support of x . We get

$$W(L - x) = \frac{M}{R} = \frac{M}{I} \frac{d^2\zeta}{dx^2}$$

or

$$\frac{d^2\zeta}{dx^2} = \frac{W}{I} (L - x). \quad (38.40)$$

This one we can integrate without any tricks; we get

$$\zeta = \frac{W}{I} \left(\frac{Lx^2}{2} - \frac{x^3}{3} \right), \quad (38.41)$$

using our assumptions that $\zeta(0) = 0$ and that $d\zeta/dx$ is also zero at $x = 0$. That is the shape of the beam. The displacement at the end is

$$\zeta(L) = \frac{W}{I} \frac{L^3}{3}, \quad (38.42)$$

the displacement of the end of a beam increases as the cube of the length.

In deriving our approximate beam theory, we have assumed that the axes section of the beam did not change when the beam was bent. When the thickness of the beam is small compared to the radius of curvature, the cross section changes very little and our result is O.K. In general, however, this effect cannot be neglected, as you can easily demonstrate for yourselves by bending a stiff rubber eraser in your fingers. If the cross section was originally rectangular, you will find that when

it is bent or bended at the bottom (see Fig. 38-15). This happens because when we compress the bar/bam, the material expands sideways – as described by Poisson's ratio. But then it's easy to bend or even to break it, as you only take a liquid in that, it's hard to change the volume – as it's only when you bend the eraser. For an incompressible material, Poisson's ratio would be exactly $1/2$ – for rubber it's nearly that.

38-5 Buckling

We want now to use our beam theory to understand the theory of the "buckling" of beams, or columns, or rods. Consider the situation sketched in Fig. 38-16 in which a rod that would normally be straight is held in its bent shape by two opposite forces that push on the ends of the rod. We would like to calculate the shape of the rod and the magnitude of the force on the ends.

Let the deflection of the rod from the straight line between the ends be $r(x)$, where x is the distance from one end. The bending moment $M(x)$ at the point P in the figure is equal to the force F multiplied by the moment arm, which is the perpendicular distance r ,

$$M(x) = Fr. \quad (38.42)$$

Using the beam equation (38.36), we have

$$\frac{d^2r}{dx^2} = F/EI. \quad (38.43)$$

For small deflections, we see that $1/R = -d^2r/dx^2$ (the minus sign because the curvature is downward). We get

$$\frac{d^2r}{dx^2} = -\frac{F}{EI}/R. \quad (38.44)$$

which is the differential equation of a sine wave. So for small deflections, the curve of such a beam looks is a sine curve. The "wavelength" λ of the sine wave is twice the distance L between the ends. If the bending is small, this is just twice the unbent length of the rod. So the curve is

$$r = R \sin(\pi x/L).$$

Taking the second derivative, we get

$$\frac{d^2r}{dx^2} = -\frac{\pi^2}{L^2}R.$$

Comparing this to Eq. (38.43), we see that the force is

$$F = \pi^2 EI/L^2. \quad (38.45)$$

For small enough the force is *independent of the bending displacement*!

We know that the building is physically – if the force is less than the F , given in Eq. (38.45), there will be no bending at all. But if F is slightly greater than F , the beam will suddenly bend a large amount – that is, it's most often the applied force F that's "YET" (prior to last the "stability") the beam will "buckle". If the bending on the second floor of a building exceeds the Euler angle for the supporting elements, the building will collapse. Another place where the buckling force is used is in space rockets. On one hand, the rocket must be able to withstand own weight on the launching pad and ensure the stresses during steady motion; on the other hand, it is important to keep the weight of the structure to a minimum, so that the pay load and the economy may be made as large as possible.

Actually a beam will just *collapse* completely when the force reaches the Euler force. When the displacements get large, the force is larger than

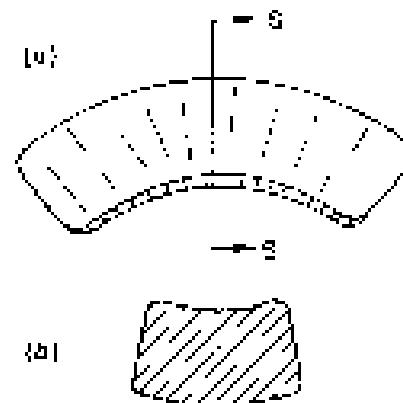


Fig. 38-15. (a) A bent eraser, (b) cross section.

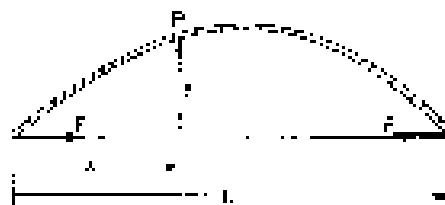


Fig. 38-16. A buckled beam.

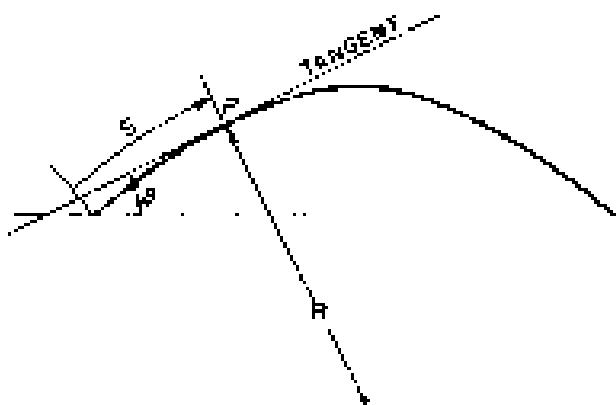


Fig. 38-17. The coordinates S and λ for the curve of a bent beam.

what we have found because of the terms in $1/R$ in Eq. (38.38) that we have neglected. To find the forces for a large bending of the beam, we have to go back to the exact equation, Eq. (38.44), which we had before we used the approximate relation between R and y . Equation (38.44) has a rather simple geometrical property.* It's a little complicated to work out, but rather interesting. Instead of describing the curve in terms of x and y , we can use two new variables: S , the distance along the curve, and λ , the slope of the tangent to the curve. See Fig. 38-17. The curvature is the rate of change of angle with distance:

$$\frac{1}{R} = -\frac{d\lambda}{dy}.$$

We can, therefore, write the exact equation (38.4) as

$$\frac{dy}{dx} = -\frac{F}{pI} \sin \lambda.$$

If we take the derivative of this equation with respect to S and replace dy/dxS by $d\lambda/dS$, we get

$$\frac{d^2\lambda}{dS^2} = -\frac{F}{pI} \sin \lambda. \quad (38.47)$$

[If λ is small, we get back Eq. (38.46). Every line is O.K.]

Now it may or may not delight you to know that Eq. (38.47) is exactly the same one you get for the large-amplitude oscillations of a pendulum with F/I replaced by another constant, of course. We learned how much in Chapter 11. We know how to find the solutions of such an equation by a numerical calculation—but the answers you get are some fascinating curves—known as the curves of the “Elastic.” Figure 38-18 shows three curves for different values of F/I .

* The exact equation appears, interestingly, in the problem of curved structures. In example 10 you saw all the surface of a liquid contained between parallel planes—and the geometric solution can be used.

† The solutions can also be expressed in terms of so-called functions called the “Jacobi elliptic functions,” the so-called Jacobi elliptic functions.

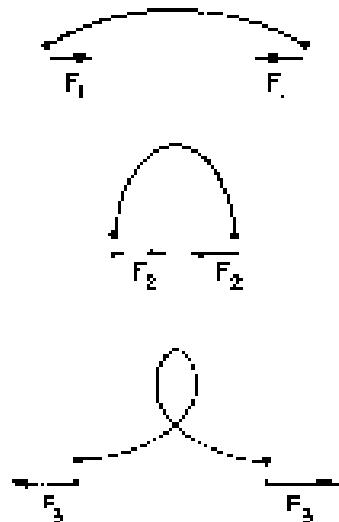


Fig. 38-18. Curves of a bent rod.

Elastic Materials

39-1 The tensor of strain

In the last chapter we talked about the distortions of various elastic objects. In this chapter we want to look at what can happen in general inside an elastic material. We would like to be able to describe the components of stress and strain inside some big blob of stuff which is roughly an sphere or in some complicated way. To do this, we need to be able to extend the idea of strain to every point in an elastic body; we can do this by giving a set of numbers—which are the components of a symmetric tensor—for each point. Earlier, we spoke of the stress tensor (Chapter 31); now we need the tensor of strain.

Imagine that we start with the material initially undeformed and watch the motion of a small speck of "dirt" embedded in the material when the strain is applied. A speck that was at the point P located at $r = (x, y, z)$ moves to a new position P' at $r' = (x', y', z')$ as shown in Fig. 39-1. We will call \mathbf{r} the vector displacement from P to P' . Then

$$\mathbf{r}' = \mathbf{r} + \mathbf{u} \quad (39-1)$$

The displacement \mathbf{u} depends, of course, on which point P we start with, so it is a vector function of r —or, if you prefer, of (x, y, z) .

Let's look first at a simple situation in which the strain is constant over the material—so we have what is called a homogeneous strain. Suppose, for instance, that we have a block of material and we stretch it uniformly. We just change all dimensions uniformly in one direction—say, in the x -direction, as shown in Fig. 39-2. The motion u_x of a speck at x is proportional to x . In fact,

$$\frac{u_x}{x} = \frac{\Delta l}{l}$$

We will write it this way:

$$u_x = \epsilon_{xx} x,$$

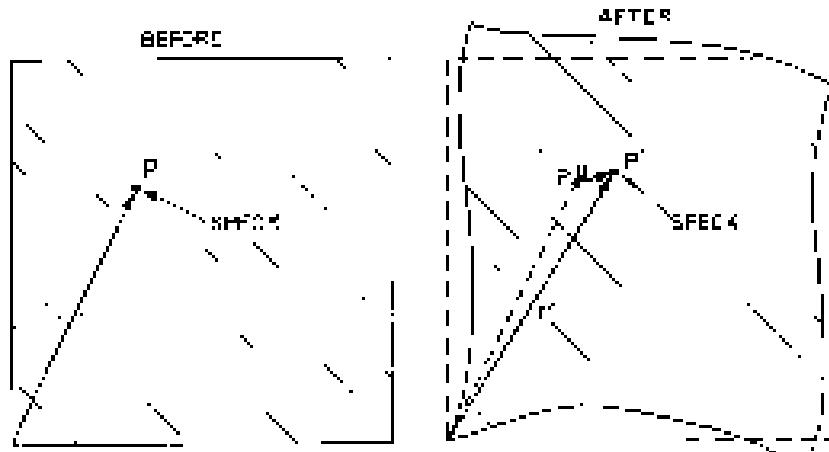


Fig. 39-1. A speck of the material at the point P in an undeformed block moves to P' where the block is strained.

39-1 The tensor of strain

39-2 The tensor of elasticity

39-3 The modulus in an elastic body

39-4 Nonelastic behavior

39-5 Calculating the elastic constants

Reference: C. Kittel, *Introduction to Solid State Physics*, John Wiley and Sons, Inc., New York, third ed., 1976.

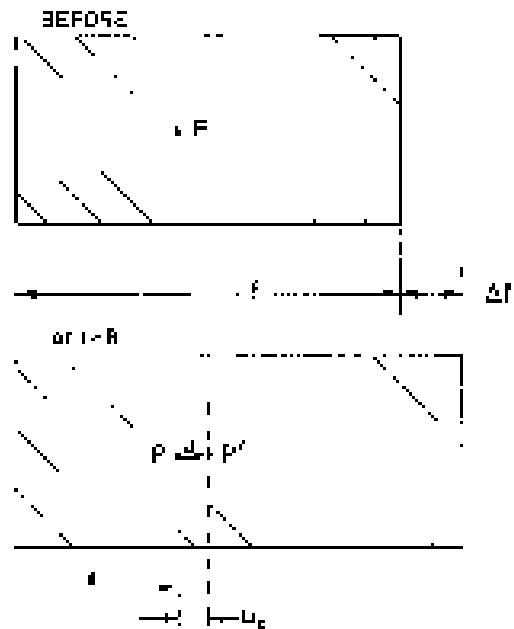


Fig. 39-2. A homogeneous stretch-type strain.

The proportionality constant ϵ_{xx} is, of course, the stretch ratio A/J . (You will see shortly why we use a double subscript.)

If the strain is not uniform, the relation between ϵ_x and ϵ will vary from place to place in the sample. For the general situation, we define the ϵ_{xy} by a kind of local δ/J , namely by

$$\epsilon_{xy} = \frac{\partial u_y}{\partial x}, \quad (39.2)$$

This number—which is now a function of x , y , and z —describes the amount of stretching in the y -direction throughout the block of jello. There may, of course, also be stretching in the x - and z -directions. We describe them by the numbers

$$\epsilon_{xz} = \frac{\partial u_z}{\partial x}, \quad \epsilon_{yz} = \frac{\partial u_z}{\partial y}. \quad (39.3)$$

We now turn to the inhomogeneous shear-type strain. Suppose we imagine a little cube snipped out of the initially unstrained jello. When the jello is passed through a shear, this cube may get elongated in a parallelepipedon, as sketched in Fig. 39.3. Up the back of each side of the cube an equal amount of stretching is proportional to y -displacement.

$$u_x = \frac{\delta}{2} y, \quad (39.4)$$

And there is also a rotation proportional to x ,

$$u_y = \frac{\theta}{2} x, \quad (39.5)$$

So we can describe such a shear-type strain by writing

$$u_x = \epsilon_{xy} y, \quad u_y = \theta x + \epsilon_{yz} x$$

with

$$\epsilon_{xz} = \epsilon_{yz} = \frac{\theta}{2}.$$

Now you might think that when the strains are not homogeneous, we could get rid of the generalized shear strain by defining the quantities ϵ_{xy} and ϵ_{yz} by

$$\epsilon_{xy} = \frac{\partial u_y}{\partial y}, \quad \epsilon_{yz} = \frac{\partial u_z}{\partial y}. \quad (39.6)$$

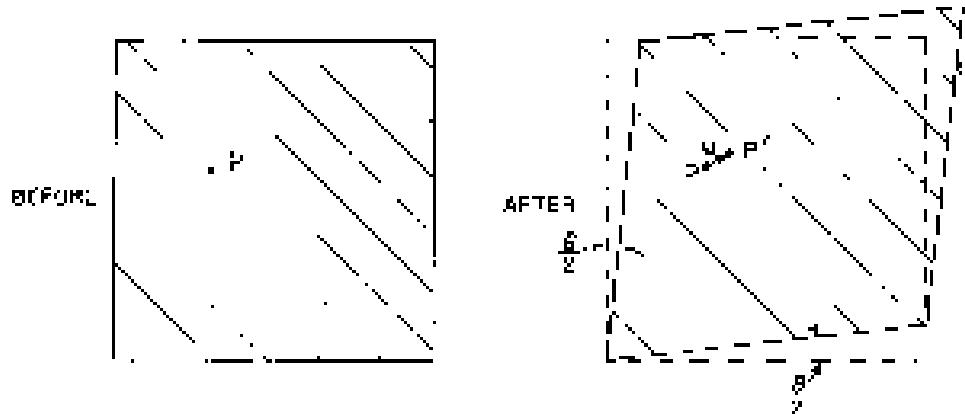


Fig. 39.3. A homogeneous shear strain.

But there is a difficulty. Suppose that the displacements u_x and u_y were given by

$$u_x = \frac{\theta}{2} y, \quad u_y = -\frac{\theta}{2} x,$$

* We propose for the moment to split the total shear angle θ into two equal parts and make the x-axis symmetric with angles $\pi/6$ and $-\pi/6$.

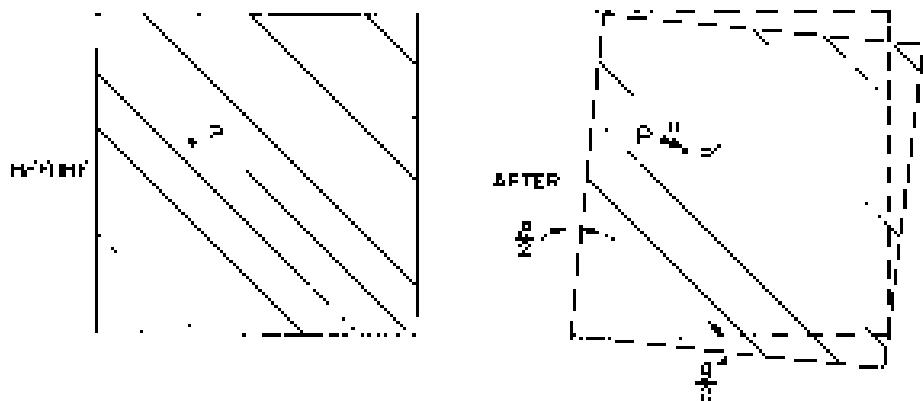


Fig. 39-4. A homogeneous rotation—there is no strain.

They are like Eqs. (39.4) and (39.5) except that the sign of ω_y is reversed. With these displacements a little cube in the jelly simply gets shifted by the angle $\omega_z/2$, as shown in Fig. 39-4. There is no strain at all; just a rotation in space. There is no distortion of the material, the relative positions of all the atoms are unchanged at all. We must somehow make our calculations so that plane rotations are not included in our definitions of a shear strain. The key point is that if ω_x/ω_z and ω_y/ω_z are equal and opposite, that is desirable, so we can lie things up by defining

$$\epsilon_{xy} = \epsilon_{yx} = \frac{1}{2}(\partial u_y / \partial x - \partial u_x / \partial y).$$

For a pure rotation they are both zero, but for a pure shear we get that ϵ_{xy} is equal to ϵ_{yx} , as we would like.

In the most general distortion—which may include stretching or compression as well as shear—we define the state of strain by giving the nine numbers

$$\begin{aligned}\epsilon_{xx} &= \frac{\partial u_x}{\partial x}, \\ \epsilon_{yy} &= \frac{\partial u_y}{\partial y}, \\ \epsilon_{zz} &= \frac{\partial u_z}{\partial z}, \\ \epsilon_{xy} &= \frac{1}{2}(\partial u_y / \partial x - \partial u_x / \partial y), \\ \epsilon_{yz} &= \frac{1}{2}(\partial u_z / \partial y - \partial u_y / \partial z), \\ \epsilon_{xz} &= \frac{1}{2}(\partial u_z / \partial x - \partial u_x / \partial z).\end{aligned}\quad (39.7)$$

These are the terms of a tensor of strain. Because it is a covariant tensor, our definitions make $\epsilon_{xy} = \epsilon_{yx}$ always; there are really only six different numbers. You remember (see Chapter 11) that the general characteristic of a tensor is that the terms transform like the products of the components of two vectors. (If A and B are vectors, $C_{ij} = A_i B_j$ is a tensor.) Each term of ϵ_{ij} is a product (or linear combination of products) of the components of the vector $\omega = (\omega_x, \omega_y, \omega_z)$, and of the vector $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$, which we know transforms like a vector. Let's let x_1 , x_2 , and x_3 stand for x , y , and z ; and u_1 , u_2 , and u_3 stand for u_x , u_y , and u_z ; then we can write the general term ϵ_{ij} of the strain tensor as

$$\epsilon_{ij} = \frac{1}{2}(\partial u_j / \partial x_i + \partial u_i / \partial x_j), \quad (39.8)$$

where i and j can be 1, 2, or 3.

When we have a homogeneous strain, which may include both stretching and shear, all of the ϵ_{ij} are constants, and we can write

$$u_i = \epsilon_{ix} x + \epsilon_{iy} y + \epsilon_{iz} z. \quad (39.9)$$

(We choose our origin of x_1 , x_2 , x_3 at the point where ω is zero.) In this case, the strain tensor ϵ_{ij} gives the ratio λ (shear) between two vectors: the coordinate vector $r = (x_1, x_2, x_3)$ and the displacement vector $u = (u_1, u_2, u_3)$.

When the strains are not homogeneous, any piece of the jelly may also get somewhat twisted - there will be a local rotation. If the distortions are all small, we would have

$$\Delta\omega_i = \sum_j (\epsilon_{ij} - \omega_{ij}) \Delta x_{ij}, \quad (39.10)$$

where ω_{ij} is an unperturbed tensor,

$$\omega_{ij} = \frac{1}{2}(\partial u_j / \partial x_i + \partial u_i / \partial x_j), \quad (39.11)$$

which describes the rotation. We will, however, not worry any more about rotations, but only about the strains described by the symmetric tensors.

39-2. The tensor of elasticity

Now that we have described the strains, we want to relate them to the internal forces: the stresses in the material. For the small piece of the material, we assume Hooke's law holds and write just the stresses and proportional to the strains. In Chapter 11 we defined the stress tensor S_{ij} as the P.I. (the product of the force across a unit area) perpendicular to the j -axis. Hooke's law says that each component of S_{ij} is linearly related to each of the components of strain. Since 5 and ϵ each have nine components, there are $9 \times 9 = 81$ possible coefficients which describe the elastic properties of the material. They are constants if the material itself is homogeneous. We write these coefficients as C_{ijkl} and define them by the equation

$$S_{ij} = \sum_k C_{ijkl} \epsilon_{kl}, \quad (39.12)$$

where i, j, k, l all take on the values 1, 2, or 3. Since the coefficients C_{ijkl} relate one tensor to another, they also form a tensor - a tensor of the fourth rank. We can call it the tensor of elasticity.

Suppose that all the C 's are known and that you put a complicated force on an object of some peculiar shape. There will be all kinds of distortion, and the thing will settle down with some twisted shape. What are the displacements? You can see that it is a complicated problem. If you knew the strains, you could find the stresses from Eq. (39.12) - or vice versa. But the stresses and strains you end up with at any point depend on what happens in all the rest of the material.

The easiest way to get at the problem is by thinking of the energy. When there is a force F proportional to a displacement x , say $F = kx$, the work required for any displacement x is $kx^2/2$. In a similar way, the work W that goes into each cubic volume of a distorted material turns out to be

$$W = \frac{1}{2} \sum_{ijkl} C_{ijkl} \epsilon_{kl} \Delta x_{ij}, \quad (39.13)$$

The total work W done in deforming the body is the integral of W over all the

$$W = \int \frac{1}{2} \sum_{ijkl} C_{ijkl} \epsilon_{kl} \Delta x_{ij} d\text{Vol}. \quad (39.14)$$

This is then the potential energy stored in the internal stresses of the material. Now when a body is in equilibrium, this internal energy must be at a minimum. So the problem of finding the strains in a body can be solved by finding the set of displacements throughout the body which will make W a minimum. In Chapter 19 we gave some of the general ideas of the calculus of variations. Let us look at buckling minimization problems like this. We cannot go into the problem in any more detail here.

What we are mainly interested in now is what we can say about the general properties of the tensor of elasticity. First, it is clear that there are not really 81 different terms in C_{ijkl} . Since just S_{ij} and ϵ_{ij} are symmetric tensors, and with only six different terms, there can be at most 36 different terms in C_{ijkl} . There are, however, usually many fewer than this.

Let's look at the special case of a cubic crystal. In it, the energy density is given as follows:

$$\begin{aligned} \epsilon &= \frac{1}{2}(C_{xx}e_{xx}^2 + C_{yy}e_{yy}^2 + C_{zz}e_{zz}^2) \\ &= C_{xy}e_{xy}^2 + C_{yz}e_{yz}^2 + \dots + C_{xz}e_{xz}^2 + \dots \quad (39.15) \end{aligned}$$

with δ_i 's running in all! Now a cubic crystal has certain symmetries, etc. In particular, if the crystal is rotated 90° , it has the same physical properties—it has the same stiffness for stretching in the x -direction as for stretching in the y -direction, etc. Therefore, if we change our definition of the coordinate directions x and y in Eq. (39.15), the energy wouldn't change. It must be true for a cubic crystal

$$C_{xx} = C_{yy} = C_{zz} \quad (39.16)$$

Next we can show that the terms like C_{xy} must be zero. A cubic crystal has the property that it is symmetric under a reflection in any plane perpendicular to one of the axes. If we replace y by $-y$, nothing is different. But changing y to $-y$ changes e_{xy} to $-e_{xy}$, a displacement which was toward y is now toward $-y$. If the energy is not to change, C_{xy} must go into $-C_{xy}$ when we make a reflection. But a reflected crystal is the same as before, so C_{xy} must be the same as C_{-xy} . This can happen only if $C_{xy} = 0$.

You say, "But the same argument will make $C_{xy} = 0$!" No, because there are four y 's. The sign changes once for each y , and four times make a plus. If there are *odd* number's, the term does not have to be zero. It is zero only when there is *even*, etc. Now, for a cubic crystal, any nonzero term of C will have only an even number of identical subscripts. (The arguments we have made for y obviously hold also for x and z .) We might then have terms like C_{xxy} , C_{xyz} , C_{xyz} , and so on. We have already shown, however, that if we change all x 's to y 's and vice versa (or all z 's and x 's, and so on) we must get—for a cubic crystal—the same number. This means that there are only four different nonzero possibilities:

$$\begin{aligned} C_{xxx} &= C_{yyy} = C_{zzz}, \\ C_{xyy} &= C_{yyx} = C_{zyz}, \text{ etc.} \\ C_{xzy} &= C_{zyx} = C_{xyz}, \text{ etc.} \end{aligned} \quad (39.17)$$

For a cubic crystal, then, the energy density will look like this:

$$\begin{aligned} \epsilon &= \frac{1}{2}(C_{xxx}e_{xx}^2 + e_{yy}^2 + e_{zz}^2 \\ &\quad + 2C_{xyy}e_{xy}^2 + e_{yyx}^2 + e_{yyz}^2) \\ &\quad + 4C_{xzy}e_{xz}^2 - e_{yy}^2 - e_{yyz}^2). \end{aligned} \quad (39.18)$$

For an isotropic—that is, monocrystalline—material, the symmetry is still higher. The C 's must be the same for any choice of the coordinate system. Then it turns out that there is another relation among the C 's, namely, that

$$C_{xxx} = C_{yyy} = C_{zzz} \quad (39.19)$$

We can see that this is so by the following general argument. The stress tensor S_{ij} has to be related to e_{ij} in a way that doesn't depend at all on the coordinate directions—it must be related only by *scalar* quantities. "That's easy," you say. "The only way to obtain S_{ij} from e_{ij} is by multiplication by a scalar constant. It's just Hooke's law. It must be that $S_{ij} = (\text{const.})e_{ij}$." But that's not quite right: there could also be the *wave vector* k , multiplied by some scalar, linearly related to e_{ij} . The only invariant you can make that is linear in the e 's is $\sum e_{ii}$. (It transforms like $x^2 + y^2 + z^2$, which is a scalar.) So the most general form for the equation relating S_{ij} to e_{ij} —for isotropic materials—is

$$S_{ij} = 2\mu e_{ii} + \lambda \left(\sum_k e_{kk} \right) \delta_{ij}. \quad (39.20)$$

(The first constant is usually written as two times μ ; then the coefficient λ is equal

to the shear modulus we defined in the last chapter). The constants μ and λ are called the Timoshenko elastic constants. Comparing Eq. (39.20) with Eq. (39.13), you see that

$$\begin{aligned} C_{111111} &= \lambda, \\ C_{222222} &= 2\mu, \\ C_{121212} &= 2\mu + \lambda. \end{aligned} \quad (39.21)$$

So we have proved that Eq. (39.10) is indeed true. You also see that the elastic properties of an isotropic material are completely given by two constants, μ and λ , as in the last chapter.

The C 's can be put in terms of any two of the elastic constants we have used earlier—for instance, in terms of Young's modulus E and Poisson's ratio ν . We will leave it for you to show that

$$\begin{aligned} C_{111111} &= \frac{E}{1 - \nu} \left(1 + \frac{\nu}{2\mu} \right), \\ C_{222222} &= \frac{E}{1 - \nu} \left(\frac{\nu}{2\mu} \right), \\ C_{121212} &= \frac{E}{(1 + \nu)^2}. \end{aligned} \quad (39.22)$$

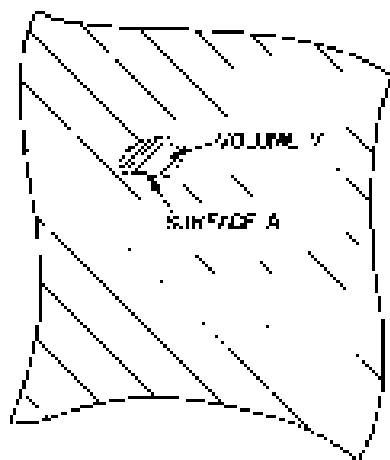


Fig. 39-5. A small volume element V bounded by the surface A .

39-3 The motions in an elastic body

We have pointed out that for an elastic body in equilibrium the internal stresses adjust themselves to make the energy minimum. Now we take a look at what happens when the internal forces are not in equilibrium. Let's say we have a small piece of the material inside some surface A . See Fig. 39-5. If the piece is in equilibrium, the total force F acting on it must be zero. We can think of this force as being made up of two parts. There could be one part due to "external" forces like gravity, which act from a distance on the matter in the piece to produce a force per unit volume f_{ext} . The total external force F_{ext} is the integral of f_{ext} over the volume of the piece:

$$F_{ext} = \int f_{ext} dV. \quad (39.23)$$

In equilibrium, this force would be balanced by the total force F_{int} from the neighboring material which acts across the surface A . When the piece is not in equilibrium—if it is moving—the sum of the internal and external forces is equal to the mass times the acceleration. We would have

$$F_{tot} = F_{int} = \int \rho r dV, \quad (39.24)$$

where ρ is the density of the material, and r is its acceleration. We can now combine Eqs. (39.23) and (39.24) writing

$$F_{int} = \int (f_{ext} - \rho r) dV. \quad (39.25)$$

We will simplify our writing by defining

$$f = -f_{ext} + \rho r. \quad (39.26)$$

Then Eq. (39.25) is written

$$F_{int} = \int f dV. \quad (39.27)$$

What we have called F_{int} relates to the stress in the material. The stress tensor S_{ij} was defined (Chapter 37) so that the i -component of the force of an area dA on a surface element dA , whose unit normal is a_i , is given by

$$dF_i = (S_{xx})_{ij} a_i + (S_{yy})_{ij} a_j + (S_{zz})_{ij} a_k dA. \quad (39.28)$$

The decomposition of F_{ij} on our little piece is then the integral of dF , over the surface. Substituting this into the ϵ_{ijk} -component of Eq. (39.27), we get

$$\int_{\partial V} (S_{ij}\epsilon_{ijk} - S_{kj}\epsilon_{ijk} - S_{ki}\epsilon_{ijk}) d\sigma = \int_V f_j dV. \quad (39.29)$$

We have a surface integral related to a volume integral—and that reminds us of something we learned in electricity. Were I to verify you might see the first component S_{ij} on each of the S 's in the left-hand side of Eq. (39.29); it looks just like the integral of a quantity “ N ” $\cdot \mathbf{n}$, that is, the normal component of a vector over the surface. It would be the flux of “ N ” out of the volume. And this could be written, using Gauss law, as the volume integral of the divergence of “ N ”. It is, in fact, true whether the x -subscript i is there or not; it's just a mathematical theorem you get by integrating by parts. In other words, we can change Eq. (39.29) into

$$\int_V \left(\frac{\partial S_{ij}}{\partial x_i} + \frac{\partial S_{ik}}{\partial x_j} + \frac{\partial S_{jk}}{\partial x_i} \right) dx = \int_V f_j dV. \quad (39.30)$$

Now we can leave off the volume integral and write the differential equation for the general component ϵ_{ijk}^i as

$$f_j = \sum_i \frac{\partial S_{ij}}{\partial x_j}. \quad (39.31)$$

This tells us how the force per unit volume is related to the stress tensor S_{ij} .

The theory of the motions inside a solid works this way. If we start out knowing the initial displacement—given by, say, \mathbf{u} —we can work out the strains ϵ_{ij} . From the strains we can get the stresses from Eq. (39.12). From the stresses we can get the force density f in Eq. (39.31). Knowing f we can go to Eq. (39.16), the acceleration \mathbf{r} of the material, which tells us how the displacements will be changing. Putting everything together, we get the familiar equation of motion for an elastic solid. We will also write down the results that come out for an isotropic material. If you use (39.20) for S_{ij} , and write the ϵ_{ij} as $\partial u_i / \partial x_j = \partial u_j / \partial x_i$, you end up with the vector equation

$$f = (\alpha + \omega) \nabla(\nabla \cdot \mathbf{u}) + \nu \nabla^2 \mathbf{u}. \quad (39.32)$$

You can, in fact, see that the equation relating f and \mathbf{u} must have this form. The force must depend on the second derivatives of the displacement \mathbf{u} . What second derivatives of \mathbf{u} are there that are vectors? One is $\nabla(\nabla \cdot \mathbf{u})$; that's a true vector. The only other one is $\nabla^2 \mathbf{u}$. So the most general form is

$$f = \alpha \nabla(\nabla \cdot \mathbf{u}) + \nu \nabla^2 \mathbf{u},$$

which is just (39.42) with a different relation of the constants. You may be wondering why we can't have a third term using $\nabla \times \nabla \times \omega$, which is also a vector. But remember that $\nabla \times \nabla \times \omega$ is the same thing as $\nabla^2 \omega - \nabla(\nabla \cdot \omega)$, so this linear combination of the relations we have. Adding it would add nothing new. We have proved once more that isotropic material has only two elastic constants.

For the equation of motion of the material, we can set (39.32) equal to $\rho \ddot{\mathbf{u}}^2 / \partial t^2$, neglecting for now any body forces like gravity, and get

$$\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = (\lambda + \nu) \nabla(\nabla \cdot \mathbf{u}) + \nu \nabla^2 \mathbf{u}. \quad (39.33)$$

It looks something like the wave equation we had in electrostatics, except that there is an additional complicating term. For materials whose elastic properties are everywhere the same we can see what the material looks like in the following way. You will remember that any vector field can be written as the sum of two vectors: one whose divergence is zero, and the other whose curl is zero. In

other words, we can put

$$\mathbf{a} = \mathbf{u}_1 + \mathbf{u}_2, \quad (39.34)$$

where

$$\nabla \cdot \mathbf{u}_1 = 0, \quad \nabla \times \mathbf{u}_1 = 0. \quad (39.35)$$

Substituting $\mathbf{u}_1 + \mathbf{u}_2$ for \mathbf{a} in (39.34), we get

$$\rho \partial^2/\partial t^2[\mathbf{u}_1 + \mathbf{u}_2] = (\lambda + \mu) \nabla(\nabla \cdot \mathbf{u}_2) + \mu \nabla^2(\mathbf{u}_1 + \mathbf{u}_2). \quad (39.36)$$

We can eliminate \mathbf{u}_1 by taking the divergence of this equation,

$$\rho \partial^2/\partial t^2(\nabla \cdot \mathbf{u}_2) = (\lambda + \mu) \nabla^2(\nabla \cdot \mathbf{u}_2) - \mu \nabla \cdot \nabla^2 \mathbf{u}_2.$$

Since the operators (∇^2) and $(\nabla \cdot)$ can be interchanged, we can factor out the divergence to get

$$\nabla \cdot \{\rho \partial^2 \mathbf{u}_2 / \partial t^2 - (\lambda + 2\mu) \nabla^2 \mathbf{u}_2\} = 0. \quad (39.37)$$

Since $\nabla \times \mathbf{u}_2$ is zero by definition, the curl of the bracket $\{\}$ is also zero; so the bracket itself is identically zero, and

$$\rho \partial^2 \mathbf{u}_2 / \partial t^2 = (\lambda + 2\mu) \nabla^2 \mathbf{u}_2. \quad (39.38)$$

This is the vector wave equation for waves which move at the speed $C_0 = \sqrt{\lambda + 2\mu}/\rho$. Since the curl of \mathbf{u}_2 is zero, there is no shearing associated with this wave, it's wave is just the compressional longitudinal wave we discussed in the last chapter, and the velocity is just what we found for C_{long} .

In a similar way—by taking the curl of Eq. (39.36)—we can show that \mathbf{u}_1 satisfies the equation

$$\rho \partial^2 \mathbf{u}_1 / \partial t^2 = \mu \nabla^2 \mathbf{u}_1. \quad (39.39)$$

This is again a vector wave equation for waves with the speed $C_0 = \sqrt{\lambda + 2\mu}/\rho$. Since $\nabla \cdot \mathbf{u}_1$ is zero, \mathbf{u}_1 produces no changes in density; the vector \mathbf{u}_1 corresponds to the transverse, or shear-type, wave we saw in the last chapter, and $C_0 = C_{shear}$.

If we wished to know the static stresses in an isotropic material, we could, in principle, find them by solving Eq. (39.33) with \mathbf{f} equal to zero—or equal to the static body forces from gravity such as \mathbf{g} —under certain conditions which are related to the forces acting on the surfaces of our large block of material. This is somewhat more difficult to do than the corresponding problems in electromagnetism. It is more difficult first, because the equations are a little more difficult to handle, and second, because the shape of the elastic bodies we are likely to be interested in are usually much more complicated. In electromagnetism, we are often interested in solving Maxwell's equations around relatively simple geometric shapes such as cylinders, spheres, and so on, since these are conventional shapes for electrical devices. In elasticity, the objects we would like to analyze may have quite complicated shapes—like a crane hook, or an automobile crankshaft, or the rotor of a gas turbine. Such problems can sometimes be worked out approximately by numerical methods, using the minimum energy principle we mentioned earlier. Another way is to use a model of the object and measure the stresses using experiments, using polarized light.

It works this way. When a transparent isotropic material—like example, a clear plastic like cellophane—is put under stress, it becomes birefringent. If you pass polarized light through it, the plane of polarization will be rotated by an amount related to the stress; by measuring the rotation, you can measure the stress. Figure 39-6 shows how such a setup might look. Figure 39-7 is a photograph of a dielectric model of a complicated shape under stress.

39-4 Nonelastic behavior

In all that has been said so far, we have assumed that stress is proportional to strain; in general, that is not true. Figure 39-8 shows a typical stress-strain curve for a ductile material. For small strains, the stress is proportional to the strain

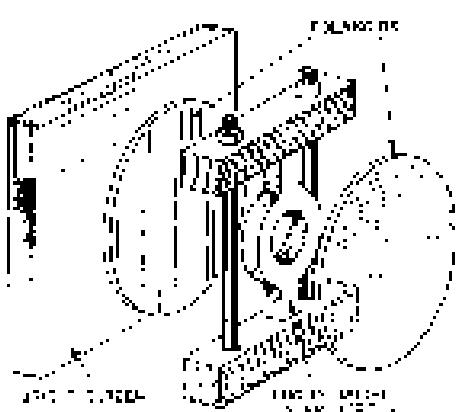


Fig. 39-6. Measuring internal stresses with polarized light.

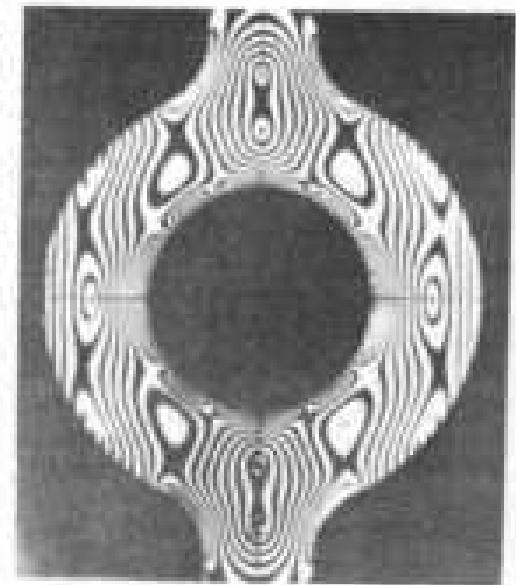


Fig. 39-7. A stressed plastic model, seen between crossed polarizers (From E. W. Sere, Optics, Addison Wesley Publishing Co., Reading, Mass., 1949.)

strain. Eventually, however, after a certain point, the relationship between stress and strain begins to deviate from a straight line. For many materials, the ones we would call "brittle," the object breaks for strains only a little above the point where the curve starts to bend over. In general, there are other complications in the stress-strain relationship. For example, if you strain an object, the stresses may be high at first, but decrease slowly with time. Also, if you go to high stresses, but still just to the "breaking" point, when you lower the strain the stress will return along a different curve. There is a small hysteresis effect (like the one we saw between β and γ in magnetic materials).

The stress at which a material will break varies widely from one material to another. Some materials will break when the maximum tensile stress reaches a certain value. Other materials will fail when the maximum shear stress reaches a certain value. Chalk is an example of a material which is much weaker in tension than in shear. If you pull on the ends of a piece of blackboard chalk, the chalk will break perpendicular to the direction of the applied stress, as shown in Fig. 39-9(a). It breaks perpendicular to the applied force because it is only a bunch of particles packed together which are easily pulled apart. The material is, however, much better in shear, because the particles go in each other's way. Now you will remember that when we had a rod in torsion there was a shear all around it. Also, we showed that shear was equivalent to a combination of a tension and compression at 45° . For these reasons, if you twist a piece of blackboard chalk, it will break along a complicated surface which starts out at 45° to the axis. A photograph of a piece of chalk broken in this way is shown in Fig. 39-9(b). The chalk breaks where the material is in maximum tension.

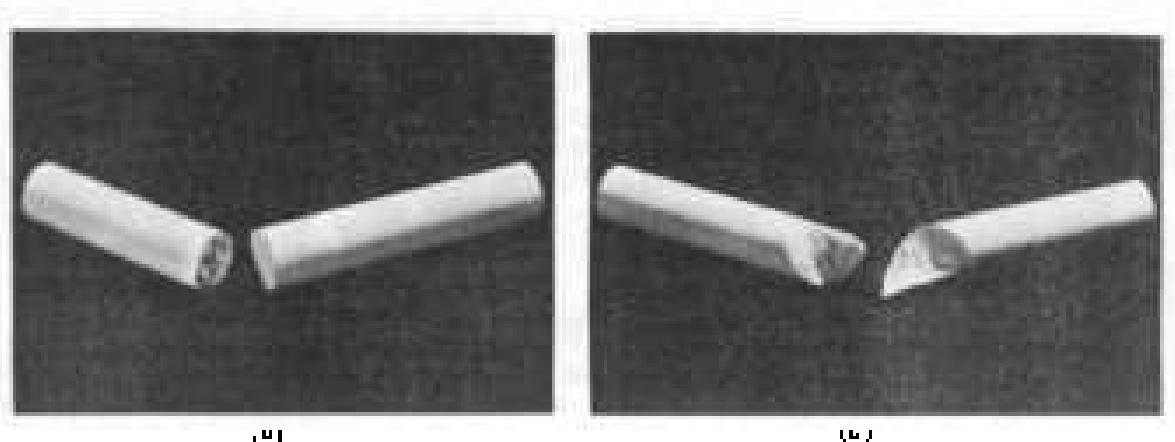


Fig. 39-9. (a) A piece of chalk broken by pulling on the ends; (b) a piece broken by twisting.

Other materials behave in strange and complicated ways. The more complicated the materials are, the more interesting their behavior. If we take a sheet of "Scotch-Wrap" and crumple it up into a ball and throw it on the table, it slowly unfolds itself and returns toward its original flat form. At first sight, we might be tempted to think that it is merely "lazy" and prefers it's returning to its original form. However, a simple calculation shows that the energy is several orders of magnitude too small to account for the effect. There appear to be real important complicating factors: "Something" inside the material "remembers" the shape it had initially and "wants" to get back there, but something else "prefers" the new shape and "resists" the return to the old shape.

We will not attempt to describe the mechanism at play in the Scotch plastic, but you can get an idea of how such an effect might come about from the following model. Suppose you imagine a material made of long, flexible, but strong, fibers mixed together with some hollow cells filled with a viscous liquid. Imagine also that there are narrow pathways from one cell to the next so the liquid can leak slowly from a cell to its neighbor. When we examine a sheet of this stuff, we stretch the long fibers, squeezing the liquid out of the cells in one place and forcing it into other cells which are being stretched. When we let go, the long fibers try to

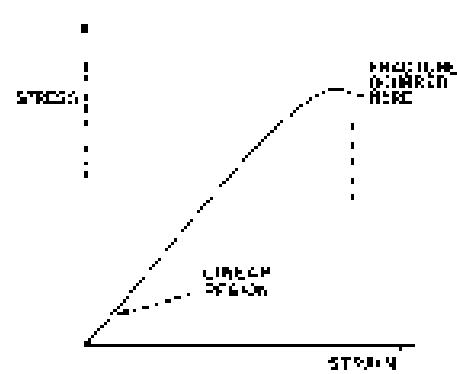


Fig. 39-8. A typical stress-strain relation for large strains.

return to their original shape. But to do that, they have to force the liquid back to its original location, which will happen very slowly because of the viscosity. The forces we apply in stretching a sheet are much larger than the forces exerted by the fibers. We can calculate the sheet quickly, but it will return much slowly. It is fundamentally a consideration of large stiff molecules and smaller, more mobile ones at the *Brown-Wiech* that is responsible for ice behavior. This also fits with the fact that the material returns more quickly to its original shape when it's warmed up than when it's cold—the heat increases the mobility (decreases the viscosity) of the smaller molecules.

Although we have been discussing how Hooke's law breaks down, the remarkable thing is perhaps not that Hooke's law breaks down for large strains but that it should be so *surprisingly* true. We can get some idea of why this might be by looking at the strain energy of a material. To say that the stress is proportional to the strain is the same thing as saying that the strain energy *varies* as a *square* of the strain. Suppose we had a rod and we bent it through a small angle θ . If Hooke's law holds, the strain energy would be proportional to the square of θ . Suppose we were to assume that the energy were some arbitrary function of the angle; we could write it as a Taylor expansion around zero angle:

$$U(\theta) = U(0) + U'(0)\theta - \frac{1}{2}U''(0)\theta^2 + \frac{1}{3}U'''(0)\theta^3 \dots \quad (39.40)$$

The torque τ is the derivative of U with respect to angle; we would have

$$\tau(\theta) = U'(0) + U''(0)\theta - \frac{1}{2}U'''(0)\theta^2 + \dots \quad (39.41)$$

Now if we increase our angles from the equilibrium position, the last term is zero, so the first term, $U'(0)$, is proportional to θ ; and for small enough of θ , this will dominate the term at θ^2 . (Actually, materials are sufficiently symmetric laterally to find $\tau(0) = -\tau(-0)$, the term in θ^3 will be zero, and the dependence from linearity would come only from the θ^2 term.) There is, however, no reason why this should be true for compressions and tensions. [See Fig. 39-10, we have not explained *why* materials usually break soon after the higher-strain terms become significant.]

39-5 Calculating the elastic constants

As our last topic on elasticity we would like to show how one could try to calculate the elastic constants of a material, starting with some knowledge of the properties of the atoms which make up the material. We will take only the simple case of an *ideal* cubic crystal like sodium chloride. When a crystal is strained, its volume or its shape is changed. Such changes result in an increase in the potential energy of the crystal. To calculate the change in strain energy, we have to know where each atom goes. In complicated crystals, the atoms will rearrange themselves in the lattice in very complicated ways to make the total energy as small as possible. This makes the computation of the strain energy rather difficult. In the case of a simple cubic crystal, however, it is easy to see what will happen. The extensions made the crystal will be geometrically similar to the distortions of the cubic host crystal.

We can calculate the elastic constants for a cubic crystal in the following way. First, we can express the law between each pair of atoms in the crystal. Then, we calculate the change in the potential energy of the crystal when it is distorted from its equilibrium shape. This gives us a relation between the energy and the strain which is quadratic in all the strains. Comparing the energy obtained this way with Eq. (39.13), we can identify the coefficient of each term with the elastic constants C_{ijkl} .

For our example we will assume a simple force law: that the force between neighboring atoms is a central force, by which we mean that it acts along the line between the two atoms. We would expect the forces in ionic crystals to be like this, since they are just primarily Coulomb forces. (The forces of covalent bonds are usually more complicated, since they can exert a sideways push on a nearby atom.)

atom; we will leave out this complication.) We are also going to include only the forces between each atom and its nearest and next-nearest neighbors. In other words, we will make no approximations which neglects all forces beyond the next nearest neighbor. The forces we will include are shown for the xy -plane in Fig. 19-10(a). The corresponding forces in the yz - and zx -planes also have to be included.

Since we are only interested in the elastic conditions which apply to small strains, and therefore only want the forces to be linearly related quadratically with the strains, we can imagine that the force between each atom pair varies linearly with the displacements. We can also imagine that each pair of atoms is joined by a linear spring, as drawn in Fig. 39-10(a). All of the springs between a sodium atom and a chlorine atom should have the same spring constant, say k_1 . The springs between two sodiums and between two chlorines would have different constants, but we will make our discussion simpler by taking them equal; we call them k_1 . (We could come back later and make them different after we have seen how the calculations go.)

Now we assume that the crystal is deformed by a homogeneous strain described by the strain tensor ϵ_{ij} . In general, it will have components involving x , y , and z ; but we will consider now only a shear with the three components ϵ_{12} , ϵ_{23} , and ϵ_{13} so that it will be easy to visualize. If we pick one atom at our origin, the displacement of every other atom is given by equations like Eq. (31) by:

$$\begin{aligned} u_2 &= c_{21}x + c_{22}v \\ u_3 &= c_{31}x + c_{32}v \end{aligned} \quad (19.42)$$

Suppose we call the beam or $x = y = 0$ "beam 1" and imagine it's negligible in the xy -plane as shown in Fig. 39-11. Calling the lattice constant a , we get the x and y displacements \bar{u} and \bar{v} , listed in Table 39-1.

Now we can calculate the energy stored in the springs, which is $E^2/3$ times the square of the extension for each spring. For example, the energy in the horizontal spring between atom 1 and atom 2 is

$$\frac{k_1(k_1, \omega)^L}{2}, \quad (39.43)$$

Note that to first order, the y -displacement of atom 2 does not change the length of the spring between atom 1 and atom 2. To get the total energy in a diagonal displacement, such as atom 1 to atom 3, however, we need to calculate the change in length due to both the horizontal and vertical displacements. For small displacements from the

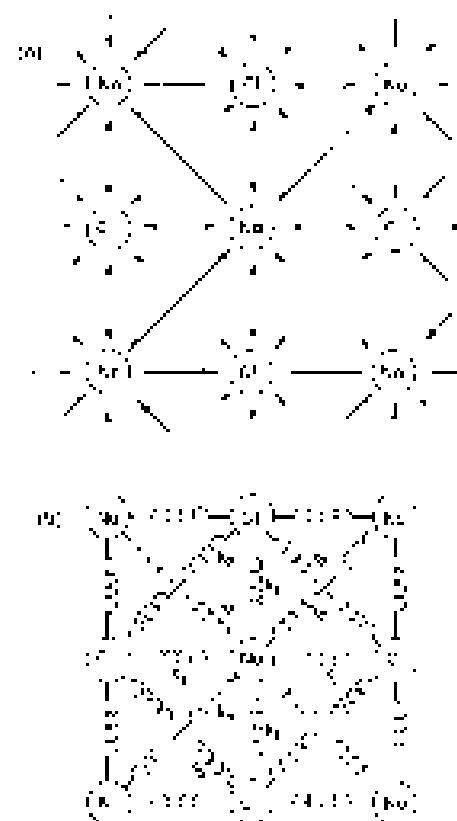


Fig. 99-12. (a) The interatomic forces we are taking into account; (b) a model in which the atoms are connected by springs.

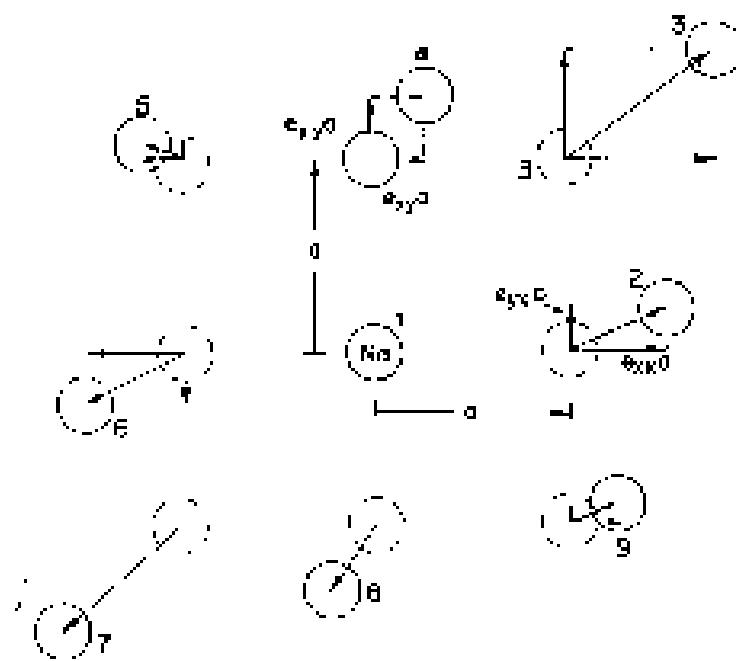


Fig. 39-11. The displacements of the nearest and next-nearest neighbors of whom I (cross-hatched).

Table 39-1

Atom	Location x, y	α_x	α_y	δ
1	0, 0	0	0	—
2	$a, 0$	$c_{xx}v$	$c_{yy}v$	k_1
3	a, a	$(c_{xx} + c_{yy})v$	$(c_{yy} - c_{xx})v$	k_2
4	$0, a$	$c_{yy}v$	$c_{xx}v$	k_1
5	$-a, 0$	$(-c_{xx} - c_{yy})v$	$(-c_{yy} + c_{xx})v$	k_2
6	$-a, a$	$c_{yy}v$	$c_{xx}v$	k_1
7	$-a, -a$	$(c_{xx} - c_{yy})v$	$(c_{yy} + c_{xx})v$	k_1
8	$0, -a$	$-c_{yy}v$	$-c_{xx}v$	k_1
9	$a, -a$	$(c_{yy} - c_{xx})v$	$(c_{xx} + c_{yy})v$	k_2

original cube, we can write the change in the distance to atom 3 as the sum of the components of α_x and α_y in the diagonal direction, namely as

$$\frac{1}{\sqrt{2}}(\alpha_x + \alpha_y).$$

Using the values of α_x and α_y from the table we get the energy

$$\frac{k_2}{2} \left(\frac{\alpha_x + \alpha_y}{\sqrt{2}} \right)^2 = \frac{k_2 v^2}{4} (c_{xx} + c_{yy} + c_{xy} + c_{yz})^2. \quad (39.43)$$

For the total energy for all the springs in the xy -plane, we need the sum of eight terms like (39.43) and (39.44). Calling this energy E_1 , we get

$$\begin{aligned} E_1 = & \frac{a^2}{2} \left[k_1 c_{xx}^2 + \frac{k_2}{2} (c_{xx} + c_{yy} + c_{xy} + c_{yz})^2 \right. \\ & + k_1 c_{yy}^2 + \frac{k_2}{2} (c_{xx} + c_{yy} + c_{xy} + c_{yz})^2 \\ & + k_1 c_{xy}^2 + \frac{k_2}{2} (c_{xx} + c_{yy} + c_{xy} + c_{yz})^2 \\ & \left. + k_1 c_{yz}^2 + \frac{k_2}{2} (c_{xx} + c_{yy} + c_{xy} + c_{yz})^2 \right]. \end{aligned} \quad (39.44)$$

To get the total energy of all the springs connected to atom 1, we must make one addition to the values in Eq. (39.43). Even though we have only x - and y -components of the strain, there are still various energies associated with the next-nearest neighbours of the atom. Thus additional energy is

$$k_2(c_{xx}^2 + c_{yy}^2). \quad (39.45)$$

The elastic constants are related to the energy density ϵ by Eq. (39.11). The energy we have calculated is the energy associated with one atom, or rather, it is twice the energy per atom, since one-half of the energy of each spring should be assigned to each of the two atoms it joins. Since there are 10^{23} atoms per unit volume, ϵ and C_{ijkl} are related by

$$\epsilon = \frac{C_{ijkl}}{2a^4}.$$

To find the elastic constants C_{ijkl} , we need only to expand out the springs in Eq. (39.44), adding the terms of (39.45)—and compare the coefficients of v^2 with the corresponding coefficient in Eq. (39.13). For example, collecting the terms

in ϵ_{12} and in ϵ_{23} , we get the factor

$$(k_1 + 2k_2)\sigma^2.$$

so

$$C_{xxz} = C_{yyz} = \frac{k_1 + 2k_2}{\sigma}.$$

For the remaining terms, there is a slight complication. Since we cannot distinguish the product of two terms like $\epsilon_{12}\epsilon_{23}$ from $\epsilon_{23}\epsilon_{12}$, the coefficient of such terms in our energy is equal to the sum of two terms in Eq. (39.13). The coefficient of $\epsilon_{12}\epsilon_{23}$ in Eq. (39.13) is $2k_3$, so we have that

$$(C_{xxx} + C_{yyz}) = \frac{2k_3}{\sigma}.$$

But because of the symmetry in our crystal, $C_{xxx} = C_{yyz}$, so we have that

$$C_{xxx} = C_{yyz} = \frac{k_3}{\sigma}.$$

By a similar process, we can also get

$$C_{xxy} = C_{yyx} = \frac{k_3}{\sigma}.$$

Finally, you will notice that any term which involves either ϵ_x or ϵ_y only once is zero—as we concluded earlier from symmetry arguments. Summarizing our results:

$$\begin{aligned} C_{xxx} &= C_{yyz} = \frac{k_1 + 2k_2}{\sigma}, \\ C_{xxy} &= C_{yyx} = \frac{k_3}{\sigma}, \\ C_{xyx} &= C_{yxy} = C_{xyz} = C_{zyx} = \frac{k_3}{\sigma}, \\ C_{zyx} &= C_{xyz} = 0. \end{aligned} \quad (39.49)$$

We have been able to relate the bulk elastic constants to the atomic properties which appear in the constants k_1 and k_2 . In our particular case, $C_{xyx} = C_{xyz}$. It turns out—as you can perhaps see from the way the calculations went—that these terms are always equal for a cubic crystal, no matter how many force terms are taken into account, provided only that the forces act along the line joining each pair of atoms—that is, so long as the forces between atoms are like springs and don't have a sideways component as you might get from a cantilevered beam (and yet the net is covariant overall).

We can check this conclusion with the experimentally measured values of the elastic constants. In Table 39-2 we give the observed values of the three elastic coefficients for several cubic crystals.* You will notice that C_{xxx} and C_{xyx} are, in general, not equal. This means σ that in metals like sodium and potassium the interatomic forces are not along the line joining. In other words, as we assumed in our model, Diamond does not work either, because the forces in carbon are equivalent forces and have some directional properties—the bonds would prefer to be at the tetrahedral angle. The cubic crystals like Lithium, Boron, and iron boride, and so on, do have nearly all the physical properties consistent in our model, and the table shows that the constants C_{xxx} and C_{xyx} are almost equal. It is not clear why silver sulfide should not satisfy the condition that $C_{xyx} = C_{xyz}$.

* In the literature you will often find that a different notation is used. For instance, people may write $C_{xxx} = C_{111}$, $C_{xyx} = C_{112}$, and $C_{xyz} = C_{113}$.

Table 39-2

Elastic Moduli of Cubic Crystals
in 10^{12} dyn/cm²

	C_{xxx}	C_{xyx}	C_{xyz}
Na	0.053	0.043	0.029
K	0.046	0.037	0.026
Be	2.37	1.41	1.13
Diamond	0.76	1.24	5.73
Al	1.00	0.62	0.24
LiF	1.12	0.54	0.33
NaCl	0.482	0.129	0.125
KCl	0.43	0.062	0.052
NaBr	0.33	0.13	0.13
KJ	0.27	0.072	0.042
AgCl	0.60	0.36	0.162

* J. R. C. Smart, *Introduction to Solid State Physics*, John Wiley and Sons, Inc., New York, 2nd ed., 1976, p. 97.

The Flow of Dry Water

40-1 Hydrostatics

The subject of the flow of fluids, and particularly of water, fascinates everybody. We can all remember, as children, playing in the bath tub or in mud puddles with the streams of water. As we get older, we watch streams, waterfalls, and whirlpools, and we are fascinated by the substance which seems so mysterious relative to solids. The chapter on fluid mechanics may very well profit a student using it as the subject of his chapter project. The action of a child trying to catch a stream flowing in the street and his surprise at the strange way the water works is very old, but it's amazing how our attempts over the years to understand the flow of fluids. We have tried to nail the water up in our understanding by getting the laws and the equations that describe the flow. We will describe these attempts in this chapter. In the next chapter, we will describe the unique way in which water has broken through the chain and escaped our attempts to understand it.

We suppose that the elementary properties of water are already known to you. The main property that distinguishes a fluid from a solid is that a fluid cannot maintain a shear stress for any length of time. If a shear is applied to a fluid, it will move under the shear. Thicker liquids like honey move less easily than fluids like air or water. The measure of the ease with which a fluid yields is its viscosity. In this chapter we will consider only situations in which the viscous effects can be ignored. The effects of viscosity will be taken up in the next chapter.

We begin by considering hydrostatics, the theory of liquids at rest. When liquids are at rest, there are no shear forces (even the viscous liquids). The law of hydrostatics, therefore, is that the stresses are always normal to any surface inside the fluid. The normal force per unit area is called the pressure. From the fact that there is no shear in a static fluid it follows that the pressure stress is the same in all directions (Fig. 40-1). We will let you entertain yourself by proving that if there is no shear on any plane in a fluid, the stress must be the same in any direction.

The pressure in a fluid may vary from place to place. For example, in a static fluid at the earth's surface the pressure will vary with depth, because of the weight of the fluid. If the density ρ of the fluid is considered constant, and if the pressure at some arbitrary zero level is called p_0 (Fig. 40-2), then the pressure at a height h above this point is $p = p_0 + \rho gh$, where g is the gravitational force per unit mass. The equation

$$p = p_0 + \rho gh$$

is, therefore, a constant in the static fluid. This relation is familiar to you, but we will now derive a more general result of which it is a special case.

If we take a small cube of water, what is the net force on it from the pressure? Since the pressure at any place is the same in all directions, there can be a net force per unit volume only because the pressure varies from one place to another. Suppose that the pressure is varying in the x -direction, and we make the coordinate directions parallel to the cube edges. The pressure on the face at x gives the force $p_x A_x$ (Fig. 40-3), and the pressure on the face at $x + Ax$ gives the force $p_{x+Ax} A_x$.

$[p_{x+Ax}/(p_x/A_x)] A_x \approx Ax$, so that the resultant force is $(\partial p/\partial x) Ax^2 p_x/2$. If we take the remaining pairs of faces of the cube, we easily see that the pressure force per unit volume is ∇p . If there are other forces in addition, such as gravity, then the pressure must balance them to give equilibrium.

40-1 Hydrostatics

40-2 The equations of motion

40-3 Steady flow—Bernoulli's theorem

40-4 Circulation

40-5 Vortex tubes

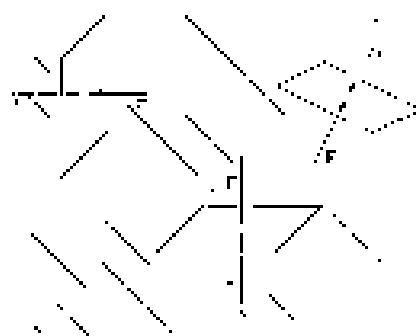


Fig. 40-1. In a static fluid the force per unit area across any surface is normal to the surface and is the same for all orientations of the surface.

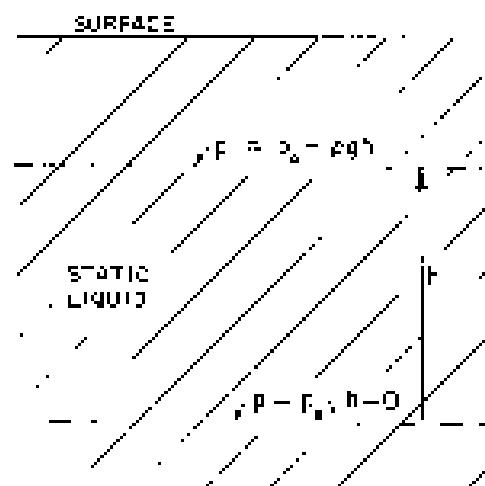


Fig. 40-2. The pressure in a static liquid.

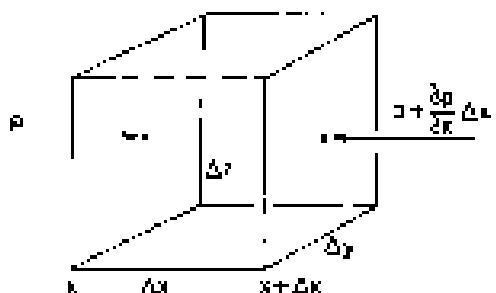


Fig. 40-3. The net pressure force on a cube $\sim = \nabla p$ per unit volume.

Let's take a circumstance in which such an additional force can be described by a potential energy, as would be true in the case of gravitation; we will let ϕ stand for the potential energy per unit mass. (For gravity, for instance, ϕ is just g .) The force per unit mass is given in terms of the potential by $-\nabla\phi$, and if ρ is the density of the fluid, the force per unit volume is $-\rho\nabla\phi$. For equilibrium this force per unit volume added to the pressure force per unit volume must give zero:

$$-\nabla p - \rho\nabla\phi \sim 0. \quad (40.1)$$

Equation (40.1) is the equation of hydrostatics. In general, it has no solution. If the density varies in space in an arbitrary way, there is no way for the forces to be in balance, and the fluid cannot be in static equilibrium. Convection currents will start up. We can see this from the equation since the pressure term is a pure gradient, whereas for variable ρ the other term is not. Only when ρ is a constant is the potential term a pure gradient. Then the equation has a solution:

$$\rho + \rho\phi = \text{const}$$

Another possibility which allows hydrostatic equilibrium is for ρ to be a function only of p . However, we will leave the subject of hydrostatics because it is not nearly so interesting as the situation when fluids are in motion.

40-2. The equations of motion

First, we will discuss fluid motions in a purely abstract, theoretical way and then consider special examples. To describe the motion of a fluid, we must give its properties at every point. For example, at different places, the water filet or call the fluid "water" is moving with different velocities. To specify the character of the flow, therefore, we must give the three components of velocity at every point and for any time. If we can find the equations that determine the velocity, then we would know how the liquid moves at all times. The velocity, however, is not the only property that the fluid has which varies from point to point. We have just discussed the variation of the pressure from point to point. And there are still other variables. There may also be a variation of density from point to point. In addition, the fluid may be a conductor and carry an electric current whose density varies from point to point, its magnitude and direction. There may be a temperature which varies from point to point, or a magnetic field, and so on. So a number of fields needed to describe the complex situation will depend on a few hydrodynamic problems. There are instabilities, phenomena when vortices and instabilities play a dominant role in determining the behavior of the fluid, the subject is called hydrodynamics, and great attention is being paid to it at the present time. However, we are not going to consider these more complicated situations because there are already interesting phenomena at a lower level of complexity, and even the more elementary level will be considered enough.

We will take the situation where there is no unipolar field, our no conductivity, and we will not worry about the temperature because we will suppose that the density and pressure determine the unique feature, the temperature at any point. As a matter of fact, we will reduce the complexity of our work by making the assumption that the density is a constant. We imagine that the fluid is essentially incompressible. Putting it another way, we are supposing that the variations of pressure are so small that the changes in density produced thereby are negligible. If that is not the case, we would encounter phenomena in addition to the ones we will be discussing here. For example, the propagation of sound or of shock waves. We have already discussed the propagation of sound and shocks to some extent, so we will now make our consideration of hydrodynamics from these other phenomena by making the approximation that the density ρ is a constant. It is easy to determine when the approximation of constant ρ is a good one. We can say that if the velocities of flow are much less than the speed of a sound wave in the fluid, we do not have to worry about variations in density. The escape that water makes in our attempt to understand it is not related to the approximation of

constant density. The equations that do permit the escape will be discussed in the next chapter.

In the zeroth theory of fluids one must begin with an equation of state for the fluid which connects the pressure to the density. In our approximation, this equation of state is simply

$$\rho = \text{const.}$$

This idea is the D'Alembert relation for one variable. The new relation expresses the conservation of mass. - if fluid flows away from a point, there must be a decrease in the amount left behind. If the fluid velocity is v_i , then the mass which flows in a unit time across a unit area of surface is the component of v normal to the surface. We have had a similar relation in electricity. We also know from electricity that the divergence of such a quantity gives the rate of decrease of the density per unit time. In the same way, the equation

$$\nabla \cdot (v \rho) = - \frac{\partial \rho}{\partial t} \quad (40.1)$$

expresses the conservation of mass for a fluid. (A third axiomatic equation of continuity.) In one representation, which is the most appropriate fluid approximation, ρ is a constant, and the equation of continuity is simply

$$\nabla \cdot v = 0. \quad (40.2)$$

The fluid velocity v like the magnet + light B has zero divergence. (The hydrodynamic variables are often closely analogous to the electrodynamic variables; that's why we studied electricity first. Some people argue the other way: they think that one should study hydrodynamics first so that it will be easier to understand electricity afterwards. But electrodynamics is really much easier than hydrodynamics.)

We will get our next equation from Newton's law which tells us how the velocity changes because of the forces. The mass of an element of volume of the fluid times its acceleration must be equal to the force on the element. Taking an element of unit volume, and writing the force per unit volume as f , we have

$$\rho \times (\text{acceleration}) \leftrightarrow f.$$

We will write the force density as the sum of three terms. We have already considered the pressure force per unit volume, $-p v$. Then there are the "external" forces which act at a distance—the gravity or electricity. When they are conservative forces with a potential per unit mass, ϕ , they give a force density $-\rho \nabla \phi$. (If the external forces are not conservative, we would have to write f_{ext} for the calculated force per unit volume.) Then there is another "internal" force per unit volume, which is due to the fact that in a flowing fluid there can also be a shearing stress. This is called the viscous force, which we will write f_{vis} . Our equation of motion is

$$v \times (\text{acceleration}) \leftrightarrow -\nabla p - \rho \nabla \phi + f_{vis}. \quad (40.3)$$

For this chapter we are going to suppose that the liquid is "thin" in the sense that its viscosity is negligible, so we will omit f_{vis} . When we drop the viscosity term, we will obtain a simple approximation, which describes some ideal stuff rather than real water. John von Neumann was well aware of the tremendous difference between what happens when you don't know the various terms and when you do, and he was also aware that, during most of the development of hydrodynamics until about 1930, the main interest was in solving useful nonviscous problems with the approximation which he almost nothing to do with, except, like the mathematician T. C. J. Stoker who wrote such articles as a man who should "say what?" Such articles have not an enormous majority of the time. It is because we are leaving this present set of our calculations in the chapter, that we have given it the title "The Flow of Dry Water." We are postponing a discussion of viscosity to the next chapter.

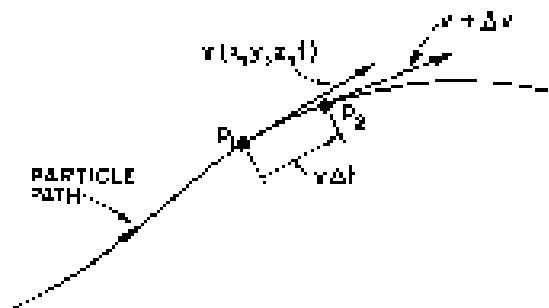


Fig. 40-4. The acceleration of a fluid particle.

If we leave out f_{ext} , we have in Eq. (40.4) everything we need except an expression for the acceleration. You might think that the formula for the acceleration of a fluid particle would be very simple, for it seems obvious that if v is the velocity of a fluid particle at some place in the fluid, the acceleration would just be $\partial v / \partial t$. It is not, and for a rather simple reason. The derivative $\partial v / \partial t$ is the rate at which the velocity $v(x, y, z, t)$ changes at a fixed point in space. What we need is how fast the velocity changes for a particular piece of fluid. Imagine that we mark one of the drops of water with a colored speck so we can watch it. In a small interval of time Δt , this drop will move to a different location. If the drop is moving along some path as sketched in Fig. 40-4, it might in Δt move from P_1 to P_2 . In fact, it will move in the x -direction by an amount Δx , in the y -direction by the amount Δy , and in the z -direction by the amount Δz . We see that, if $v(x, y, z, t)$ is the velocity of the fluid particle which is at (x, y, z) at the time t , then the velocity of the same particle at the time $t + \Delta t$ is given by $v(x + \Delta x, y + \Delta y, z + \Delta z, t + \Delta t) = v(t)$

$$\Delta x = v_x \Delta t, \quad \Delta y = v_y \Delta t, \quad \text{and} \quad \Delta z = v_z \Delta t.$$

From the definition of the partial derivatives—recall Eq. (2.7)—we have, to this order, that

$$\begin{aligned} \Delta v &= v(x + \Delta x, y + \Delta y, z + \Delta z, t + \Delta t) - v(x, y, z, t) \\ &= \Delta v(x, y, z, t) + \frac{\partial v}{\partial x} \Delta x + \frac{\partial v}{\partial y} \Delta y + \frac{\partial v}{\partial z} \Delta z + \frac{\partial v}{\partial t} \Delta t. \end{aligned}$$

The acceleration $\Delta v / \Delta t$ is

$$= \frac{\partial v}{\partial t} + v_x \frac{\partial v}{\partial x} + v_y \frac{\partial v}{\partial y} + v_z \frac{\partial v}{\partial z}.$$

We can write this symbolically—treating ∇ as a vector—as

$$(v \cdot \nabla)v + \frac{\partial v}{\partial t}. \quad (40.5)$$

Note that there can be an acceleration even though $\partial v / \partial t = 0$ so that velocity at a given point is not changing. As an example, water flowing in a circle at a constant speed is accelerating even though the velocity at a given point is not changing. This occurs, of course, when the velocity of a particular piece of water which is initially at one point on the circle has a different direction a moment later; there is a centripetal acceleration.

The rest of our theory is just mathematics. Finding solutions of the equation of motion we get by putting the acceleration (40.5) into Eq. (40.1). We get

$$\frac{dv}{dt} = (v \cdot \nabla)v + \frac{\partial v}{\partial t} + \nabla \cdot v, \quad (40.6)$$

where viscosity has been omitted. We can rearrange this equation by using the following identity from vector analysis:

$$(v \cdot \nabla)v = (\nabla \times v) \times v + \frac{1}{2}\nabla(v \cdot v).$$

If we now define a new vector field Ω as the curl of v ,

$$\Omega = \nabla \times v. \quad (40.7)$$

The vector identity can be written as

$$(\nabla \cdot \nabla) v = \Delta v - \frac{1}{2} \nabla v^2,$$

and our equation of motion, (40.6) becomes

$$\frac{\partial v}{\partial t} + \Omega \times v - \frac{1}{2} \nabla v^2 = -\frac{\nabla p}{\rho} - \nabla \phi. \quad (40.8)$$

You can verify that Eqs. (40.6) and (40.8) are equivalent by checking that the components of the two sides of the equation are equal—and making use of (40.7).

The vector field Ω is called the vorticity. If the vorticity is zero everywhere, we say that the flow is irrotational. We have already defined in Section 3-3 a thing called the circulation of a vector field. The circulation around any closed loop in a fluid is the line integral of the fluid velocity, at a given instant of time, around that loop:

$$(\text{Circulation}) = \oint_C v \cdot d\mathbf{s}$$

The circulation per unit area for an infinitesimal loop is then using Stokes' theorem—equal to $\nabla \times v$. So the vorticity Ω is the circulation around a unit area, perpendicular to the direction of Ω . It also follows that if you get a little piece of driftwood floating about, at any place in the liquid it will rotate with the angular velocity $\Omega/2$. Try to see if you can prove that. You can also check it out for a number of waves on a swimming pool; it is equal to twice the local tangential velocity of the water.

If we are interested only in the velocity field, we can eliminate the pressure from our equations. Taking the curl of both sides of Eq. (40.8), remembering that p is a constant and that the curl of any gradient is zero, and using Eq. (40.7), we get

$$\frac{\partial \Omega}{\partial t} + \nabla \times (\Omega \times v) = 0. \quad (40.9)$$

This equation, together with the equations

$$\Omega = \nabla \times v \quad (40.10)$$

and

$$\nabla \cdot v = 0, \quad (40.11)$$

describes completely the velocity field v . Mathematically speaking, if we know Ω at some time, then we know the curl of the velocity vector, and we also know that its divergence is zero, so given the physical situation we have all we need to determine v everywhere. (It is just like the situation in magnetism, where we had $\nabla \cdot B = 0$ and $\nabla \times B = \mu_0 i e^2$.) Thus, a given Ω determines v just as a given j determines B . Then, knowing v , Eq. (40.9) tells us how to change Ω from which we can get the new Ω for the next instant. Using Eq. (40.10), again we find the new v , and so on. You see how these equations reduce all fluid dynamics to calculating the flow. Note, however, that this procedure gives the velocity field only; we have lost all information about the pressure.

We point out one special consequence of our equations. If $\Omega = 0$ everywhere at any time t , v after that vanishes, so that Ω is still zero everywhere at $t + \Delta t$. We have a solution to the equations; the flow is permanently irrotational. If a flow was started with zero rotation, it would always have zero rotation. The equations to be solved then are

$$\nabla \cdot v = 0, \quad \nabla \times v = 0.$$

They are just like the equations for the electric or magnetic fields in free space. We will come back to them and look at some special problems later.

40-2 Steady flow—Bernoulli's theorem

Now we want to return to the equation of motion, Eq. (40.5), but limit ourselves to situations in which the flow is "steady." By steady flow we mean that at any one place in the fluid, the velocity never changes. The fluid at any point is always replaced by new fluid moving in exactly the same way. The velocity picture always looks the same—it is a static vector field. In the same way that we drew "field lines" in magnetostatics, we can now draw lines which are always tangent to the fluid velocity as shown in Fig. 40-5. These lines are called *streamlines*. For steady flow, they are exactly the actual paths of fluid particles. The steady flow line structure pattern changes in time, and the streamline pattern at any instant does not represent the path of a fluid particle.

A steady flow does not mean that nothing is happening; atoms in the fluid are moving and changing their velocities. It is only that $\partial u/\partial t = 0$. That is, if we take the dot product of $\partial u/\partial t$ with the equation of motion, the term $(\partial u/\partial t) \cdot u$ drops out, and we are left with

$$u \cdot \nabla \left[\frac{p}{\rho} + \phi + \frac{1}{2} u^2 \right] = 0. \quad (40.12)$$



Fig. 40-5. Streamlines in steady flow form.

The equation says that for a small displacement in the direction of the fluid velocity, the quantity inside the brackets doesn't change. Now in steady flow all components are along streamlines, so Eq. (40.12) tells us that for all the points along a streamline, we can write

$$\frac{p}{\rho} + \frac{1}{2} u^2 - \phi = \text{const (streamline).} \quad (40.13)$$

This is *Bernoulli's principle*. The constant may in general be different for different streamlines, and we know is that the left-hand side of Eq. (40.13) is the same all along a given streamline. Unfortunately, we may notice that for steady irrotational motion, for which $\phi = 0$, the equation of motion (40.8) gives us the relation

$$\nabla \left[\frac{p}{\rho} + \frac{1}{2} u^2 \right] = 0,$$

so that

$$\frac{p}{\rho} + \frac{1}{2} u^2 + \phi = \text{const (everywhere).} \quad (40.14)$$

It's just like Eq. (40.13) except that now the constant has the same value throughout the fluid.

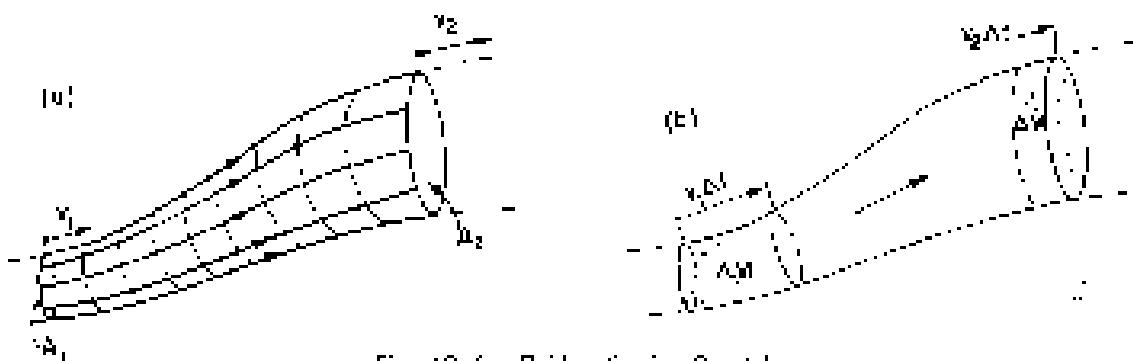


Fig. 40-6. Fluid motion in a flow tube.

The theorem of Bernoulli is in fact nothing more than a statement of the conservation of energy. A conservation law such as this gives us a lot of information about a flow without ever actually having to solve the equations of motion. Bernoulli's theorem is an important and simple idea that we would like to show you how it can be derived in a way that is different from the formal calculations we have just used. Imagine a bundle of adjacent streamlines which form a stream tube as sketched in Fig. 40-6. Since the walls of the tube consist of streamlines, no fluid flows out through the walls. Let's call the area at one end of the stream

mass ΔM , the fluid velocity there v_1 , the density of the fluid ρ_1 , and the potential energy ϕ_1 . At the other end of the tube, we have the corresponding quantities ρ_2 , v_2 , ϕ_2 , and ΔM . Now after a short interval of time Δt , the fluid at A_1 has moved a distance $v_1 \Delta t$, and the fluid at A_2 has moved a distance $v_2 \Delta t$ [Fig. 10-6(a)]. The conservation of mass requires that the mass which enters through A_1 must be equal to the mass which leaves through A_2 . These masses ΔM at the two ends must be the same:

$$\Delta M = \rho_1 A_1 v_1 \Delta t = \rho_2 A_2 v_2 \Delta t.$$

So we have the equality

$$\rho_1 A_1 v_1 = \rho_2 A_2 v_2 \quad (10.15)$$

This equation tells us that the velocity varies inversely with the area of the stream tube if c is constant.

Now we calculate the work done by the fluid pressure. The work done on the fluid entering at A_1 is $\rho_1 P_1 A_1 \Delta t$, and the work given up at A_2 is $\rho_2 P_2 A_2 \Delta t$. The net work per unit mass between A_1 and A_2 is, therefore,

$$\rho_2 P_2 A_2 \Delta t - \rho_1 P_1 A_1 \Delta t$$

which must equal the increase in the energy of a mass ΔM of fluid in going from A_1 to A_2 . In other words,

$$\rho_2 P_2 A_2 \Delta t - \rho_1 P_1 A_1 \Delta t = \Delta M(E_2 - E_1), \quad (10.16)$$

where E_1 is the energy per unit mass of fluid at A_1 , and E_2 is the energy per unit mass at A_2 . The energy per unit mass of the fluid can be written as

$$E = \frac{1}{2} v^2 + \phi + U,$$

where $\frac{1}{2} v^2$ is the kinetic energy per unit mass, ϕ is the potential energy per unit mass, and U is an additional term which represents the internal energy per unit mass of fluid. The internal energy might also be called the thermal energy in a compressible fluid, or mechanical energy in a solid. All these quantities are ΔM from point to point. Using this form for the energy in (10.16) we have

$$\frac{\rho_2 d(M) \Delta t}{\Delta M} = \frac{\rho_2 d(v_2) \Delta t}{\Delta M} = \frac{1}{2} (v_2^2 - v_1^2) + \phi_2 - \phi_1 + U_2 - U_1 = C,$$

But we have seen that $\Delta M = \rho A \Delta t$, so we get

$$\frac{\rho_2}{\rho_1} + \frac{1}{2} \frac{v_2^2}{v_1^2} - \phi_1 + U_1 = \frac{\rho_2}{\rho_1} + \frac{1}{2} \frac{P_2^2}{P_1^2} + \phi_2 - U_2, \quad (10.17)$$

which is the Bernoulli result with an additional term for the internal energy. If the fluid is incompressible, the internal energy term is the same on both sides, and we see again that Eq. (10.14) holds along any streamline.

We consider now some simple examples in which the Bernoulli integral gives us a description of the flow. Suppose we have water flowing out of a hole near the bottom of a tank, as drawn in Fig. 10-7. We take a situation in which the flow speed v_{top} at the hole is much larger than the flow speed near the top of the tank. In other words, we imagine that the diameter of the tank is so large that we can neglect the drop in the liquid level. (We could make a more accurate calculation if we wished.) At the top of the tank the pressure is P_0 , the atmospheric pressure, and the pressure at the sites of the jet is also P_0 . Now we write an ordinary equation for a streamline, such as the one shown in the figure. At the top of the tank, we take ϕ equal to zero and we also take the gravity potential ϕ to be zero. At the open end, and $\phi = -gh$, so that

$$P_0 = P_0 - \frac{1}{2} \rho v_{top}^2 - \rho g h,$$

or

$$v_{top} = \sqrt{\frac{2 P_0}{\rho}}. \quad (10.18)$$

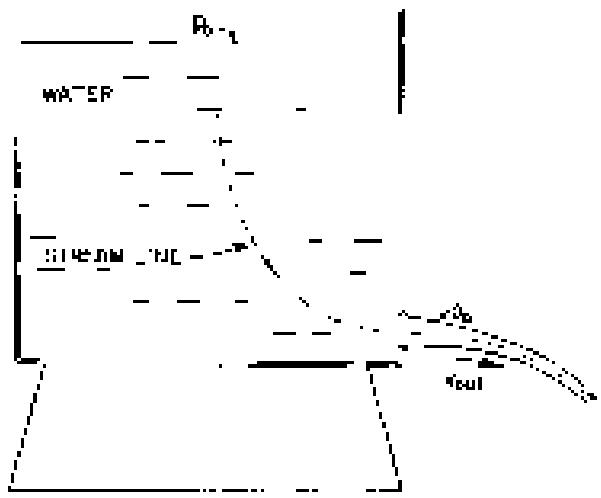


Fig. 40-7. Flow from a tank.

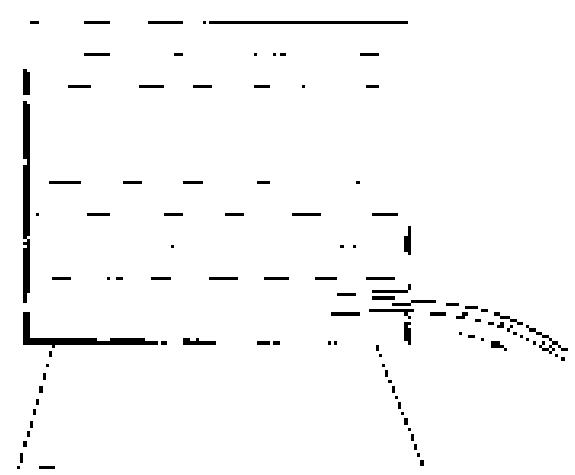


Fig. 40-8. With a constricted discharge tube, the stream contracts to maintain the area of the opening.

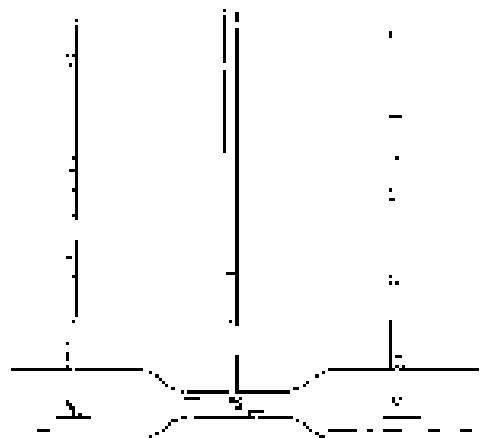


Fig. 40-9. The pressure is lowest where the velocity is highest.

This velocity is just what we would get for something which falls the distance h . It is not too surprising, since the water at the top gains kinetic energy at the expense of the potential energy of the water at the top. Do not get the idea, however, that you can figure out the rate that the fluid flows out of the tank by multiplying this velocity by the area of the hole. The fluid velocities as the jet leaves the hole are not all parallel to each other but have components directed toward the center of the stream—the jet is converging. After the jet has gone a little way, the contraction stops and the velocities do become parallel. So the total flow is the velocity times the area of the jet. In fact, if we have a discharged opening which is just a round hole with a sharp edge, the jet contracts to 62 percent of the area of the hole. The reduced effective area of the discharge varies for different shapes of discharge tubes, and experimental contractions are available as tables of efflux coefficients.

If the discharge tube is constricted, as shown in Fig. 40-8, it is possible to prove in a most beautiful way that the efflux coefficient is exactly 50 percent. We will give you a hint of how the proof goes. We have used the measureless of energy to get the velocity, Eq. (40-18), but there is also a measure of conservation to consider. Since there is an outflow of momentum in the discharge jet, there must be a force applied over the cross section of the discharge tube. Where does the force come from? The force must come from the pressure on the walls. As long as the efflux hole is an all-around flow free of walls, the fluid velocity and the walls of the tank will be very small. Therefore the pressure on every fiber is almost exactly the same as the static pressure for a fluid at rest—*from* Eq. (30-14). Then the static pressure at any point on the side of the tank must be matched by the static pressure at the point on the opposite wall, except at the points on the wall opposite the discharge tube. If we calculate the static pressure exerted on through the jet by the pressure, we can show that the efflux coefficient is 1/2. We cannot use this method for a discharge tube like that shown in Fig. 40-9, however, because the velocity increases along the wall right near the discharge area gives a pressure fall which we are unable to eliminate.

Let's look at another example—a horizontal pipe with changing cross section, as shown in Fig. 40-9, with water flowing in one end and out the other. The law of conservation of energy, namely Bernoulli's formula, says that the pressure is lower in the constricted part where the velocity is higher. We can easily demonstrate this effect by measuring the pressure at different areas sections with small vertical columns of water attached to the flow tube through holes small enough so that they do not disturb the flow. The pressure is then measured by the height of water in these vertical columns. The pressure is found to be less at the constriction than it is on either side. If the area beyond the constriction comes back to the same value it had before the constriction, the pressure rises again.

Bernoulli's formula would predict that the pressure downstream of the constriction should be the same as it was upstream, but actually it is noticeably less. The reason that our prediction is wrong is that we have neglected the frictional viscous forces which cause a pressure drop along the tube. Despite this pressure drop the pressure is definitely lower at the constriction. This is because the water goes faster than it is on either side of it, as predicted by Bernoulli. The speed must certainly exceed v_1 to get the same amount of work through the constriction, since ΔE is zero. So the water accelerates as it goes from the wide to the narrow part. The force that gives this acceleration comes from the drop in pressure.

We can check our results with another simple demonstration. Suppose we have an orifice discharge嘴 which throws a jet of water outward as shown in Fig. 40-10. If the efflux velocity were exactly $\sqrt{2gh}$, the distance x from the mouth to a level even with the surface of the water in the tank, experimentally, it fails somewhat short. Our prediction is roughly right, but again viscosity friction which has not been included in our energy conservation formula has resulted in a loss of energy.

Have you ever held two pieces of paper close together and tried to blow them apart? Try it! They come together. The reason, of course, is that the air has a higher speed going through the restricted space between the sheets than it does when it gets outside. The pressure between the sheets is lower than atmospheric pressure, so they come together rather than separating.

40-4 Circulation

We saw at the beginning of the last section that if we want to move up past the fluid with no circulation, the flow satisfies the following two equations:

$$\nabla \cdot v = 0, \quad \nabla \times v = 0. \quad (40.29)$$

They are the same as the equations of electrodynamics or magnetostatics in empty space. The divergence of the electric field is zero when there are no charges, and the curl of the electrostatic field is always zero. The curl of the magnetic field is zero if there are no currents, and the divergence of the magnetic field is always zero. Therefore, Eqs. (40.2) have the same solutions as the equations for E in electrodynamics or for B in magnetostatics. As a matter of fact, we have already solved the problem of the flow of a fluid past a sphere, as an electrostatic analogy, in Section 2-5. The electrostatic analog is a uniform electric field with a dipole field. The dipole field is so adjusted that the flow velocity normal to the surface of the sphere is zero. The same problem for the flow past a cylinder can be worked out in a similar way by using a suitable line dipole with a uniform flow field. This solution holds for a situation in which the field velocity at large distances is constant—both in magnitude and direction. The solution is sketched in Fig. 40-11(a).

There is another solution for the flow around a cylinder when the conditions are such that the fluid at large distances moves in circles around the cylinder. The flow is, then, circular everywhere, as in Fig. 40-11(b). Such a flow has a circulation around the cylinder, although $\nabla \times v$ is still zero in the fluid. How can there be circulation without a curl? We have a circulation around the cylinder because the line integral of v around any long cylinder the cylinder is not zero. At the same time, the line integral of v around any closed path which does not include the cylinder is zero. We saw the same thing when we found the magnetic field around a wire. The curl of B was zero outside of the wire, although it has a stepup in its rotational circulation around a cylinder as precisely the same as the magnetic field around a wire. For a circular path with its center at the center of the cylinder, the line integral of the velocity is

$$\oint v \cdot d\ell = 2\pi rv.$$

For inviscid flow the integral must be independent of r . Let's add the constant

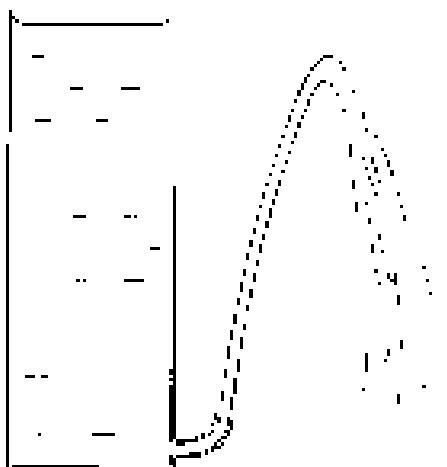


Fig. 40-10. Orifice discharge嘴. $h \approx \sqrt{2gh}$.

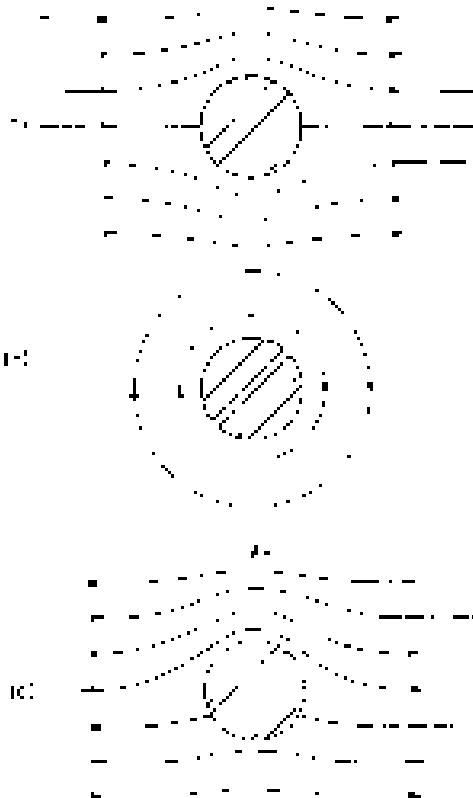


Fig. 40-11. (a) Inv. fluid flow past a cylinder. (b) Circular flow around a cylinder. (c) The superposition of (a) and (b).

value C , then we have that:

$$v = \frac{C}{2\pi r} r \quad (40.20)$$

where v is the tangential velocity, and r is the distance from the axis.

This is a nice demonstration of a fluid circulating around a hole. You take a transparent cylindrical tank with a hole near the center of the bottom. You fill it with water, stir up some circulation with a stick, seal plug the down plug. You get the pretty effect shown in Fig. 40-12. (You can add a similar though less dramatic effect by adding your hand, from just beginning, it strengthens because of viscosity and the flow becomes turbulent.) Although you put no net force on the cylinder, it moves, so there is net circulation, it rotates clockwise because of viscosity and the flow becomes turbulent, although still with some circulation around the hole.

From the theory, we can calculate the shape of the free surface of the water. As a particle of the water moves toward it picks up speed. From Eq. (40.20) the tangential velocity goes as $1/r$. It's just like the conservation of angular momentum, like the skater picking up more. Also the radial velocity goes as $1/r$. Ignoring the tangential motion, we have water going radially toward toward a hole; from $\nabla \cdot v = 0$, it follows that v is radial, so v_r is proportional to $1/r$. So the radial velocity also increases as $1/r$, and the water goes in along Archimedes spirals. The air/water surface is all at atmospheric pressure, so it must move. From Eq. (40.14), the property that

$$gx + \frac{1}{2} \rho v^2 = \text{const.}$$

But v is proportional to $1/r$, so the shape of the surface is

$$(z - z_0) = \frac{x}{r_0}$$

An interesting point, which *most* people don't know but is true for incompressible, irrotational flow, is that if we have one soliton and a second soliton, then the sum is also a soliton. This is true because the equations in (40.19) are linear. The complete equations of hydrodynamics, Eqs. (40.8), (40.9), and (40.10), are not linear, which makes a vast difference. For the irrotational flow about the cylinder, however, we can superpose the flow of Fig. 40-11(c) on the flow of Fig. 40-11(b) and get the new flow pattern shown in Fig. 40-11(d). This flow is of special interest. The flow velocity is higher on the upper side of the cylinder than on the lower side. The pressures are therefore lower on the upper side than on the lower side. So when we have a combination of a circulation around a cylinder and a net horizontal flow, there is a net reduced force on the cylinder—it is called a lift force. Of course, if there's no circulation, there is no net force on any body according to our theory of "dry" water.

40-5. Vector Equations

We have already written down the pair of equations for the flow of an incompressible fluid with zero body resistance. They are

$$\text{I. } \nabla \cdot v = 0,$$

$$\text{II. } \mathbf{Q} = \nabla \times \mathbf{v},$$

$$\text{III. } \frac{\partial \mathbf{Q}}{\partial t} + \nabla \times (\mathbf{Q} \times \mathbf{v}) = 0.$$

The physical content of these equations has been described in words by Helmholtz in terms of three theorems. First, imagine that in the fluid we were to draw some lines other than streamlines. By vector lines we mean field lines that have the direction of \mathbf{Q} and have a density in any region proportional to the magnitude of \mathbf{Q} . From II the divergence of \mathbf{Q} is always zero (see end of Section 2-7) that the divergence of a curl is always zero. So vector lines are like lines of \mathbf{Q} . They never start or stop, and will tend to go in closed loops. Now Helmholtz described (in 40-10)

In words, by the following statement: the vortex lines move with the fluid. This means that if you were to mark the fluid particles along some vortex lines by coloring them with ink, for example, then as the fluid moves and carries those particles along, they will always mark the new positions of the vortex lines. In whatever way the atoms of the liquid move, the vortex lines move with them. That is one way to describe the flow.

It also suggests a method for solving any problems. Given the initial flow pattern—say v everywhere—then you can calculate Ω . From the Ω you can also tell where the vortex lines are going to be at the later time t , they move with the speed v . With the new Ω you can use Eqs. 11 to find the new v . (That's just like the problem of finding Ω given the current.) If we are given the flow pattern at one instant we can in principle calculate it for all subsequent times. We have the general solution for nonviscous flow:

We would like to show how Helmholtz's statement—and therefore III—can be at least partly understood. It is really just the law of conservation of angular momentum applied to the fluid. Suppose we imagine a small cylinder whose axis is parallel to the vortex lines, as in Fig. 40-13(a). At some time later, this same piece of fluid will be somewhere else. Generally it will occupy a cylinder with a different diameter and be in a different place. It may also have a different orientation, say as in Fig. 40-13(b). If the diameter has changed, however, the length will have increased to keep the volume constant. (Since we are assuming an incompressible fluid.) Also, since the vortex lines are stuck with the material, their density will go up as the cross-sectional area goes down. The product of the vorticity Ω and area A of the cylinder will remain constant, so according to Bernoulli, we should have

$$\Omega_1 A_1 = \Omega_2 A_2 \quad (40.21)$$

Now just as Ω will zero viscously all the forces on the surface of the cylindrical volume (or any volume, for that matter) are perpendicular to the surface. The pressure force can cause the volume to expand from point to point, or can cause it to change shape; but with no tangential forces, the orientation of the exterior momentum of the material inside stays constant. The angular momentum of the fluid in the little cylinder is its moment of inertia I times the angular velocity of the liquid, which is proportional to the vorticity Ω . But a cylinder, the moment of inertia is proportional to $m r^2$. So from the equality for the angle of orientation, we would conclude that

$$(M_1 R_1^2) \Omega_1 = (M_2 R_2^2) \Omega_2.$$

But the mass is the same, $M_1 = M_2$, and the areas are proportional to R^2 , so we get again just Eq. (40.21). Bernoulli's statement—which is equivalent to III—is just a consequence of the fact that in the absence of viscosity the angular momentum of an element of the fluid cannot change.

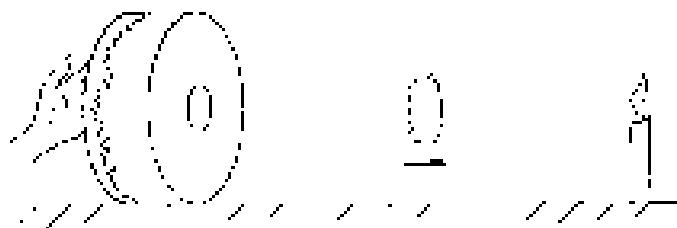


Fig. 40-14. Making a travelling vortex ring.

There is a nice demonstration of a moving vortex which is made with the simple apparatus of Fig. 40-14. It is a "drum" two feet in diameter and two feet high made by stretching a thick rubber sheet over the open end of a cylindrical "box." The "bottom"—the drum—is typed on its side. It is solid except for a 3-inch diameter hole. If you give a sharp kick on the rubber drumogen with your hand, a vortex ring is projected out of the hole. Although the vortex is invisible, you can tell it's there because it will blow out a candle 10 or 20 feet away. By the delay in

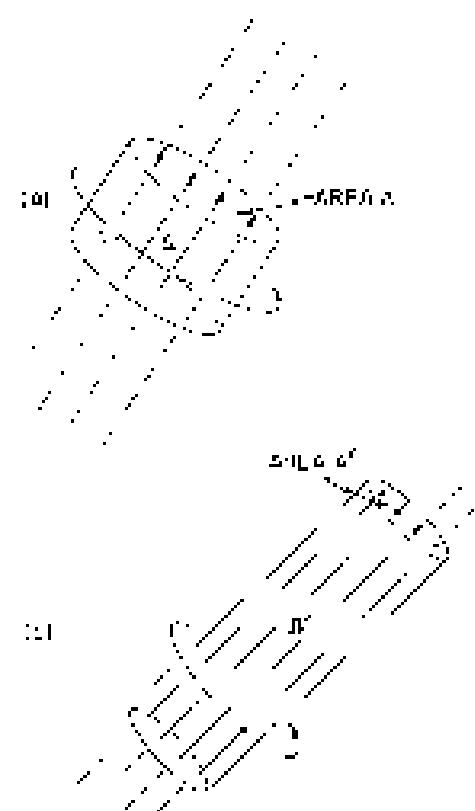


Fig. 40-13. (a) A group of vortex lines at t ; (b) the same lines at a later time t' .

the effect you can tell that "something" is traveling at a linear speed. You can see better what is going on if you look now at the other pictures. Then you see the vortex as a beautiful round "smoke ring."

The smoke ring is a torus-shaped bundle of vortex lines, as shown in Fig. 40-15(a). Since $\mathbf{B} = -\nabla \times \mathbf{v}$, these vortex lines represent, as a cross section of \mathbf{B} as shown in part (b) of the figure. We can understand the forward motion of the ring in the following way: The circulating velocity around the center of the ring depends on the size of the ring, having been a known function. Since the lines of \mathbf{B} curve with the fluid, they also move there with the velocity v_r . Consequently, the circulation is increased in one part of the ring, as a consequence of the forward motion of the center line of the bubble.

We must now understand something differently. We have already stated that Eq. (40-8) says that $\mathbf{v} \cdot \mathbf{B} = 0$ for the flow \mathbf{v} to be irrotational. This just like Eq. (39-6) is one of the three "no-slip" rules. Let us, for a moment, consider \mathbf{B} as zero and analyze what it is impossible to produce by "rotating" under "no-slip" conditions. Yet, in the simple example given with Fig. 40-15, we can produce a vortex ring starting with $\mathbf{B} = 0$, which we typically call "no-slip." (Remember, $\mathbf{B} = 0$ everywhere in the background except at the "ring," we will note that we can ignore some subtlety in a later chapter.) Coming from your theory of "no-slip" wedge problems complete non-slipping of \mathbf{v} is to be expected at a point.

Another feature of the "no-slip" theory which is incorrect is the superimposition we make regarding the flow at the boundary between fluid and the surface of a solid. When we discussed the flow past a cylinder, as in Fig. 40-14, for example, we permitted the fluid to slide along the surface of the solid. In such theory, the velocity at a solid surface could have any value depending on how it got there, and we did not consider any "friction" between the fluid and the solid. It is an experimental fact, however, that the velocity of real fluid always goes to zero at the surface of a solid object. Therefore, our solution for the cylinder, with or without circulation, is wrong. As is our view regarding the generation of vorticity. We will tell you about the more accurate theories in the next chapter.

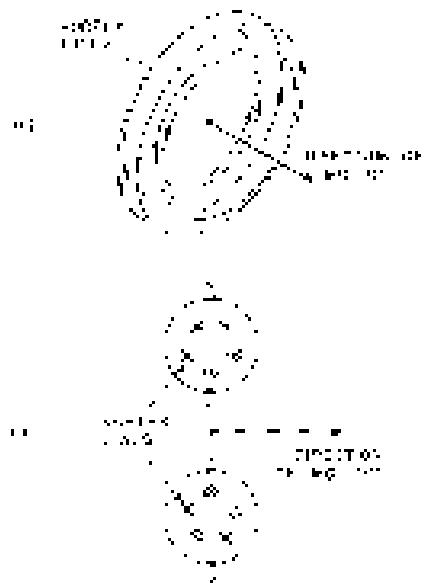


Fig. 40-15. A moving vortex ring from above (a), and the vortex lines (b) a cross-section of the ring.

The Flow of Viscous Water

41-1 Viscosity

In the last chapter we discussed the behavior of water, disregarding the phenomena of viscosity. Now we would like to discuss the phenomena of the flow of fluids, including the effects of viscosity. We want to look at the *real behavior* of fluids. We will describe qualitatively the normal behavior of the fluids under various different circumstances so that you will get some feel for the subject. Although you will see some complicated equations and hear about some complicated things, it is not our purpose that you should learn all these things. This is, in a sense, a "cultural" chapter which will give you some idea of the way the world is. There is only one item which is worth learning, and that is the simple definition of viscosity which we will come to in a moment. The rest is only for your entertainment.

In the last chapter we found that the laws of motion of a fluid are contained in the equation

$$\frac{d\mathbf{v}}{dt} - (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{\nabla P}{\rho} - \mathbf{F}_T - \frac{f_{visc}}{\rho} \mathbf{v}. \quad (41.1)$$

In our "dry" water pipe experiment we left out the last term, so we were neglecting all viscous effects. Also, we sometimes made an additional approximation by considering the fluid as *incompressible*; then we had the additional equation

$$\nabla \cdot \mathbf{v} = 0.$$

This last approximation is often quite good—particularly when flow speeds are much slower than the speed of sound. But in real liquids it is almost never true that we can neglect the internal friction that we call viscosity; most of the interesting things that happen come from it in one way or another. For example, we saw that in "dry" water the circulation never changes—if there is none to start out with, there will never be any. Yet, circulation in fluids is an everyday occurrence. We must dig up this theory.

We begin with an important experimental fact. When we work out the flow of "dry" water around a post or cylinder—the so-called "potential flow"—we had no reason not to permit the water to have a velocity tangent to the surface, only a normal component had to be zero. We took no account of the possibility that there might be a shear force between the liquid and the solid. It turns out, although it is not at all self-evident,—Let's say all circumstances where it has been experimentally checked, the velocity of a fluid is *strictly zero* on the surface of a solid. You have probably, no doubt, felt the place where you will collect a "long train of dust"—and that is still there after the bus has been cleaned, up the side. You have seen the same effect even on the great tail of a comet to trail. Why isn't the dust blown off the air? In spite of the fact that the particles, moving at high speed through the air, the speed of the air relative to the bus, always goes to zero right at the surface. So the very *strange* fact that particles are not disturbed.⁷ We must modify the theory to agree with the experiments that that in all ordinary fluids, the molecules next to a solid surface have zero velocity (relative to the surface).†

⁷ You know that larger dust particles than a table top, but not the very finest ones. The large ones stick up into the air.

[†] You can imagine circumstances when it is not true: glass is *exceptionally* a "solid," but it can certainly do much damage to a steel surface. So our assertion must break down somewhere.

41-2 Viscosity

41-2.1 Poiseuille flow

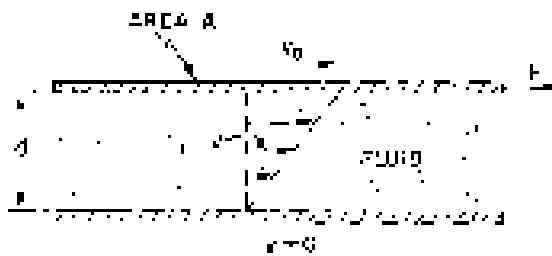
41-2.2 The Reynolds number

41-2.3 Flow past a circular cylinder

41-2.4 The limit of zero viscosity

41-2.5 Couette flow

Fig. 41-1. Viscous drag between two parallel plates.



We originally characterized a liquid by the fact that if you put a shearing stress on it, it would flow until it would give way. In static situations, there are no shear stresses. But before separation is reached—as long as you still push on it—there can be shear forces. This is evidence for those shear forces which exist in a moving fluid. To get a measure of the shear forces during the motion of a fluid, we consider the following kind of experiment. Suppose that we have two solid plane surfaces with water between them, as in Fig. 41-1, and we keep one stationary while moving the other parallel to it at the slow speed v_0 . If you measure the force required to keep the upper plate moving, you find that it is proportional to the area of the plates and to v_0/d , where d is the distance between the plates. So the shear stress σ_{xy} is proportional to v_0/d :

$$\frac{\sigma_{xy}}{v_0} = \eta \frac{1}{d},$$

The constant of proportionality η is called the coefficient of viscosity.

In a less cumbersome notation, we can always consider a line. The rectangular cell in the sense with its base dx is shown in Fig. 41-2. The shear force across the cell is given by

$$\frac{\Delta F}{\Delta x} = \eta \frac{\Delta u}{\Delta y} = \eta \frac{u_y}{dy}. \quad (41-2)$$

Note, du/dy , is the "rate of change" of the shear strain we defined in Chapter 28, so for a liquid, the shear stress is proportional to the rate of change of the shear strain.

In the general case we write

$$\sigma_{xy} = \eta \left(\frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x} \right). \quad (41-3)$$

If there is a uniform rotation of the fluid, du/dy is the negative of du/dx , and $\sigma_{xy} = \eta u_x$ —that is, since there are no stresses in a uniformly rotating fluid. (We did a similar thing in deriving σ_{xx} in Chapter 39.) There are, of course, the corresponding expressions for σ_{yy} and σ_{xy} .

As an example of the application of these ideas, we consider the problem of a fluid between two coaxial cylinders. I.e., the inner one has the radius a and the peripheral velocity v_a , and let the outer one have radius b and velocity v_b . See Fig. 41-3. We might ask, what is the velocity distribution between the cylinders? To answer this question, we begin by finding a formula for the viscous shear in the fluid at a distance r from the axis. From the symmetry of the problem, we can assume that the flow is always tangential and that its magnitude decreases only as $v = v(r)$. If we watch a speck in the water at radius r , its motion must be a function of time t ,

$$x = r \cos \omega t, \quad y = r \sin \omega t,$$

where $\omega = v/r$. Then the x - and y -components of velocity are

$$v_x = -r \omega \sin \omega t = -\omega r \quad \text{and} \quad v_y = r \omega \cos \omega t = \omega r. \quad (41-4)$$

From Eq. (41-3), we have

$$\sigma_{xy} = \eta \left[\frac{\partial}{\partial r} (v_x) + \frac{\partial}{\partial y} (v_y) \right] = \eta \left[\frac{\partial v_x}{\partial r} + \frac{\partial v_y}{\partial y} \right]. \quad (41-5)$$

For a point at $y = 0$, $\partial \omega / \partial r = 0$, and x would be the same as $r \partial \omega / \partial r$. So at that point

$$(S_{xy})_{r=0} = \sigma \frac{\partial \omega}{\partial r}. \quad (4.1.6)$$

(It is remarkable that λ should depend on $\partial \omega / \partial r$; when there is no change in ω with r , the liquid is in uniform rotation and there are no stresses.)

The stress we have calculated is the tangential shear which is the same all around the cylinder. We can get the shear acting across a radial slice from the equation for τ by multiplying the shear stress by the distance from r and the area term. We get

$$\tau = 2\pi r^2 (S_{xy})_{r=a} = 2\pi \rho r^2 \frac{d\omega}{dr}. \quad (4.1.7)$$

Since the rotation of the water is steady, there is no angular acceleration—the net torque on the cylindrical shell of water between r and $r + dr$ must be zero; that is, the moment τ must be balanced by an equal and opposite torque at $r + dr$, so τ must be independent of r . In other words, $d\omega/dr$ is equal to some constant, say A , and

$$\frac{d\omega}{dr} = \frac{A}{r^2}. \quad (4.1.8)$$

Integrating, we find that ω varies with r as

$$\omega(r) = \frac{A}{2r^2} + B. \quad (4.1.9)$$

The constants A and B are to be determined to fit the condition that $\omega = \omega_0$ at $r = a$, and $\omega = 0$ as $r \rightarrow b$. We get that

$$\begin{aligned} A &= \frac{2a^2b^2}{b^2 - a^2} (\omega_0 - \omega_b), \\ B &= \frac{b^2\omega_b - a^2\omega_0}{b^2 - a^2}. \end{aligned} \quad (4.1.10)$$

So we know ω as a function of r , and from $\tau = r\omega$,

If we want the torque, we can get it from Eqs. (4.1.7) and (4.1.9):

$$\tau = 2\pi a^4 A$$

or

$$\tau = \frac{4\pi a^2 b^2}{b^2 - a^2} (\omega_0 - \omega_b). \quad (4.1.11)$$

It is proportional to the relative angular velocities of the two cylinders. One standard apparatus for measuring the coefficient of viscosity is built this way—the cylinder—say the outer one—is on pivots but is held stationary by a spring balance which measures the torque on it, while the inner one is rotated at a constant angular velocity. The coefficient of viscosity is then determined from Eq. (4.1.11).

From its definition, you see that the units of τ are newton-meters/m². For water at 20°C,

$$\tau = 10^8 \text{ newton-meters/m}^2.$$

It is usually more convenient to use the specific viscosity, which is η divided by the density ρ . The values for water and air are then comparable:

$$\begin{aligned} \text{water at } 20^\circ\text{C}, \quad \eta/\rho &= 10^{-4} \text{ m}^2/\text{sec}, \\ \text{air at } 20^\circ\text{C}, \quad \eta/\rho &= 1.5 \times 10^{-6} \text{ m}^2/\text{sec}. \end{aligned} \quad (4.1.12)$$

Viscosities usually depend strongly on temperature. For instance, for water just below the freezing point, η/ρ is 1.5 times larger than it is at 20°C.

41-2 Viscous flow

We now go to a general theory of viscous flow – at least in the most general form known to man. We already understand that the shear stress components are proportional to the spatial derivatives of the various velocity components such as $\partial v / \partial x_1$, or $\partial w / \partial x_2$. However, in the general case of a compressible fluid there is another term in the stress which depends on other derivatives of the velocity. The general expression is

$$S_{ij} = \mu \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) + \gamma' \delta_{ij} (\nabla \cdot v), \quad (41.1)$$

where v_i is any one of the rectangular coordinates x_1 , x_2 , or x_3 , and x_j is any one of the rectangular coordinates of the velocity v . (The symbol δ_{ij} is the Kronecker delta which is 1 when $i = j$ and 0 for $i \neq j$.) The additional term at δ_{ij} is to add the diagonal elements S_{ii} of the stress tensor. If the liquid is incompressible $\nabla \cdot v = 0$, and this extra term doesn't appear. So it has to do with internal forces during compression. So two constants are required to describe the liquid, just like we had two constants to describe a homogeneous elastic solid. The constant μ is the "ordinary" coefficient of viscosity which we have already encountered. It is also called the first coefficient of viscosity or the "shear viscosity coefficient," and the new coefficient γ' is called the second coefficient of viscosity.

Now we want to determine the viscous force per unit volume f_{visc} , so we can substitute Eq. (41.1) to get the equation of motion for a real fluid. The force on a small cubical volume element of a fluid is the resultant of the forces on all the six faces. Taking them two at a time, we will get differences that depend on the derivatives of the stresses, and, therefore, on the second derivatives of the velocity. This is nice because it will get us back to a vector equation. The component of the viscous force per unit volume in the direction of the rectangular coordinate x_i is

$$\begin{aligned} f_{visc,i} &= \sum_{j=1}^3 \frac{\partial S_{ij}}{\partial x_j} \\ &= \gamma' \sum_{j=1}^3 \frac{\partial}{\partial x_j} \left[\mu \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) \right] + \frac{\partial}{\partial x_i} (\mu \nabla \cdot v). \end{aligned} \quad (41.4)$$

Usually, the viscosity and the viscosity coefficients with respect to $\nabla \cdot v$ is not simplified to just $\gamma' \nabla^2 v$. Then, the viscous force per unit volume becomes $\gamma' \nabla^2 v$ plus the effect of the divergence of the velocity. We saw in Chapter 39 that a second derivative of v is called the Laplacian $\nabla^2 v = \nabla^2 v$, and a term in the gradient of divergence $(\nabla \nabla \cdot v)$. Equation (41.4) is now much easier with the coefficients μ and $(\gamma' + \nu')$. We get

$$f_{visc,i} = \gamma' \nabla^2 v_i - (\nu' + \gamma') \nabla (\nabla \cdot v), \quad (41.5)$$

In the incompressible case, $\nabla \cdot v = 0$, and the viscous force per unit volume is just $\gamma' \nabla^2 v$. That is all that many people use; however, if you should want to calculate the absorption of sound in a fluid, you would need the second term.

We can now complete our general equation of motion for a real fluid. Substituting Eq. (41.5) into Eq. (41.1), we get

$$\rho \frac{\partial v_i}{\partial t} + \rho (\nabla \cdot v) v_i - \nu' \nabla p - \rho \nabla \phi = \gamma' \nabla^2 v_i - (\nu' + \gamma') \nabla (\nabla \cdot v)$$

We complicated. But that's the way nature is.

If we introduce the velocity $\Omega = \nabla \times v$, as we did before, we can write our equation as

$$\begin{aligned} \rho \left[\frac{\partial v_i}{\partial t} + \Omega \times v + \frac{1}{2} \nabla^2 v_i \right] &= -\nabla p - \rho \nabla \phi - \gamma' \nabla^2 v_i \\ &= -\nu' v - \gamma' \nabla (\nabla \cdot v). \end{aligned} \quad (41.15)$$

We are supposing again that the only body forces acting are conservative forces like gravity. To see what the new term means, let's look at the incompressible fluid case. Then, if we take the curl of Eq. (41.16), we get

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla \times (\mathbf{v} \times \mathbf{B}) = \frac{\eta}{\rho} \nabla^2 \mathbf{v}. \quad (41.17)$$

This is like Eq. (40.9) except for the new term on the right-hand side. When the right-hand side was zero, we had the Helmholtz theorem that the vorticity stays with the fluid. Now, we have the rather complicated nonzero term on the right-hand side which, however, has straightforward physical consequences. If we disregard for the moment the term $\nabla \times (\mathbf{v} \times \mathbf{B})$, we have a diffusion equation. The new term means that the vorticity \mathbf{v} diffuses through the fluid. If there is a "vortex gradient," in the vorticity \mathbf{v} , it will spread out into the neighboring fluid.

This is the term that causes the smoke ring to get thicker as it goes along. Also, it shows up nicely if you send a "clear" vortex (a "smokeless" ring) made by the apparatus described in the last chapter through a cloud of smoke. When it comes out of the cloud, it will have picked up some smoke, and you will see a hollow shell of a smoke ring. Some of the \mathbf{B} will pass outward into the smoke, while still maintaining its forward motion with the vortex.

41-3 The Reynolds number

We will now clear up the changes which are made in the character of fluid flow as a consequence of the new viscosity term. We will look at two problems in some detail. The first of these is the flow of a fluid past a cylinder—a flow which we tried to calculate in the previous chapter using the theory for nonviscous flow. It turns out that the viscous equations can be solved by hand only for a few special cases. So some of what we will tell you is based on experimental measurements, assuming that the experimental model satisfies Eq. (41.17).

The mathematical problem is this: We would like the solution for the flow of an incompressible, viscous fluid past a long cylinder of diameter D . The flow should be given by Eq. (41.17) and by

$$\mathbf{B} = \nabla \times \mathbf{v} \quad (41.18)$$

with the conditions that the velocity at large distances is some constant velocity, say V (parallel to the x -axis), and at the surface of the cylinder is zero. That is,

$$v_x = v_y = 0 \quad (41.19)$$

so

$$x^2 + y^2 = \frac{D^2}{4}.$$

That specifies completely the mathematical problem.

If you look at the equations, you see that there are four different parameters in the problem: η , ρ , D , and V . You might think that we would have to give a whole series of cases for different V 's, different D 's, and so on. However, that is not the case. All the different possible solutions correspond to different values of one parameter. This is the most important general thing we can say about viscous flow. To see why this is so, notice first that the viscosity and density appear only in the ratio η/ρ , the dynamic viscosity. That reduces the number of independent parameters to three. Now suppose we measure all distances in the only length that appears in the problem, the diameter D of the cylinder; that is, we substitute for x , y , z the new variables x' , y' , z' with

$$x' = x/D, \quad y' = y/D, \quad z' = z/D.$$

Then D disappears from (41.19). In the same way, if we measure all velocities in terms of V —that is, we set $v = v'/V$, we get rid of the V , and v' is just equal to 1 at large distances. Since we have after our units of length and velocity, our unit

of time is now D/V , so we should set

$$t = \tau \frac{D}{V} \quad (4.20)$$

With our new variables, the derivatives in Eq. (4.1-8) get changed from $\partial/\partial t$ to $(1/D)\partial/\partial t$ and, since $v = \dot{x}$, Eq. (4.1-8) becomes

$$\omega = \nabla \times v = \frac{1}{\rho} \nabla^2 \times v' = \frac{\rho}{D} \Omega'. \quad (4.21)$$

Our main equation (4.1-7) then reads

$$\frac{\partial \Omega'}{\partial r} + \nabla \times (\Omega \times v') = \frac{1}{\rho D} \nabla^2 \Omega'. \quad (4.22)$$

All the constants condense into one factor which we write, following tradition, as $1/\rho L$,

$$\alpha_1 = \frac{1}{\rho} \rho D. \quad (4.23)$$

If we just substitute α_1 in all of our equations we can work with all quantities in the *new* units, yes even on all the sources. Our equations for the flow are then

$$\frac{\partial \Omega}{\partial r} + \nabla \times (\Omega \times v) = \frac{1}{\alpha_1} \nabla^2 \Omega. \quad (4.24)$$

and

$$\Omega = \nabla \times v$$

and the conditions

$$v = 0$$

at

$$r' = r'_{\infty} = 1.0 \quad (4.25)$$

and

$$v_x = 1, \quad v_y = v_z = 0$$

far

$$v' = v^2 + w^2 \approx 1$$

What this all means physically is very interesting. It means, for example, that if we solve the problem of the flow for a velocity V_1 and a certain cylinder diameter D_1 , and then calculate the flow for a different diameter D_2 and a different fluid, the flow will be the same for the velocity V_2 which gives the same Reynolds number—that is, when

$$\alpha_1 = \frac{D_1}{\rho} V_1 D_1 = \alpha_2 = \frac{D_2}{\rho} V_2 D_2. \quad (4.26)$$

For any two situations which have the same Reynolds number, the flows will "look" the same—in terms of the appropriate scaled r' , v' , w' , and Ω' . This is an important proportionality because it means that we can determine what the behavior of the flow of air past an airplane wing will be without having to build an airplane and test it. We can, instead, make a model and make measurements using a velocity that gives the same Reynolds number. This is the principle which allows us to apply the results of "the wind tunnel," measurements at small-scale airplanes, or "model" tests, right down scale-model boats, to the full-size objects. Remember, however, that we can only do this provided the compressibility of the fluid can be neglected. Otherwise, a flow quantity enters the picture of wind. And different situations will rarely correspond to each other only if the ratio of V to the sound speed is also the same. This latter ratio is called the *Mach number*. So, for velocities near the speed of sound or above, the flows are the same in two situations if both the Mach number and the Reynolds number are the same for both situations.

4-7

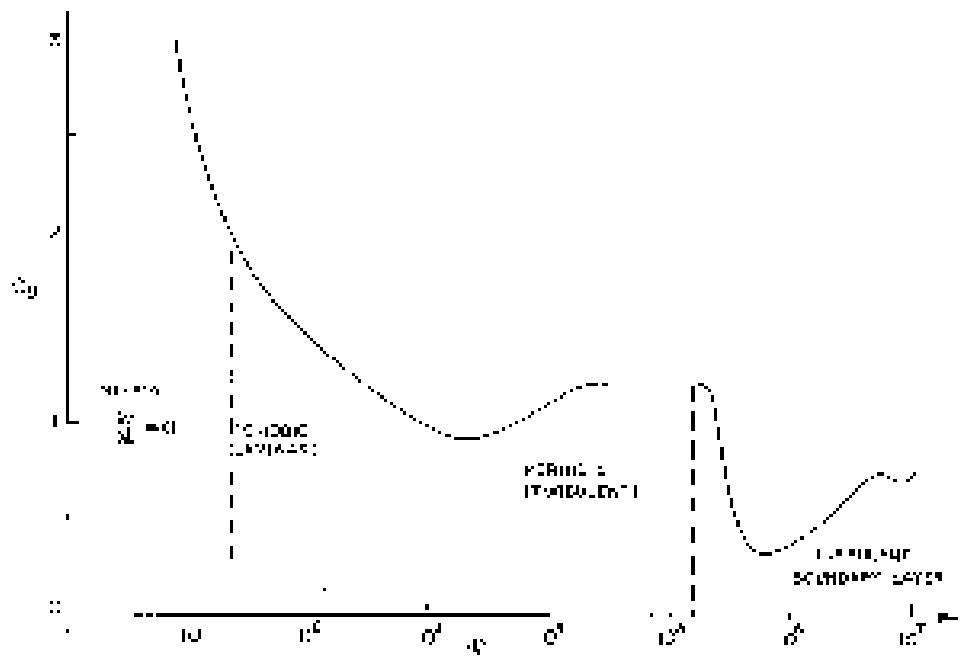


Fig. 41-4. The drag coefficient C_D of a circular cylinder as a function of the Reynolds number.

41-4 Flow past a circular cylinder

Let's go back to the problem of low-speed (nearly incompressible) flow over the cylinder. We will give a qualitative description of the flow of a real fluid. There are many things we might want to know about such a flow—how much heat, what is the drag force on the cylinder? The drag force on a cylinder is plotted in Fig. 41-1 as a function of ν —which is proportional to the air speed V if everything else is held fixed. What is actually plotted is the so-called drag coefficient C_D , which is a dimensionless number equal to the force divided by $\frac{1}{2}\rho V^2 D$. Here D is the diameter, V is the length of the cylinder, and ρ is the density of the fluid.

$$C_D = \frac{F}{\frac{1}{2}\rho V^2 D}$$

The coefficient of drag varies in a rather complicated way, giving us a preview that something rather interesting and complicated is happening in the flow. We will now describe the nature of flow for the different ranges of the Reynolds number. First, when the Reynolds number is very small, the flow is quite steady; that is, the velocity is constant at any place, and the flow goes around the cylinder. The usual distribution of the flow lines is, however, not like it is in potential flow. They are solutions of a somewhat different equation. When the velocity is very low or, what is equivalent, when the viscosity is very high so the fluid's like honey, then the inertial terms are negligible and the flow is described by the equation

$$\nabla^2 u = 0.$$

This equation was first solved by Stokes. He also solved the same problem for a sphere. If you have a small sphere moving under such conditions of low Reynolds number, the force needed to drag it is equal to $6\pi\mu rV$, where r is the radius of the sphere and V is its velocity. That is a very useful formula because it tells the speed at which tiny grains of sand (or other particles which can be approximated as spheres) move through a fluid under a given force—as, for instance, in a centrifuge, in sedimentation, or diffusion. In the low Reynolds number region—for ν less than 1—the lines of flow around a cylinder are as drawn in Fig. 41-5.

If we now increase the fluid speed to get a Reynolds number somewhat greater than 1, we find that the flow is different. There is a circulation behind the sphere, as shown in Fig. 41-6(b). It is still an open question as to whether there is a way

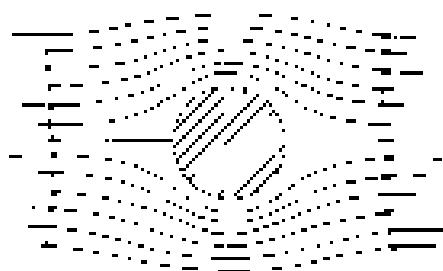


Fig. 41-5. Viscous flow (low velocity) around a circular cylinder.

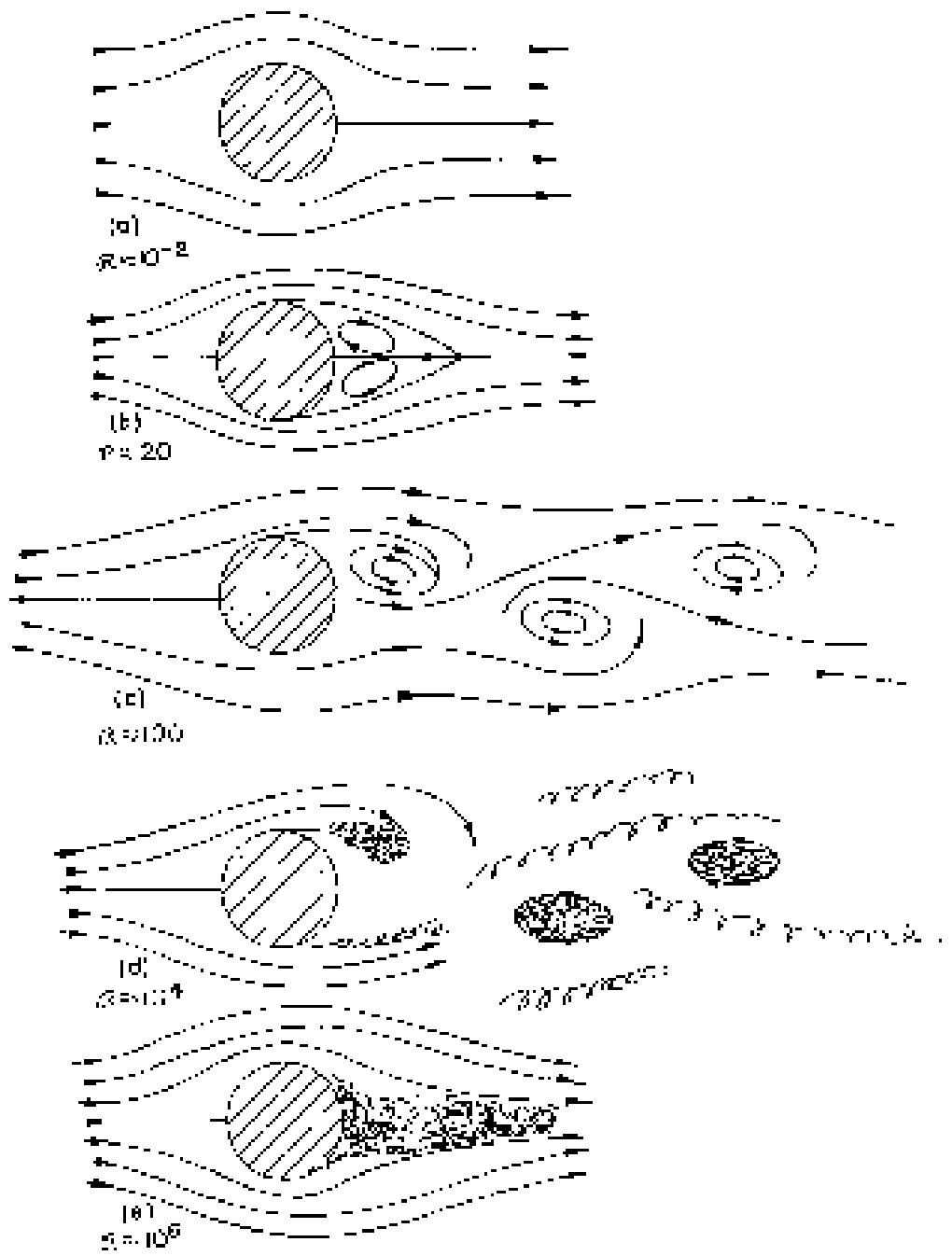


Fig. 41-6. Flow past a cylinder for various Reynolds numbers.

a circulation there even at the smallest Reynolds number or whether things suddenly change at a certain Reynolds number. It used to be thought that the transition was continuous. But it is now thought that it appears suddenly, and it is certain that the circulation increases with R . In any case, there is a sudden change to the flow for R in the region from about 10 to 30. There is a pair of vortices behind the cylinder.

The flow changes again by the time we get to a number of 40 or so. There is suddenly a complete change in the character of the motion. What happens is that one of the vortices behind the cylinder gets so large that it breaks off and travels downstream with the fluid. Then the fluid rotates around behind the cylinder, and makes a new vortex. The vortices peel off alternately on each side, so the instantaneous view of the flow looks roughly as shown in Fig. 41-6(e). The numerical

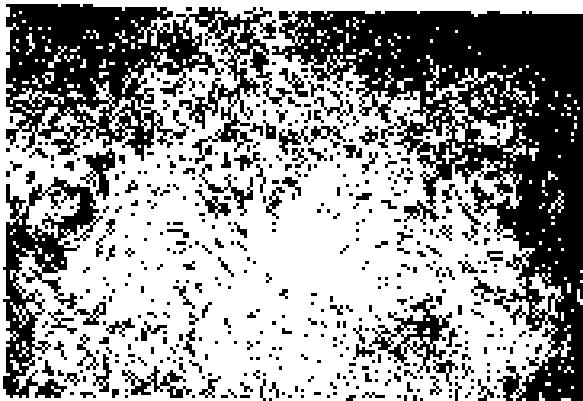


Fig. 41-7. Photograph by Ludwig Prandtl of the "vortex street" in the flow behind a cylinder.

vortices is called a "Kármán vortex street." They always appear for $R > 40$. We show a photograph of such a flow in Fig. 41-7.

The differences between the two flows in Fig. 41-6(a) and 41-6(b) or 41-6(c) is a most complete difference in regularity. In Fig. 41-6(a) or (b), the velocity is constant, whereas in Fig. 41-6(c), the velocity at any point varies with time. There is a steady solution above at $R = 40$ —which we have marked on Fig. 41-1 by a dashed line. For these higher Reynolds numbers, the flow varies with time but in a regular, cyclic fashion.

We can get a physical idea of how these vortices are produced. We know that the local velocity must be zero at the surface of the cylinder and that it also increases rapidly away from that surface. Vorticity is created by this large local variation in local velocity. Now where the mean stream velocity is low enough, there is sufficient time for ω to wanting to roll these out of the thin region near the solid surfaces where ω is produced and to propagate a long segment of vorticity. This physical picture should help us predict what the next change in the nature of the flow as the mean stream velocity, U_∞ , is increased still more.

As the velocity goes higher and higher, there is less time for the vorticity to diffuse over a larger region in front. By the time we reach a Reynolds number of several hundred, the vorticity begins to fill up a thin band, as shown in Fig. 41-8(a). At this stage, the flow is elongated and irregular. The region is called the boundary layer and the irregular flow region works its way further and further upstream until it becomes wider. In the boundary layer, the velocities are very much like U_∞ only after the flow is very fully developed, except that twists and turns in all three dimensions. This is called the regular separation, because superimposed on the turbulent one.

As the Reynolds number is increased further, the turbulent wake grows, moving forward until it reaches the point where the flow lines never rejoin again. The flow is somewhat above $R = 10^3$. The flow is as shown in Fig. 41-8(b), and we have what is called a "turbulent boundary layer." Also, there is a sharp jump in the drag force, D , due to a large factor, as shown in Fig. 41-4. In the speed region, the drag force not only decreases with increasing speed, there seems to be little evidence of periodicity.

What happens for still larger Reynolds numbers? As we increase the speed further, the wake increases in size again and the drag increases. The latest experiments, which go up to $R = 10^7$ or so, indicate that a new periodicity appears in the wake, either because the wake wake is oscillating periodically, or in a gross motion, or because some new kind of vortex is appearing together with an irregular, rotating motion. The details are as yet not clearly clear, and must be left to future experiments.

41-8 The limit of zero viscosity

We would like to point out that none of the flows we have described are anything like the potential flow section we found in the preceding chapter. This is, at first sight, quite surprising. After all, it is important that the flow goes to zero, equivalent to η going to infinity. And if we take the limit of large R ,

Eq. (41.20), we get rid of the right-hand side and get just the equations of the last chapter. Yet you would find it hard to believe that the highly turbulent flow at $\alpha = 10^7$ was approaching the smooth flow computed from the equations of "dry" water. How can it be that as we approach $\alpha = \infty$, the flow described by Eq. (41.20) gives a completely different solution than the one we obtained taking $\alpha = 0$ to start out with? The answer is very interesting. Note that the right-hand term of Eq. (41.25) has $1/\alpha$ times a second derivative. It is a higher derivative than any other derivative in the equation. What happens is that although the coefficient $1/\alpha$ is small, there are very rapid variations of α in the space near the surface. These rapid variations compensate for the small coefficient, and the product does not go to zero with increasing α . The solutions do not approach the limiting case as the coefficient of $1/\alpha$ goes to zero.

You may be wondering, "What is the line-gauge turbulence and how does it originate itself? How can the vorticity which is made somewhere at the edge of the cylinder generate so much noise in the background?" The answer is again interesting. Vorticity has a tendency to amplify itself. If we forget for a moment about the diffusion of vorticity which causes a loss, the laws of flow say (as we have seen) that the vortex lines are carried along with the fluid, at the velocity v . We can imagine a certain number of lines of ω which are being distorted and twisted by the complicated flow pattern of v . This pulls the lines closer together and mixes them all up. Lines that were straight before will get knotted and pulled close together. They will be longer and tighter together. The strength of the vorticity will increase and its irregularities—the plumes and ripples—will, in general, increase. So the magnitude of vorticity in three dimensions increases as we twist the fluid sheet.

You might well ask, "When is the potential flow a satisfactory theory at all?" In the first place it is satisfactory outside the turbulent region where the vorticity does not spread appreciably by diffusion. By making special streamlined bodies, we can keep the turbulent region as small as possible; the flow around airplane wings—which are carefully designed—is almost entirely a potential flow.

41-6 Couette Flow

It is possible to demonstrate that the complex and shifting character of the flow past a cylinder is not special but just the great variety of flow possibilities occurs generally. We have worked out in Section 1 a situation for two rotating fluid between two cylinders, and we can compare the results with what actually happens. If we take two concentric cylinders with a rod in the space between them and put a fine aluminum powder as a suspension in the oil, the flow is easy to see. Now if we turn the outer cylinder slowly, nothing unexpected happens; see Fig. 41-9(a). Alternatively, if we turn the inner cylinder slowly, nothing very surprising occurs. However, if we turn the inner cylinder at a higher speed, we get a surprise. The fluid breaks into horizontal bands, as indicated in Fig. 41-9(b). When the inner cylinder rotates at a similar rate with the outer one at rest, no such effect occurs. How can it be that there is a difference between rotating the inner or the outer cylinder? After all, the flow pattern we derived in Section 1 depended only on $\omega_i - \omega_o$. We can get the answer by looking at the cross sections shown in Fig. 41-9. When the inner layers of the fluid are moving more rapidly than the outer ones, they tend to move outward; the centrifugal force is larger than the pressure holding them in place. A whole layer cannot move out uniformly because the outer layers are in the way. It must break into cells and circulate, as shown in Fig. 41-9(b). It is like the convection currents in a room which has hot air at the bottom. When the inner cylinder is at rest and the outer cylinder has a high velocity, the centrifugal forces build up a pressure gradient which keeps everything in a jet. Marangoni—see Fig. 41-9(c) (as in a room with hot air at the top).

Now let's speed up the inner cylinder. At first, the number of bands increases. Then suddenly you see the bands become wavy, as in Fig. 41-9(d), and the waves travel around the cylinder. The speed of those waves is easily measured. For high rotation speeds they approach $1/2$ the speed of the inner cylinder. And no one

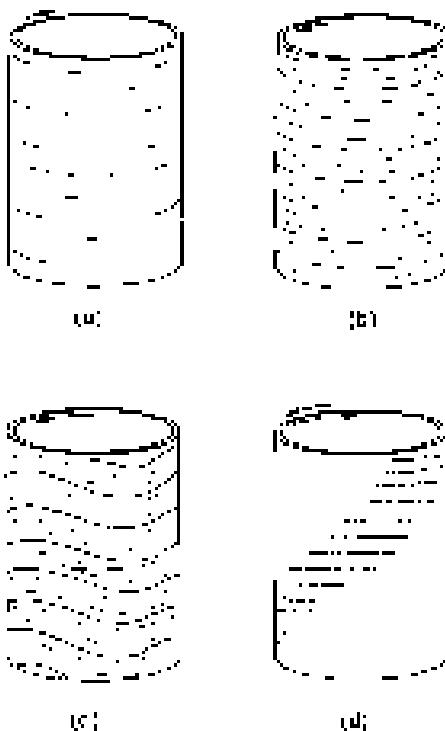


Fig. 41-9. Liquid flow patterns between two transparent rotating cylinders.

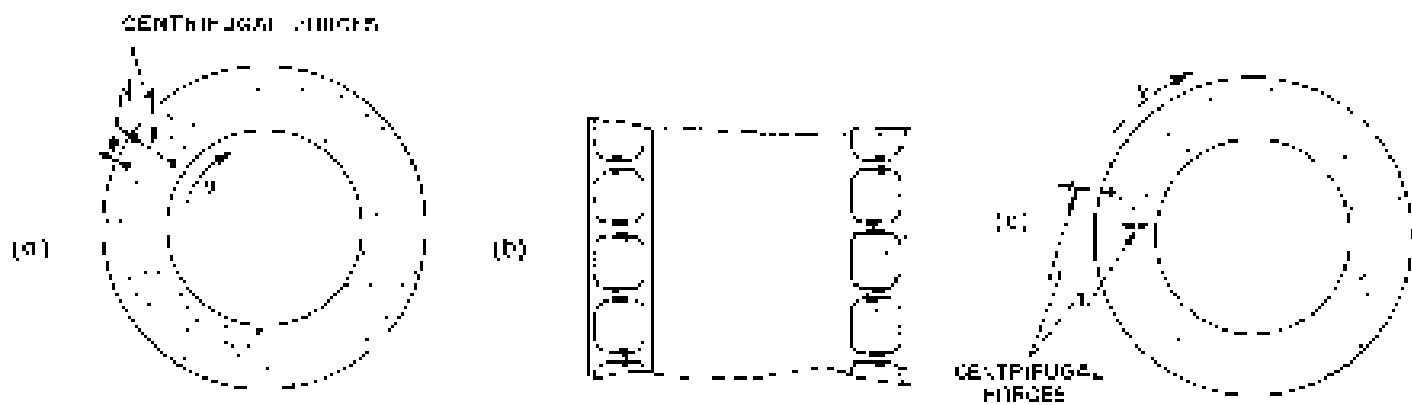


Fig. 41-2. Why the flow breaks up into jets.

knows why! There's a challenge. A simple number like λ_1 , and its relation to Ω know, the whole mechanism of the wave formation is not very well understood, yet it is steady laminar flow.

If we now start rotating the outer cylinder also – but in the opposite direction – the flow pattern starts to break up. We get wavy regions alternating with apparently quite random as sketched in Fig. 41-2(d), making a spotty pattern. In these "spotted" regions, however, we can see that the flow is really quite regular; it is, in fact, definitely turbulent. The wavy regions also begin to show a regular turbulent flow. If the cylinders are rotated still more rapidly, the whole flow becomes completely turbulent.

In this simple experiment we see many interesting regimes of flow which are quite different, and yet, which are all contained in our simple equations for various values of the one parameter Ω . With our rotating cylinders, we can see many of the effects we observed in the flow past a cylinder: first, there is steady flow, second, a few sets of wavy bands but in a regular, smooth way; finally, the flow becomes completely irregular. You have all seen the same effect in the column of smoke rising from a cigarette in your hand. There is a smooth steady column followed by a series of wiggles as the column of smoke begins to break up, ending finally in complete disarray about which I am sure.

The lesson I guess to be learned from all of this is that a tremendous variety of behavior is hidden in the simple set of equations in (41.23). All the solutions are for the same equations, only with different values of Ω . We have no reason to think that there are any further missing, free-flow equations. The only difficulty is that we do not have the necessary power today to solve them except for very small Reynolds numbers, ~ 1 , in the completely viscous case. That we have nothing in equal or cases not so near from the flow of Hales is a chance or mystery or its surprise.

If such variety is possible in a simple situation with only one parameter, how much more is possible with more complex situations! Perhaps the fundamental equation that does best the world, include sun, the continents, revolving, and exploding stars and galaxies, is just a simple equation like the hydrodynamics behavior of many plasma hydrogen gas. Often people in some unspecified part of physics say you can't write an equation for it. Well, perhaps we can. As a matter of fact, we very possibly already have the equation, it's just that it's too complicated when we write the expansion of quantum mechanics:

$$H\psi = -\frac{\hbar^2 \partial^2}{2m} \psi.$$

We have just seen that the complexities of things can so easily and dramatically escape the simplicity of the equations which describe them. And one of the secret of simple equations, that has often fascinated me, is that being about it good, and nice equations, is required to explain the complexities of the world.

We have written the equations of water flow. From experiment, we find a set of concepts and approximations to use to discuss the solution: vortex streaks, turbulent wakes, boundary layers. When we have similar equations in a less familiar situation, and one for which we cannot yet experiment, we try to solve the equations in a primitive, halting, and confused way to try to determine what new qualitative features may come out, or what new qualitative forms are a consequence of the equations. Our equations for the sun, for example, as a ball of hydrogen gas, describe a sun without sunspots, without the nice grain structure of the surface, without prominences, without coronas. Yet, all of these are really in the equations, we just haven't found the way to get them out.

There are those who are going to be disappointed when no life is found on other planets. Not I—I want to be reminded and delighted and surprised once again, through interplanetary exploration, with the infinite variety and novelty of phenomena that can be generated from such simple principles. The test of science is its ability to predict. Had you never visited the earth, could you predict the thunderstorms, the volcanoes, the ocean waves, the auroras, and the celestial surfaces? A salutary lesson it will be when we learn of all that goes on on each of those dead planets—those eight or ten balls, each generated from the same dust cloud and each obeying exactly the same laws of physics.

The next greatest era of awakening in human intellect may well produce a wealth of understanding the qualitative content of equations. Today we cannot fully see that the water flow equations contain such things as the border pole structure of turbulence that one sees between rotating cylinders. Let us, we cannot see whether Schrödinger's equation contains freights, musical compositions, or morality—or whatever it does not. We cannot say whether some entity beyond a like God is needed, or not. And so we can still hold strong opinions either way.

Index

- Absolute, I-22-7, I-34-6
Absolute zero, I-1-5
Absorption, II-51-2 ff
Absorption coefficient, II-12-3
Acceleration, I-4-5 ff
 component, I-2-3
 of gravity, I-3-4
Acceleration, g, of field, II-29-1 ff
Active fer energy, I-47-7
Active circuit element, II-22-5
Adams, J., I-7-5
Adiabatic compression, I-38-5
Adiabatic expansion, II-39-1 ff
Adiabatic expansion, I-13-2
Adiabatic nature, I-17-4
Aeromag, II-7-12
Air density, I-10-7
Aigaku, I-22-1 ff
All-vacuum wave circuits, II-22-1 ff
Accepting current densities:
 II-12-6 ff
Akiba S., II-37-10
Amber, II-1-10
Amsterdam, II-10-4
Ampere, A., II-1-3
Andreev's law, II-13-4
Andronov's oscillation, II-36-3
Anisotropic oscillation, I-21-5
Antimagnetic length, I-2-5
Analogy computer, I-22-8
Anderson, C. D., I-53-10
Angle, π radians, I-26-3
 of precession, II-32-4
 of rotation, I-26-3
Angular motion, I-2-3
Angular frequency, I-21-1, I-29-2
Angular momentum, I-7-2, I-28-5 ff,
 I-30-4,
 conservation of, I-2-7, I-18-6 ff,
 I-30-5
 of rigid body, I-20-8
Angular vibration, I-25-9 ff
Antiferromagnetic material, II-37-11
Antineutron, I-52-10
Antivortex, I-2-8
Aristotle, I-5-1
Aitken, I-4-3
 convergence, I-12-10
 Kutta-McCay method, II-5-5
 stability, II-5-8
 Thomson model, II-5-9
Aromatic clock, I-5-5
Atomic current, II-1-5 ff
Atomic hypothesis, I-1-2
Atomic units, II-1-8
Atomic particles, I-2-9 ff
Aromatic pi-electrons, II-32-2
 Kondo process, I-1-5-7
Attenuation, I-4-3
Aufbau's, A., I-29-2
- Average work, number, I-41-10
Axial vector, I-32-6
- Babinet, II-22-6
Bekerman, A. H., I-28-3
Bell, A. G., II-16-3
Benzene molecule, III-46-10 ff
Benzene ring current, II-40-6 ff
Besse, A., I-29-2 ff
Bertrand, II-17-5
Bragg-Williams law, II-14-10
Birstein, I-32-2 ff
Blackbody radiation, I-47-5 ff
Boeum, I-51-10
Bohm, II-4-7, II-12-12
Bohr, N., I-42-3, II-5-3
Bohr magneton, II-31-1 ff
Bohr radius, I-38-6
Bohr atom, I-1-4-2
Bohr-Zucker's law, I-30-3 ff
Boop, II-28-3
Booz, M., I-37-1, I-38-9, II-28-7
Boundary laws, II-1-2
Boundary-value problems, II-7-1
Boose's law, I-10-8
"Brookhaven," II-9-10
Brugg, L., II-30-9
Brugy-Nevo crystal model, II-30-9 ff
Breakthrough theory, II-5-9
Breakthrough, I-7-6-6
Brewster's angle, I-20-6
Briggs, P., I-22-5
Brown R., I-4-4-1
Brownian motion, I-1-6, I-6-5,
 I-11-1 ff
Brush discharge, II-30-9
Buck modulus, II-38-5
- Circular, differential, I-6-1, II-2-1 ff
 degree, II-3-1 ff
 of variation, II-16-3
C. index of beam, II-38-10
Capacitor, I-53-8
 current, II-22-7
Capacitor, I-1-9, I-28-5, II-22-2 ff,
 II-23-7
 parallel plate, I-14-6, II-5-11 ff,
 II-48-2
Capillary, II-5-11
 as a condenser, II-5-2
Capillary series, I-51-8
Carroll, S., I-4-2, I-4-8-2 ff
Case of cycles, I-44-5 ff, I-15-2
Copper sign, I-18-5
Coulomb, I-12-3
Coulomb, P., I-7-2
Cavendish experiment, I-7-2
Cavity resonator, II-29-1 ff
- Central forces, I-18-1, I-1-34-7
Centrifuge, force, I-1-5, I-22-11
Cerenkov, P. A., I-51-2
Cerenkov radiation, I-1-2
Charge conservation, I-1-5 ff,
 II-13-11
 current, I-10-2
 rate of, II-5-3 ff
 velocity, II-29-1 ff
 short, I-1-1
 sphere of, II-5-4 ff
Charge, density, II-5-4
Charge separation, II-2-7 ff
Changed condition, II-8-7 ff
Cronical energy, I-2-2
Chemical kinetics, I-42-1 ff
Chemical reaction, I-1-6 ff
Chronometry, I-1-5-6 ff
Coaxial, alternating current, II-22-1 ff
 impedance, II-32-10 ff
Circular elements, II-12-1
 series, II-22-5
 series, II-22-9
Circular motion, I-21-1
Circular wave, II-13-8 ff
Circular wave, II-13-8 ff
Classical electron radius, II-5-2-7
Climate, K., I-1-2, I-15-3
Clayton-Gibney equation, I-25-6 ff
Clausius-Messel's equation, II-11-6 ff,
 II-12-7
Cloudy plane, II-10-4
Conical line, II-24-1
Coefficient absorption, II-32-8
 of coupling, II-17-4
 of friction, I-12-4
 gravitation, I-7-9
 of viscosity, II-4-42
Collinear, I-16-5
 angle, I-10-7
Collective particles, II-7-8 ff
Color violet, I-38-1 ff
 absorption capacity, I-38-1 ff
Complex impedance, I-1-1-7
Complex numbers, I-22-9 ff, I-23-1 ff
Complex variable, II-7-2 ff
Compton eye, I-36-5 ff
Compton radiation, I-38-5
 law, I-4-4-5
Compton's parallel plate, I-13-2,
 II-6-1, II-7, II-8-4
Conductivity, II-3-2 ff
 current, II-2-8, II-12-2
Conductor, II-4-2
Cones, I-37-1
Conservation, of angular momentum,
 I-4-7, I-5-5 ff, I-20-5
 of charge, I-1-3, II-10-1 ff
 of energy, I-3-2, I-4-1 ff, II-29-1 ff
 of linear momentum, I-1-7,
 I-11-1 ff

- two dimensions, II 2–2 II
 vacuum, II 1–4 5, II 2–5–7
 Field energy, II 17–1 7
 of a point charge, II 28 1 r
 Field index, II 7–8–9
 Field ion microscope, II 6–14
 Field lines, II 4–1
 Field momentum, II 27–9 T
 of a moving charge, II 28 2–5
 Field strength, II 2–2
 Force, law, II 22 14 7
 Force field, II 2–8–9 II
 gravitational, II 23 5
 electro., II 4–7 2–7
 Fluid flow, II 13–5 7
 Flux, II 2–3 8 II
 electric, II 4–4
 of a vector field, II 2–2 2 II
 Function, II 12 17–7
 Focal length, II 27–1–7
 Focus, I 29 5
 Force, cent. I 2–1, II 7–5, II 17–11
 components, II 2–4 3
 conservative, II 2–2 2–7
 Forces, I 29 6 II
 Frequency, I 2 3 II, II 1–13 1–13–1
 Frequency, II 7–8 2
 Free fall, I 2 3
 Free space, I 2 3–4, II 18–14
 magnetic, II 1–2 3, II 22 1
 molecular, II 1–3, II 2–6 1
 incident, II 18–3
 nonconservative, I 2–4 6 II
 in elect., II 7–12
 pseudo, I 2 10 2
 Fourier, J. 1 50 2 7
 Fourier analysis, I–29 2 C
 Fourier theorem, II 7–11
 Fourier transform, I 2–4
 Four velocity, I 17–5–7, II 17–2 II,
 II 27–1 II
 Force, I 29 1
 Frank, J. 1 51 2
 Franklin, B., II 5–6
 Frequency, ultralow, I 2–1 3, II 29 2
 oscillation, II 2–3
 plasma, II 7–5, II 22 12
 Fourier's, vector theorem, I 23 8
 Function, I 1–3–8, II 1–2 3 II
 coefficient of, I 1–2 –

 Galileo, I 5 1, I 2–2, I 9 1, II 52 3
 Galilean relativity, I–17–2
 Galilean transformation, I 12 1
 Galvani, Luigi, II 1–8 12–16 2
 Faraday, II 37–17
 Faraday law, II 4–9 4, II 5–7 2–3
 Faraday's law, II 2–5–8
 Faraday effect, II 2–9 1
 Faraday, H. 5 1
 Faraday–Möller, I–12–9
 Generator, alternating current,
 II 17–6 T
 electric, II 16 1–2, II 22 5 II
 von de Genn, II 2–9, II 3–7
 Geometrical optics, I 26 1, II 27–1 /
 Gerlach, II 36 3
 Gluon, upper m., II 2–1, II 5–1
 Gravitation, I 2 A, I 7–1 II, II 12 2
 Gravity acceleration, I 2–2–4
 Gravitational coefficient, I 7 9
 Gravitational energy, I 2–2 II
 Gravitational field, I–17–8 (E, 1–1)–9 T
 Gradient, I 19 3 II
 angular, II 2–1
 Green's function, I 28 4
 Ground state, II 8–9
 Gyroscope, I 20 3 C

 Hamilton's first principle, II 1–2–5
 Hemeogenic meson, I 21 4, II 23 1 E
 Harmonic oscillator, I 17–1, II 21–1 T
 forced, I 21–5 C, II 23 2 II
 Harmonics, I 20 1 II
 Heat, I = 3, II 1–3–3
 Heat conduction, II 2–6 8
 Heat diffusion equation, II 3–8
 Heat energy, I 2–2, II 1–5, II 16–7
 II 10 5
 Heat engines, I 1–4–1 T
 Heat flux, II 2–8 II, II 12 2–7
 Heaviside, W., I 6–10, II 3–7–1,
 II 20 9, II 27 11–1 II 27 12, II 28 2
 Hecht, E., II 1–5, II 7, II 40–19
 Henry (unit), I–27–7
 Hess, II 9 2
 Hesiod's self, II 10–7
 High voltage breakdown, II 6 13 7
 Helmholtz, I 1–2 6, II 38–1 T
 Helium, C, I 2–5–2, II 2–2–7
 Hydrocarbon, II 19 2 II
 Hydrogen, II 4–6–1 II
 Hypocycloid, I 2 5 3
 Hysteresis curve, II 37–5 II
 Hydrotaxis, I 29 3–5

 Ideal gas law, I 29 10 7
 Illumination, II 2–10 7
 Image charge, II 6 4
 Impedance, I–26 8 II, II 27–1 T
 cavities, I–27–7
 Incident, angle of, I 26 3
 Incident plane, I 4–1
 Index of refraction, I 21 1–8
 Incident currents, II 16–1
 Incident ray, I–2–6, II 1–6–1 T
 II 17–12 II, II 22 2 II
 ray, II 17–9 II, II 22–16
 self, II 15 4, II 17 11 7
 Induction, law of, II 17–1–10
 Inductor, I 23 C
 Incident, I–2–5, II 2–7
 segment of, I 2–7, II 18 5 II
 zone plot of, I 2–1
 Infeld, II 26–7
 Induced radiation, I 22 8, II 25 1
 Inequal., I 4–7
 Inverted cylinder, II 2–1, II
 Invariant, II 1–2, II 20 1
 In reference, I–28–6, II 29–1 II
 Interfering waves, I 27 4
 Interference, I 1–5–5
 Internal reflection, II 22 12
 Ion, II 2–6
 Ionic bond, II 30–3
 Ionic conduction, II 24–6 II
 Ionic polarization, II 11–8

 Invariant energy, I 42 5
 In vacuo, II 2–5, II 3–3
 Isothermal zone, II 40 2
 Isotopic, II 1–2–3
 Isothermal gasphere, I–10–2
 Isothermal compression, I 44 2
 Isothermal expansion, I–6–2–3
 Isothermal surfaces, II 2–3
 Isotopic, I–1–2–3

 Iron, I 1–40–9, II 4–6 II, II 2–9–8
 Johnson noise, I 2–1 2, II 1–1 2
 Joule (unit), I 13 2
 Joule heating, II 24–3

 K–C of a series circuit, II 1–1–9
 Kepler, T., I 2–1
 Kepler's laws, I 7–1 6, II 2–1, II 18 6
 Kerr cell, II 23–8
 Knudsen (unit), II 9 2
 Kinetic energy, I–1 7, II 4–2, II 2–5 1
 II 29–4
 rotational, I 12 7 II
 Kinetic theory, I–19–7 T
 of gases, I 2–2 1 II
 Kirchhoff's laws, I–28–9, II 22–7 II
 Kronecker delta, II 31 6

 Lamb, II 7 6
 Landau–Lifshitz constants, II 39–6
 Landau vector, II 3–8–1
 Laplace, P., I 2–7 1
 Larmor frequency, II 6–1–1, II 7–1
 Laplacian operator, II 2–10
 Larmor frequency, II 24 7
 Laplace's theorem, II–34–6
 Laser, I 22 6, II 42 19
 Length, II 2–6
 Laws of electromagnetism, II 1–3 2
 of a charge, II 17–1 II
 Least action, principle of, II 15–1 T
 Least time, principle of, I 27–1 II,
 II 26–6
 Leibniz, G. W., I 5 2
 Lens formula, I–27–2
 Lens's rule, II 16–4, II 23–1
 Lenzen, U., I 2–5 5
 Length–Weyl potential, II 7–1–11
 Legendre, II 21 1 T
 momentum, II 24–10 II
 normalized, I 2–2 2
 separation eq., I–22 3–8
 speed of, I 2–2 1, II 18–8 II
 Light waves, I–2–1
 Lineariz., II 6 3–2 II
 Line of charge, II 2–3 2 II
 Line integral, II 1–2–2
 Linear momentum, conservation of,
 I–2–2, II 1–2–1 II
 Linear systems, I 28–1 II
 Loadstone, II 1–10
 Logarithm, I–23–1
 Lorentz, H. A., I 1–13 2
 Lorentz contraction, II 25 9
 Lorentz force, II 15–1, II 15 14
 Lorentz transform., II 7–12 II
 Lorentz gauge, II 18 2
 Lorentz transform., I–1–2,
 II 12–1, II 14 8, II 22 2, II 25 1
 II 23–8, II 25–1 II

Maxwellian spectrum, II-26-2
 Momentum spectrum, II-22-2
 Maxwells' eqns., I-3-5
 Mercury, I-9-10; II-3-7
 Methane, I-5-7; II-8-10
 frequency, II-36-9
 I-35-20, II-21-4
 resonance, I-14-2
 Harmonics, I-21-3, II-23-1, 3
 para value, I-8-10
 planetary, I-7-8, II-15-6, 16, 17-8
 Motors, electric, II-16-17
 Moving charge, field around, II-11
 II-23-24
 Noise, I-36-1-2
 Mutual conductances, II-17-8, 9,
 II-22-3

Neutral hydrocarbon, I-22-11
 Neutron, I-30-3, II-12-2
 Neutral gas, I-5-7
 diffusion of, II-12-3, 7
 Newton's diffraction equation, II-12-7
 Newman, L., I-38-1, II-15-1, II-37-1
 II-4-13
 Newton's law, I-10-1, II-1-1-2
 Newton's laws, I-2-3, II-2-3, III-2-11
 I-2-3-4, II-1-10, II-2-11-12
 II-2-3-4, II-26-2, II-41-1, II-46-1
 II-5-5
 Non-polar, I-3-6
 Nodes, I-3-4, 2
 Nonsim., I-36-1
 Nonpolar molecule, II-11-1
 Nucleus wave, I-3-9
 Nuclear energy, I-1-2
 Nuclear forces, I-12-2
 Nuclear fusion, II-3-10
 Nuclear interactions, II-3-7
 Nuclear magnetic resonance,
 II-36-10-6
 Nutrients, I-2-4, II-2-1-7
 Numerical analysis, I-2-6-6
 Nulliton, I-20-1
 Nyquist, J. F., II-36-9

One-dimensional, II-36-8
 Open circuit, I-28-9
 Optic axis, I-27-2-3, II-3-2
 Optics, I-2-2, II-2-1
 degenerate, II-2-2, II-3-1
 resonant, II-2-1, II-3-1
 Laserbeam, II-2-20
 water, II-2-4-5
 Optics axes, I-22-2
 Optimum size, I-45-2
 Optics, I-2-6-1, II
 seismic, I-2-6-1, II-27-1-4
 Optical memory, II-34-1
 Orientation polarization, II-11-1-7
 Oriented magnetic moments, II-35-4
 Orthorhombic cell, II-30-2
 Oscillation, simple, I-2-25
 damped, I-21-3-4
 frequency of, I-2-3
 natural, I-2-1-3
 periodic, I-2-1
 phase rule, I-21-2

Oscillation, 1-2, 2
 harmonic, 1-19-4, 1-2, 1-3, 2-3, 3-1,
 3-2, 3-3
 Polymer, rheological, 1-10-4
 Positive charge, 1-3-1
 Parabolic motion, 1-8-12
 Parallel-axis theorem, 1-11-2
 Parallel plate capacitor, 1-11-9,
 1-16-1, 1-17-8, 3
 Parallel system, 1-3-2-1, 1-11-1, 1-12
 Parallel rays, 1-29-2
 Parallel transmission, 1-14-2
 Parasite, "stranger", 1-5-2
 Parity, 1-17-11
 Parity守恒, 1-1-2-2
 Paraffin's triene, 1-6-2
 Passive circuit element, 1-17-2
 Pendulum, 1-45-6-8
 Period, 1-1-2-2
 Periodic oscillation, 1-21-3
 Periodic wave, 1-5-1-1
 Peripheral motion, 1-16-1
 Phase of oscillation, 1-21-3
 Phase shift, 1-21-1
 Phase velocity, 1-42-8
 Photon, 1-2-7, 1-26-1, 1-32-2
 Photoelectricity of cathode rays,
 1-12-3-3
 Photoelectricity, 1-11-2
 Photo, 1-1-2-2
 Photo, M., 1-4-5-6, 1-27-8, 1-43-9
 Planck's constant, 1-5-10-1, 6-10,
 1-17-8, 1-32-1
 Plane waves, 1-26-2
 Polar waves, 1-21-1, 8
 Planetary motion, 1-7-1-11, 1-25-6-7,
 1-26-5
 Plasma frequency, 10-7-2, 11-32-12
 Plastic oscillator, 10-7-5-7
 Plutonium, 1-5-6
 Plywood, 1-1-2-3, 1-18-5, 1-16-1
 Poincaré stars, 1-29-1
 Point charge, electric field energy of,
 1-9-13
 Field energy of, 11-28-1-7
 Poisson's ratio, 1-13-2
 Polar coordinate, 1-1-1, 1-11-3-6
 Polarization, 1-3-1-11, 1-12-1-11
 Polarization charge, 1-10-2-9
 Polarization vector, 1-10-2-1
 Polarized light, 1-17-9
 Polarized light (y), 1-1-2, 1-2-1, 2,
 1-14-1, 8
 Potential gradient of the atmosphere,
 10-9-2-7
 Power, 1-13-2
 Preying, 1-11-27-5
 Precession angle of, 10-34-4
 of atomic magnet, 11-31-4-5
 Present, 1-1-2-2
 Present, 1-1-2-5
 Present, 1-11-2-5
 Principle of least action, 10-19-1-8
 Principle of superposition, 11-1-2-3,
 11-4-2-2
 Probability, 1-1-2-11
 Probabilistic density, 1-6-8-1
 Probability distribution, 1-1-2-2
 Propagation factor, 1-22-1-1
 Proton, 1-2-4

- Proton spin, II-8-9
 Pseudo force, I-2, II-8 ff
 Pyrotron, I-26-2
 Purcell effect, I-17-2
 Pyroelectricity, II-11-8
- Quadrupole term**, II-7-1, II-29-6
 Quadrupole moment, II-5-8
 Quantized magnetic states, II-35-1 ff
 Quantum electrodynamics, I-2-2, I-2-4 ff,
 I-26-2
 Quantum mechanics, I-2-2, I-2-4 ff,
 I-26-2, I-27-2 ff, II-1 ff
 I-38-1 ff
- Kubo, T., II-35-4
 Kuhn molecular-bead model, II-35-1 ff
- Radiant energy**, I-4-2
 Radiation, infrared, I-33-8, I-34-1
 relativistic effects, I-35-1 ff
 x-ray spectrum, I-34-8 ff, I-34-6
 ultraviolet, I-26-1
 Radiative damping, I-35-3
 Radiation resistance, II-2-1 ff
- Radiative waves, I-5-3 ff
 Raphson, C., I-5-2
 Random walk, I-6-1 ff, I-6-4 ff
 Ratchet and pawl machine, I-36-1 ff
- Rayleigh's criterion, I-34-6
 Rayleigh's law, I-41-4
 Rayleigh waves, II-38-2
 Resistors, II-22-11
 Reciprocity principle, I-30-7
 Reification, I-50-9
 Recifier, II-22-15
 Reflected waves, II-32-7 ff
- Refraction, I-26-2 ff
 angle of, I-26-3
 incidence, II-15-12
 of light, II-23-1 ff
- Relativity, I-26-2 ff
 anomalous, I-33-9 ff
 index of, I-21-1 ff
 of light, II-31-1 ff
- Refractive index, II-22-1 ff
- Rock wave, penetrability, II-36-9
 Relativistic dynamics, I-15-9 ff
- Relativistic energy, I-16-1 ff
- Relativistic mass, I-16-1 ff
- Relativistic momentum, I-10-8-4,
 I-16-1 ff
- Relativity, at electric field, II-19-6 ff
 Galilean, I-16-2
 in magnetic field, II-13-6 ff
 special theory of, I-15-1 ff
 theory of, I-15-11, I-17-1
- Resistance, I-23-5
 Resistor, I-20-5, II-22-4
 Resonant cavity, II-23-6 ff
- Resonant circuit, II-23-6 ff
- Resonant mode, II-22-10
 Resonator, cavity, II-23-7 ff
- Resolving power, I-27-7 ff, I-30-5 ff
- Resonances, I-23-1 ff
 electrical, I-23-5 ff
 in molecule, I-23-7 ff
- Resonance polymerization, I-2-9
- Resonated time, I-28-2
- Rutherford, II-5-6
 Rutherford, J. D. :
 Rutherford nuclear, II-41-5 ff
- Rigid body, I-18-1
 angular momentum of, I-26-8
 rotation of, I-18-1 ff
- Ritz combination, zinc blue, I-38-8
- Roots, I-25-1, I-26-6
- Roemer, O., I-7-2
- Root-mean-square distance, I-5-6
- Rotation, of axes, I-11-2 ff
 plane, I-18-1
 of a rigid body, I-16-3 ff
 in space, I-20-1 ff
 in two dimensions, I-5-1 ff
- Rutherford, I-25-9
- Ruthenium, II-7-3
- Rutherford-Bohr atomic model, II-5-2
- Rydberg (unit), I-38-6
- Scalar, I-11-5
- Scalar field, II-7-3 ff
- Scalar product, II-25-3 ff
- Scattering of light, I-20-2 ff
- Schrödinger, E., I-35-5, I-37-1,
 I-38-9
- Schrödinger equation, II-5-12
- Scientific method, I-2-1 ff
- Screw drive, see, II-30-9
- Screw jack, I-4-5
- Second law, I-5-5
- Seismograph, I-5-8
- Self-inductance, II-16-1 ff, II-17-1 ff
- Shannon, C., I-44-3
- Shear modulus, II-38-5
- Shear waves, I-31-4, II-34-8
- Sheet of charge, II-3-4
- Side walls, I-28-4 ff
- Simultaneity, I-15-7 ff
- Sinusoidal waves, I-29-2 ff
- Skin depth, II-15-11
- Slip lubrication, II-30-8
- Smirnov model, I-4-8
- Smooth muscle, I-11-2
- Snell, W., I-26-5
- Snell's law, I-26-3, I-3-2, II-15-1
- Selenide, II-15-5
- Self-state physics, II-3-2
- Second, I-2-2, I-2-4 ff, I-30-1
 speed of, I-47-7 ff
- Space, I-3-2
- Space-time, I-2-2, I-15-1 ff, II-26-12
- Special theory of relativity, I-15-1 ff
- Specific heat, I-40-7 ff, I-45-2 ff, II-37-4
- Speed, I-6-1 ff, I-4-3
 of light, I-15-1, II-18-8 ff
 of sound, I-47-7 ff
- Speed of sound, II-3-4 ff
- Speed of travel, II-20-2 ff, II-21-2 ff
- Spine, I-15-1 ff
- Spin-orbit, II-5-7
- Spontaneous emission, I-5-2 ff
- Standard deviation, I-6-9
- Source, I-4-1 ff
- Statistical fluctuations, I-6-3 ff
- Statistical mechanics, I-5-1, I-40-1 ff
- Steady flow, II-40-6 ff
- Step ladder, II-9-10
- Stern, M. 35-9
- Stern-Gerlach experiment, II-35-3 ff
- Stewarite, Sr, I-4-5
- Stokes' theorem, II-2-1 ff
- Stream, II-15-2
- Stress tensor, II-31-5 ff
- Striated muscle, I-4-7
- Supermalley, II-36-9
- Superposition, II-15-1 ff
 of fields, I-4-2 ff
 principle of, I-25-2 ff, II-1-2,
 II-4-1
- Surface, equivalent**, II-1-1 ff
 isothermal, II-10-1
 isobaric, II-2-3
- Surface tension**, II-12-3
- Symmetry**, I-4-5, I-11-1 ff
 of physics laws, I-16-2, I-52-1 ff
- Szilard-Chen, I-2-5, I-15-9, I-34-9 ff,
 I-34-6, II-17-5
- Tamme, J., I-5-1-2
- Taylor expansion, II-6-7
- Temperature, I-25-6 ff
- Tensor, II-26-1, II-31-1 ff
- Testified, II-31-1 ff
- Tetrapole cell, I-3-3(=7)
- Thermal conduction, II-2-3, II-12-2
 cf. also, I-47-9 ff
- Thermal equilibrium, I-11-3 ff
- Thermal insulation, I-42-5 ff
- Thermodynamics, I-96-2, I-15-1 ff,
 II-57-4 ff
 laws of, I-15-1 ff
- Thompson, II-5-3
- Thorium atom model, II-5-1
- Thompson scattering cross section, I-32-8
- Three-body problem, I-10-2
- Three-dimensional waves, I-20-2 ff
- Thunderstorms, II-20-5 ff
- Ticks, I-7-2 ff
- Time, I-3-1, I-5-1 ff, I-8-1, I-8-2
 reversed, I-26-2
 standard, I-5-5
 transformation of, I-15-2 ff
- Time-avg, I-18-4, I-23-1 ff
- Tension bar, I-3-2 ff
- Total internal reflection, II-22-12 ff
- Transformation, Fourier, I-25-4
 constant, I-12-1
 linear, I-11-6
 Lorentz, I-15-5, I-16-1, I-20-2 ff
 I-52-7, II-25-1, II-26-1 ff
 of time, I-15-5
 of velocity, I-16-1 ff
- Transformation, II-16-1
- Constant, I-24-1 ff
 electrical, I-34-5 ff
- Transient response, I-2-5
- Translation of axes, I-11-1 ff
- Transmission, linear, I-2-2 ff
- Transmitted waves, II-23-1 ff
- Travelling field, II-16-5 ff
- Tree-like lattice, II-31-2
- Trigonal lattice, II-17-7
- Two paradoxes, I-16-3 ff

fixed reference field, II-7, 2-5
Fisher metric, I-7-8
flat, global horizon, I-28-1
flatness principle, I-2, 5, 16, 21,
I-27-9, I-37-1, I-38-3, II-8-3
flat cell, I-28-3
flat world, I-17-1, II-7-3
flowchart, II-25-16

van der Pol differential eq., II-5-6, II-8-7
Vector, I-5-9
Vector algebra, I-4-6, 1-52-2
Vector analysis, I-11-3, 1-52-2
Vector field, II-1, 2-7, I-2, 1-7
flux of, I-3-3
vector integrals, I-3-1-7
vector operation, II-3-8
vector potential, I-11-4, 1-21, 1-28-1
vector product, I-4-9-1
velocity, I-8, 3, 1-9, 2-1
 components of, I-4-5
 transformation of, I-10, 4-3
velocity potential, II-7-9
Voronoi diagram, I-3-1-2
vertical world, II-5-2, 1-4-5
viscosity, II-2-3-1
 coefficient, II-4-1-2

viscosity drag, II-21, 4-5
Viscous fluid, I-2-6-1 II
 friction, I-26-1
 roles, I-34-1, II
visual cortex, I-26-2
visual process, I-11-2
volumetric, II-28-1
volume & volume, II-18-3
volume stress, II-38-2
von Neumann, J., II-20-3
Voronoi cell, II-10-1-10
vertices, II-49-5

wall energy, II-21-6
Walsh, I-52-2
Walsh cell, I-13-3
Wave, I-21-1 ff., II-20-1 ff.
 electromagnetic, II-21-1-2
 light, I-28-1
 radio, II-50-1 ff.
 harmonic, II-33-1-4
 shear, I-31-4, II-38-5
 resonant, I-29-2-3
 soliton, II-26, 1-2, II, II-7-2-3
 three dimensions, II-20-3-7
 transient, II-4, 1-7-9
Wave equation, I-2-7-1, II-15-2 ff.
wavefunction, I-4-1-1

Weveguides, II-24-1, II
Wheeler, R., I-16-3, I-20-1
white matter, I-22-2
Whistler, II-7-6-7
Wheeler, David, I-1-3-1
Wheeler's model, I-41-1-7
Whey, M., I-11-1
Whistler, II-28-2
Wilson, C. T., R., II-20-2
Wink, J. D., I ff., I-11-1, II

X-ray, I-7-8, 1-9-
X-ray diffraction, II-20-1

Young, I-5-6-7
Young's modulus, II-2a-2
Y-axis, R., I-2-8, II-28-2-3
Yukawa potential, II-28-1-3
Yukawa, I-35-8

Zeno, I-8-3
Zeta, coordinate, .., I-5
Zero, ..., I-3-16, II-7-1-1
Zero divergence, II-2, 1-1, II-1-1
Zero mass, I-7-11



Frymann's Professor

I hope you will excuse this long letter, but I feel it is important to let you know about a supplement I wrote at Caltech. The instruments of Gauss and Schuster—by now well edited—have been extensively used in recent years. They are now an integral part of the mapping process. For which you may be grateful as much as I am. I hope you will take a look at them. They letters will then be made up into a book, a copy of which I'll send you if you like. Under the guidance of a teacher we will make it all work. There was a tremendous amount of material.

The spatial network we used to analyze the species was built to include the 70% of the species that did not have enough data to be included in the high-spatial-resolution grid. There were also 10 other species whose distributions

It's good to be here at the city quorum because we can make a difference. By virtue of who we are, our problems come to my heart heavily because they are really very big problems now, and the worst part about it is, they may make the only influence planks election ballot, and if the first one that goes down, one that's relatively big like public works or something else, then there's nothing left.

It was very hard to find many very simple topics to be considered, but after very
careful thought I think he addressed them to 24 hours, and yet it is necessary to take
time, - possibly, and even the most intelligent student was unable to go through
of 100% his exercises, but yes in the future, - by putting in a good time of applica-
tion, - it will be easy and natural in solving all exercises outside the usual life of
the book. For this reason, though, I tried very hard to make all the statements as
concise as possible, - again, you in every exercise do the equations and - estimate
into the body of physics, one new, - which the student is very likely would be
interested. - You tell that to each student, it is important to indicate what it is
that they think, if they are extremely clever, to able to indicate all the components
of what is measured there, and what is being put in to estimating new,
when new measurement. I think it is the best item, - they were collecting
information, - and, - every day I called on him, and his basic information, - he
had already learned, and which was actually used to begin research - was you
asked in

Some students described how they wanted their parents to know more about what they were doing at high school. Students tended to emphasize the value of being involved in extracurricular activities.

Foreword

A most triumphant twentieth century saying, that one of the main reasons for our nearly 10 years' work will be that eventually here, giving our students a true fundamental knowledge of physics to many students, ever lastingly with hardly more than a month's study, to this we add a great advantage of the physical work. We should be able to hear the arguments and all the experiments that were made, and secondly those of the quantum mechanics in a way that would be really incomprehensible. By approach you will find here a hard, perhaps hitherto never seen, simple course, and this is done at very much less cost. After having however easily come off the students take to it. I expect, I believe that the experiments will be success. There is, of course, no room for improvement, and it will come with more experience in the exercises. When you will find here the record of the first experiment.

The first set of lectures of the Feynman lectures on Physics which were given from September 1961 through May 1962, the introductory physics course at Caltech, the concepts of quantum mechanics brought in whatever they seem necessary for an understanding of the phenomena being discussed. In addition, the last twelve weeks of the school year were given over to a more advanced treatment of some of the concepts of classical mechanics. It became clear as the lectures drew to a close, however, that not enough time had been left for the quantum mechanics. As the material was prepared, it was clear it very deserved the extra time, and interesting topics such as noted with the early three books that had been developed. There was also danger that the students might not be Seldinger wise in which they had been installed in the twelfth lecture would not provide a sufficient bridge to the more conventional treatment of many topics the students might hope to read. It was therefore decided to go on to the next volume, although however they were due to the sophomore class in May of 1962. These lectures repeated and extended somewhat the material developed in the earlier lectures.

In this volume, we have paid attention the lectures has. Both come with some adjustment of the sequence. In addition, five lectures originally given to the sophomore class as an introduction to quantum physics have been added. In addition to the chapters may were Chapters 17, 18, 19, and placed as before, the chapters the "classical" new volume is self-contained. As a relatively independent part of its two parts the idea about the quantization of angular momentum, including a discussion of the "free-electron experiment had been introduced" in chapters 14 and 15 of Volume II, and similarly with atoms is necessary for the consequences of these, why was not done this explicitly at hand. These 19 sections are contained here as an Appendix.

This set of lectures tries to emphasize from the beginning three features of the quantum mechanics which are most basic and most general. The first feature is the head of the "invariance of a global" - a principle, the invariance of amplitudes, the absolute nature of a state, and the superposition and combination of states. And the Dirac condition is used from the start. In each exercise, the ideas are illustrated by both with a detailed discussion of some specific examples and by a sketchy, the physical ideas as real as possible. The time dependence of states involving states of definite energy comes next, and the ideas are applied in once to the theory of bound-state systems. A detailed discussion of the important cluster problem is becoming

wave for the production, transmission, absorption and release of radiation. The text also attempts to consider more complex systems, leading to a discussion of the propagation of pulses and crystals, and a brief consideration of the use of wave packets in the calculations of optical properties. The author also includes some appendices such as in Chapter 20 which discusses the Schrödinger wave function in different representations, the expansion for the hydrogen atom.

The last chapter in the volume is not intended as the principal reference, but is for "inspiration, general interest, it was never to be expected to be used". Several parts of the book have come with the "risk of offending" the students, however, view of the results of what they seem likely to be general features of classical "Dynamical Variables" cannot be passed over without some offence.

As explained in the Foreword to Volume I, the first volume of the experimental program for the development of a new introductory text was completed at the California Institute of Technology under the supervision of the Physics Division Radiation Committee, Chairman, Julian Nesterov, and Mr. Ernest Staudt. The project was made possible by a grant from the Ford Foundation. Many people helped with the task, the details of the preparation of the system. Mr. Carl Christian, June Cushing, Louis Hanks, Donald Kirby, Martin Ladd, Michael Nelson, Robert Wilson, and George Zimmerman. Preliminary experiments and data were collected, are only now available and, clearly, the material is increasingly tangibly, much of the authors' input.

But the study of quantum mechanics, even without them, is full of difficulties. Our efforts will have been well spent if we can contribute to bring a little clarity to the intellectual noise and the experimental noise in the field, which is his wildlife lectures on physics.

Douglas D. Scott

California Institute of Technology

Contents

Chapter 1. QUANTUM ELECTRODYNAMICS

- 1-1 Atoms and nuclei 1-1
- 1-2 Atoms interact with light 1-1
- 1-3 Atoms interact with waves 1-1
- 1-4 Atoms interact with electrons 1-4
- 1-5 The interaction in older ways 1-5
- 1-6 Weighting the charge 1-5
- 1-7 The particle or quantum 1-9
- 1-8 The uncertainty principle 1-11

Chapter 2. THE CLASSICAL WAVE AND PARTICLE VIEWS 2-1

- 2-1 Probability wave amplitude 2-1
- 2-2 Momentum, position and momentum 2-2
- 2-3 Phase, direction 2-4
- 2-4 The size of an atom 2-5
- 2-5 Energy levels 2-7
- 2-6 Photoelectric experiments 2-9

Chapter 3. SCATTERING AMPLITUDES

- 3-1 The sum of scattering amplitudes 3-1
- 3-2 The total scattering pattern 3-2
- 3-3 Scattering from a region 3-3
- 3-4 Identical particles 3-6

Chapter 4. LASER AND BEAMS

- 4-1 Basic principles and basic particles 4-1
- 4-2 States with two flow particles 4-3
- 4-3 States with a flow particle 4-5
- 4-4 Evolution of a collection of photons 4-7
- 4-5 Track records 4-8
- 4-6 Liquid helium 4-12
- 4-7 The exclusion principle 4-12

Chapter 5. SCATTERING

- 5-1 Inelastic electron scattering 5-1
- 5-2 Experimental evidence for inelasticity 5-3
- 5-3 Atom scattering from nuclei 5-6
- 5-4 The case 5-8
- 5-5 Interfering amplitudes 5-10
- 5-6 The mechanics of quantum mechanics 5-12
- 5-7 Incohering and coherent waves 5-15
- 5-8 Other solutions 5-15

Chapter 6. SCATTERING

- 6-1 Incohering amplitude 6-1
- 6-2 Incohering total cross section 6-3
- 6-3 Relativistic, but thermalized 6-5
- 6-4 Relations of 6-3 with 4-1 above 6-6
- 6-5 Relativistic 6-7
- 6-6 Reference numbers 6-7

Chapter 7. THE DETERMINATION OF ASYMPTOTES OF TIME

- 7-1 Apparent wave velocity 7-1
- 7-2 Uniform motion 7-1
- 7-3 Relativistic energy versus motion 7-6
- 7-4 Proper velocities 7-7
- 7-5 The “observed” of a spacetime path 7-10

Chapter 8. THE HADRONIC MATRIX

- 8-1 Amplitude and source 8-1
- 8-2 Non-interacting source 8-2
- 8-3 What are the variables in the matrix? 8-5
- 8-4 How soft a range will one 8-7
- 8-5 The form of the matrix 8-9
- 8-6 The incoming nucleon 8-11

Chapter 9. THE APPROXIMATE Nucleus

- 9-1 The state of a nucleon outside 9-1
- 9-2 The nucleon in a nucleon inside 9-3
- 9-3 Transition in a nucleon inside 9-4
- 9-4 The optical potential 9-11
- 9-5 Transition amplitudes 9-15
- 9-6 The also part of light 9-16

Chapter 10. THE PARTICLE REVIEWS

- 10-1 The hydrogen molecule 10-1
- 10-2 Nuclear forces 10-6
- 10-3 The hydrogen molecule 10-8
- 10-4 The hydrogen molecule 10-9
- 10-5 Prog. 10-11
- 10-6 By 10-11
- 10-7 By 10-11
- 10-8 By 10-11
- 10-9 By 10-11
- 10-10 By 10-11
- 10-11 By 10-11
- 10-12 By 10-11
- 10-13 By 10-11
- 10-14 By 10-11
- 10-15 By 10-11

CHAPTER 11. Mean Free-Path Calculations

- 11-1 The Poisson equation 11-1
- 11-2 The scattering cross section 11-3
- 11-3 The solution of the wave PDE equation 11-4
- 11-4 The probability density in the problem 11-5
- 11-5 The natural boundary 11-12
- 11-6 Stochasticity in Monte-Carlo 11-16

CHAPTER 12. The Kinetic Energy of Atoms

- 12-1 Kinetic energy for a system of N atoms in a cell periodic 12-1
- 12-2 The Fermi motion for a ground-state light gas 12-3
- 12-3 Currents 12-5 12-7
- 12-4 The Fermi velocity 12-9
- 12-5 Kinetic energy in a magnetic field 12-12
- 12-6 The projection velocity 12-9 12-11

CHAPTER 13. Derivation of a Contact Current

- 13-1 The electron-electron and ion-electron currents 13-1
- 13-2 Self-energy 13-1
- 13-3 The occupied states 13-3
- 13-4 An electron at the Fermi-energy 13-5
- 13-5 Occupied states 13-6
- 13-6 Selectivity interface in the Fermi 13-10
- 13-7 Transport by a source temperature 13-11
- 13-8 Symmetry breaking and current scale 13-13

CHAPTER 14. Polarizations

- 14-1 Electricity and magnetism 14-1
- 14-2 Impedance calculations 14-4
- 14-3 The Hall effect 14-7
- 14-4 Superconducting junction 14-8
- 14-5 Conductance quantization 14-10
- 14-6 The Zitterbewegung 14-11

CHAPTER 15. The Incoherent Particle Transmissions

- 15-1 Scattering 15-1
- 15-2 The wave function 15-1
- 15-3 Interference patterns 15-3
- 15-4 The wavefunction 15-7
- 15-5 More energy filtering 15-9
- 15-6 Illustration of interference 15-12

CHAPTER 16. The Diffraction of Particles

- 16-1 Amplitude and loss 16-1
- 16-2 The wave function 16-2
- 16-3 Fermi and the momentum 16-7
- 16-4 Normalization of the waves 16-10
- 16-5 The Rutherford spectrum 16-11
- 16-6 Ionization energy 16-11

CHAPTER 17. WAVEFUNCTIONS IN ONE-DIMENSIONAL BOXES

- 17-1 Symmetry 17-1
- 17-2 Boundary and a resonance 17-3
- 17-3 The Schrödinger law 17-7
- 17-4 Polarized light 17-9
- 17-5 The transmission of a 1D 1D 11
- 17-6 Summary of the localized modes 17-15

CHAPTER 18. Absolute Momentum

- 18-1 Absolute dipole radiation 18-1
- 18-2 Light scattering 18-2
- 18-3 Measurement of polarization 18-5
- 18-4 Radiation with a very large 18-9
- 18-5 Scattering in disorder 18-11
- 18-6 Conservation of angular momentum 18-14
- Added Note 18-15 Conservation of the angular momentum 18-15
- Added Note 18-16 Conservation of parity in photon emission 18-22

CHAPTER 19. The Hydrogen Atom and The Rydberg Table

- 19-1 Schrödinger's equation for the hydrogen atom 19-1
- 19-2 Schrödinger's equations solutions 19-4
- 19-3 States with angular degeneracy 19-5
- 19-4 The ground state of the hydrogen 19-7
- 19-5 The hydrogen wave function 19-12
- 19-6 Ionization table 19-13

CHAPTER 20. Operators

- 20-1 Operators and operators 20-1
- 20-2 Acting on a state 20-2
- 20-3 The Heisenberg uncertainty 20-6
- 20-4 The position operator 20-7
- 20-5 The momentum operator 20-9
- 20-6 Angular momentum 20-11
- 20-7 Comparison of energy with time 20-12

CHAPTER 21. The Schrödinger Equation in One-Dimension: A Summary of Superconductivity

- 21-1 Schrödinger's equation in one-dim. 21-1
- 21-2 The location of centers of the wavefunctions 21-3
- 21-3 The band of momentum 21-4
- 21-4 The location of the wave function 21-7
- 21-5 Superconductivity 21-7
- 21-6 The Meissner effect 21-8
- 21-7 Flux quantization 21-17
- 21-8 The generation of superconductivity 21-18
- 21-9 The Josephson junction 21-19

APPENDIX: GLOSSARY

Abbreviations

Topics

Quantum Behavior

1-1 Atomic databases

"Quantum mechanics" is the description of the behavior, of course, of light and atoms, and, in particular, of the happenings on an atomic scale. The go to a very small scale before we notice that you have any direct, objective events. They do not believe they sense, that is, *object*s like particles, may not be, over interplanetary distances or in weights or opinions, or in anything that you have ever seen.

Newton thought that light was made up of particles, yet there was something that he never saw—light. Today, however, if we look at the properties of the particle concept, it was found that light did indeed sometimes behave like a particle. Similarly, the electron, for example, was thought to behave like a particle, and then it was found that in many respects behaved like a wave. So it really behaves like either. Now we have *two* types. We say, "It's like *both*."

Let's now look back, however, about one century ago the light. The ecosystem is system of many objects—photons, protons, neutrons, photons, and so on—and the same for all, they are all quantum systems that can be used to call them. So what we learn about the properties of electrons and how shall we turn a compact will apply also to all "particles" including photons of light.

The greatest achievement of quantum theory is a description of the wave function during the first quarter of this century, when you were interested about how such things do behave—quantum mechanics & mechanics which was finally received in 1933 and '34 by Schrödinger, Heisenberg, and Born. They finally obtained a consistent description of the behavior of particles at a *cm* scale. We take a look at the quantum description in this chapter.

Because particle behavior is often not ordinary experience, it is very difficult to get used to and it appears peculiar and mysterious to everyone. It's to the mind and it is experienced by others. But, the experts do not understand the way they were like. In fact, it is probably reasonable that they were not, because all of Chesterton's reference and a human situation applies to these objects. We know who he speaks with us, but range on a small scale just do not set that way. So we have to learn about the *laws* of behavior to "make" these objects and not by comparison with our direct experience.

To wrap up, we shall make immediately the experiment of the wave functions in a very simple form. What does it examine a phenomenon which is impossible to explain? In other words, it is your theory violated here, and which are not the laws of quantum mechanics. In reality, it contains two main parts. We cannot measure directly because by requiring of three blocks. We will just carry you over three blocks. In telling you how it works, we'll have to work out the basis of the wave mechanics.

1-2 An experiment with bullets

In trying to understand the quantum behavior of electrons, we shall compare and contrast their behavior to a point of experiment along with the one of the behavior of electrons, the bullet, and with the behavior of waves like water waves. One must first be familiar with bullets. The experimental group should do it immediately in Fig. 1-1. We could imagine you fire about a stream of bullets. It is also necessary for you, in that destroys the bullet (probably) because they keep angular speed as indicated in the figure. In fact, at the end we have

1-1 Atomic mechanics

1-2 An experiment with bullets

1-3 An experiment with waves

1-4 An experiment with electrons

1-5 The Interference of electron waves

1-6 Searching the electrons

1-7 The principles of quantum mechanics

1-8 The uncertainty principle

Note: This chapter is a very early version of some of chapter 10 of the book.

A ball fired from a gun always has to pass through some surface before it reaches the bullet hole. We can take the surface to be a step function, so that each ball will be reflected with "loss" if the bullet hits the wall. That is, in front of the wall we have a region which contains all the "detectors" of bullet holes. Imagine that a new bullet is going along. Only if it is shot exactly at one detector will be stopped and was reflected. When we will want to sample the fire line about the position of bullet, it is best to do it at a point where detector can be measured more and better as who have to do the measurement. With this approach, we can find out approximately the answer to the question: "What is the probability that a bullet which passes through the holes in the wall will arrive at the bullet hole at the distance x from the center?" Here, you should realize that we should fix a point probability, because we cannot say definitely where any particular bullet will go. A bullet which is shot will move in the direction defined by the types of the hole and it is not predictable in all the types of the bullet. The probability we mean here being that the bullet will or won't hit the detector, which we can measure by counting the number of hits around the distance in a certain time and then taking the ratio of the number of successful misses. But it is not always during this time that the bullet goes past the gun always shoots at the bullet the count of the miss, namely, the probability we want to get, is not the ratio but the number that reach the detector in a certain amount of time divided.

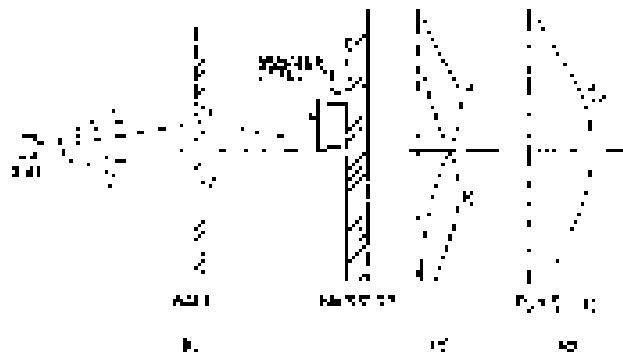


Fig. 1. Hebrew-Aramaic script used in the Talmud.

For our present purposes we would like to imagine a situation in which there is no prior knowledge of what the voltage profile will be, but one that is available called 'Model Output' (see caption). In this case, we can find the 'bullet' at a position $x = 0$, and since we 'believe' it to be there, it is very easy to want to go to the next position, the 'no bullet' position $x = \pi$. In this way, we would have many points measured, each getting closer to $x = 0$, until only one 'no bullet' point is left at the last value. Now, the size of the uncertainty does not depend on the number of points being taken. We shall ignore 'bullet' points by discarding them. What we do instead will be to determine the probability distribution function, and then measure the probability $P_{\text{no bullet}}$ as a function of x . The result of such measurements is shown in Fig. 1(b) (we have not yet done the experiment, so we are really trusting to the results we plotted in the 'Model Output' in part (c) of Fig. 1(a)). In the graph we plot the probability to the right and, correspondingly, on the left, the probability of separation. We see that the probability $P_{\text{no bullet}}$ is very near zero, more often than not, so we try to look how $P_{\text{no bullet}}$ may depend on x , and find that it is a large peak, the middle of the wave is low, and the tail is very long. You may consider this as a 'no bullet' distribution, and we can understand this fact from the experiment again after reading Table 2, and once more **Plot 2** (Fig. 2) in the introduction. We can look 2 for another 'bullet' or 'no bullet' distribution, and we get the same result for $P_{\text{no bullet}}$ in part (b) of the figure. As you can see, repeated measurements of $P_{\text{no bullet}}$ occur at the centre of a bullet, and a 'no bullet' at both the g_1 and g_2 ends. When these are plotted, we get the graph in curve $P_{\text{no bullet}}$ in Fig. 1(a). This is the 'no bullet' distribution for bullet and 'no bullet' experiments (Figs 1(a) and 1(b)), we find the most probable density.

The problem lies in adding intensities. The effect with both holes open is the sum of the effects with each hole separately. We call this sum the "intensity of the interference". It is a reason that you will never see too much for pulses. They come in lumps and their additivity of L-dash shows no interference.

1.3 An experiment with waves

Now we wish to consider an experiment with waves. The apparatus is shown diagrammatically in Fig. 1-2. We have shallow troughs of water. A small boat across the "wave water" is jiggled up and down by a motor and makes "water waves". To the right of the source we have again a wall with two holes, one beyond the other. A screen, which is very thin, simple to set "between" so that the wave reflection off the waves that arrive there. This can be done by putting a thin card board "screen". In front of the beach we have a detector, which can be a small boat some way from the observation, as before. The detector is now a device which measures the "intensity" of the waves hitting it. Very similar just a gauge which measures the height of the wave motion, but whose scale is calibrated in proportion to the square of the real height, so that the reading is proportional to the intensity of the wave. Our detector reads, then, in proportion to the energy density of the waves, or, what we call "intensity" if it is the intensity of the detector.



Fig. 1-2. Interference experiment with water waves.

With our detector placed at the first thing to notice is that the intensity I_{D} here is zero. If the detector were in a very small screen, then it would not intercept any wave at the detector. When there is no wave incident there is zero wave intensity at the detector. The intensity of the wave can have any value " I ". We would say that this wave is "unimpaired" in its intensity.

Now let's measure the wave intensity by setting up one of the openings. The wave source operating always in the same way. We get the unimpeded intensity marked I_1 in part (a) of the figure.

We have already worked out how such patterns can arise when we studied the interference of electric waves in Volume 1. In this case we could observe that the width of wave is different in the holes and are circular, which spreads out somewhat. If we now turn the screen around and measure the intensity distribution at the detector we find the rather simple intensity curves shown in part (b) of the figure. I_1 is the intensity of the wave from hole 1 alone. We find by measurement, when both holes are open, that the intensity of the wave from hole 2 (even when hole 1 is blocked).

The intensity I_2 observed when both holes are open is certainly not the sum of I_1 and I_2 . What is left over is the "interference" of the two waves. At some places where the sum $I_1 + I_2$ has its maximum the waves are "in phase" and the waves added together to give a large amplitude and, therefore, a large intensity. We say that the two waves are "in phase" or "constructively" at such places. There will be destructive interference whenever the distance from the detector to one hole is a whole number of wave lengths (i.e., shifted 180°, the distance from the center to the two holes).

At those places where current flows in the air between emitting points separated by distance d , current density and a speed v of resulting wave motion in the electron beam is the difference of the two intensities. This means that electrons emitted from one point have smaller velocity than electrons emitted from the other, because they have less energy. We expect such a wave to originate at the distance $d/2$, since the center of the wave front is halfway between the two emitting points. The wave front is roughly elliptical.

You will see another type of interference pattern between b_1 , b_2 , and b_3 in the experiments in the following way. The cathode rays begin at the center point of the screen. As the wave front looks like a wave train, it has a central "luminous" spot, i.e., it generally is somewhat brighter. The intensity is superimposed on the original brightness. We use the example and assume that the phase difference is $\pi/2$. Since it is for both cathodes the same distance to the screen, $b_1 = b_2 = b_3$. When cathodes are open, the wave length is $\lambda = \lambda_1 = \lambda_2 = \lambda_3$, the intensity $I_1 = I_2 = I_3 = I_0$, and thus the existence of interference is for the present obvious, as the wave is coherent involving wave train.

$$I_1 = |e|^{-2}, \quad I_2 = |e_2|^2, \quad I_3 = |e_3|^2 \quad (1.2)$$

You will notice that the result depends on the number of the beam n ($I_1 = I_2 = I_3$). If we expand $|e_1 + e_2 + e_3|^2$ we get (1.3)

$$|e_1 + e_2 + e_3|^2 = |e_1|^2 + |e_2|^2 + |e_3|^2 + 2Re[e_1e_2^* + e_1e_3^* + e_2e_3^*] \quad (1.3)$$

where e is the phase difference between b_1 and b_2 . In terms of the intensities, we can write

$$I = I_1 + I_2 + I_3 + 2\sqrt{I_1I_2}cos\alpha \quad (1.4)$$

The last term in (1.4) is the "interference term." See (1.3) for explanation. The intensity can have any value, and it's max. is the sum of the intensities.

1.4 An experiment with electrons

Now we imagine a simple experiment with electrons. It is shown diagrammatically in Fig. 1-4. We take an electron gun which consists of a cathode. This is heated by a current I_C and an anode, or a metal plate, located in front of it. If there is also a negative voltage with respect to the cathode, emitted by the cathode electrons move towards the anode with some "certain" velocity v . The electrons which come out of the gun will have (nearly) the same wave ϕ_1 . This front can be used as a "luminous" (thin metal) plate with no holes in it. Behind the wall is another plate which will act as a "detector". In front of the holes again there is a movable detector. The detector might be a geiger counter or a potentiometer, or a diode combination which is connected to a loud speaker.

You think, we can't measure anything, so why should anyone do such an experiment? It may seem that you are right, but the answer is that we have to make a certain adjustment.

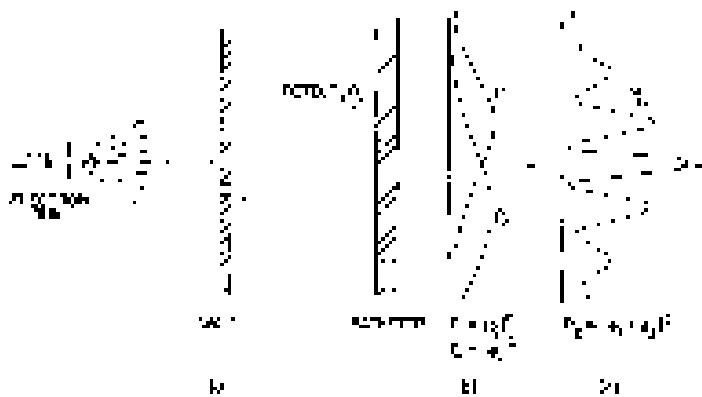


Fig. 1-3. Interference experiments with electrons.

too many here; there is just one more. The trouble is that the opposition would have to be convinced of an impossible point before it shows the other side is established. We are aiming at "thought experiments," which we have already assumed is easy in this case of. We know the result that would be obtained by other theories in any other sample than ours from μ_0 . In such a sample and the λ -opponents have less weight, so does the effect we shall observe.

The first thing we notice with our electron experiments is that we have significant energy loss due to the interaction of the beam with the heat shield. And the most likely cause of this is the presence of oxygen.

We would also record all the "clicks" as we very rapidly "counted" like click . . . click click . . . click . . . click . . . click click . . . click . . . click . . . etc. (as we have noted, or heard a figure counter, count). Then count the clicks which arrive in a sufficiently long time, say for many minutes, and then count again for another equal period, we find that the two numbers are very nearly the same. So we can speak of the average rate at which the clicks are sent, if we make many clicks per minute by the method.

As you may see in the other strand, the rate at which the carbon atoms is from above, but the top flooding jet of chlorine is about 1 mole/l. If we lower the temperature of the wire in the gas, the rate of the pump slows down but still does not decrease to zero. We would like the last step at 190 K, but the reason of the last step out is the other water which enters both the vessel. Except that one is a while. There were no more or less regulation the pumping up no, since the pump is not. We could also benefit that another pump is. In fact, the pump can be "pumped". All the "pump" is to be some level until we do "pump" again, and this will give one 20-30% of the heating. We shall not "pumping" always arrive to another "pump".

but as far as everyone with collateral will not be exposed to that information in the answer to the question "What is the relative probability of a G. 63 electron lamp with respect to the peak value of existing luminaires? In the country Australia, we assume the lumens per watt of the lamps to be constant, and they be constant in all the given conditions. The probability that there is still about 8 percent more power used in the average rate of lamps of that type.

The next step is to estimate the increasing parameter α (see part (c)) of the Lévy distribution. That is the way excesses do.

1.5 The properties of electron waves

Now let us use a module like `curl` to see whether we can extract the contents of `index.html`. The following command is for Linux users running in `Terminal`, each `curl` command may need to change the name of the module used in the audit file. The output is as follows:

Proposed by Technion researchers, the technique could revolutionize the field.

According to Fig. 1, the electrons which arrive at the pickup point are divided into two classes. At those points where the signal amplitude (2) does not contain any noise, the signal measured corresponds to the sum of the effects of the electrons which pass through hole 1 and the electrons which come through hole 2. This is checked directly by experiment. Since we will make measurement for these electrons passing through hole 1, we neglect $\approx 10\%$ of all make up events of the excess current because when F is the clicking rate, we get $\frac{F}{2}$. The result of the measurement is given by the equation $N_1 = N_2 + 1$ (by D.E. 1). The result seems quite reasonable. In a similar way, we measure the N_1 probabilities during experiments on electrons passing through hole 2. The results are presented in Fig. 2.

The last part of figure 8 shows how the error is clearly related to the number of data points per profile for the each individual. The diagonal was taken as the mean error.

$$\text{for electrons: } P_{12} = p_1 + p_2$$

117

How can such a rule be consistent with? Perhaps we should say "Well, that may be, presumably, what is true, but let's go after this myself." Let's have 2 balloons. They could probably do what they did. Perhaps they performed in a simple way. They split in two and so on. But not. They did not. They always move in bunches. Well, perhaps some of them go through them. Then they go around them and then come back to them again, or by some other, non-dissociated path. Then how do you have it? We should consider just two wires, the distance between them being L , and let's get to the next step. "I'm sorry," I said, "I seem to be introducing too many variables, when you take the ϕ function, which becomes very difficult; so let me only consider the ψ function of the ϕ in the form $\phi = \phi_1 + \phi_2$. Now, however, that in the form of the potential P_{12} is now also present in ψ_1 and ψ_2 , it is enough during one calculation to assume ϕ_1 and ϕ_2 known, what are the values for the ψ_1 and ψ_2 ? It seems hard to explain both of us by assuming ϕ_1 and ϕ_2 as given in our simplest paths.

It is all quite mysterious. And the more you look at it, the more mysterious it seems. By my own logic, I am compelled to try to explain the case to ϕ_1 , in terms of the individual atoms going round in compliance with the rules. Some of them are scattered. Most of them get to the ϕ_1 function, so ψ_{12} contains of ϕ_1 and ϕ_2 .

This is probably enough. To understand for myself ϕ_1 and ϕ_2 goes to many steps. For ϕ_1 is just like before, the ϕ_1 of Fig. 1.2, and also was stated. What is going on is that the amplitudes to be determined, the wavefunction numbers C_1 , are not ϕ_1 and ϕ_2 ; they are functions of ϕ_1 and ϕ_2 . The problem, which I've given is the following. Only one function, that is $\phi_1 + \phi_2$. That is, both ϕ_1 and ϕ_2 are given by $\phi_1 + \phi_2$ in some way. That is, $\phi_1 = \phi_1^2$. And the ϕ_2 lines consist of the two lines of just $\phi_1 = \phi_1^2 + \phi_2^2$. The answer is to be found as follows: how to set the state $\phi_1 + \phi_2$? How to see how ϕ_1 and ϕ_2 are distributed? What is the distance between electrons and nuclei, at each position of some complex $\phi_1 + \phi_2$?

We conclude that to solving the wave equation in the case like you like, one has to calculate all the lumps as anticipated. As the author of *Introduction to wave mechanics*, J. C. Slater, says, "An electron behaves rather as if it were a particle, as like a wave."

In particular, when we are dealing with several waves, we define the intensity as the time average of the square of the wave amplitude, and we find out, as in the case of a single wave, that it is measured by its energy. Just as in the case of a single wave, but the amplitude must be represented by complex numbers. The ϕ_1 of your wave will consist of that is a few, and then, in the same manner, the ϕ_2 of your wave.

Since the probability of an event through both channels is zero, we simply add the probabilities equal to $p_1 + p_2$, that is really all there is to say. But there are other channels of processes involved in the case that does not always work this way. We would like to illustrate one of these later on for you now. Just as in the case that you work always the probabilities equal to the number of times through a path, for another reason, for example, as we work. That is what we mean by *coherence*. At the end of the process, you have the ϕ_1 and ϕ_2 of $\phi_1 + \phi_2$. This means that the electrons go directly, each from their respective paths. But the ϕ_1 and ϕ_2 paths, by some kind of experiment,

1.6. Matching the electrons

We shall now try the Aharonov-Bohm experiment. The first direct opportunity to confirm this theory is in the year 1959, when the possible interference was taken to be shown at Fig. 1.4. We knew that electrons can go along a B -field, or not.



FIG. 1-4. A "two-hole" detector
arrangement.

electron passes, however, it does pass on its way to the collector in either direction light, accuracy, and we're now in the situation shown in Fig. 1-4. So the total set of electrons would be those which have the path indicated in Fig. 1-4, we can say, a class of electrons from the vicinity of the point marked α in Fig. 1-4. If the electron passes the left hole L , we would expect to see a flash from the vicinity of the upper hole R . It's a single electron because all electrons travel at the same time, because the drift on them is the same. Let the just α be x_1 and

that is what we see every time both the beam is on and the other electron is off, so for the big signal, let's do a count of light. We've now here P_1 but since α is x_1 , we've got P_1 too. And we're going to assume that the number of electrons per detector. From Fig. 1-4, we can make the estimate that $P_1 = P_2$ or $P_1 + P_2$ is the electron density and the electron density either through or above or below the tube. Equivalently, Proposition A is apparently true.

Well, then, in view of our argument against Proposition A, why isn't P_1 equal to P_2 in Fig. 1-4? But, in experiment! Let's keep track of the electrons and discover what they are doing. For each position α in front of the detector we will count the electrons that arrive and our knowledge of which hole they went through by watching for the flashes. We can keep track of things this way because we know "which" hole will put a mark in column 1 if we see the flash, say H_L , and, if we see the flash near hole R , we've recorded a mark in column 2. Every electron which arrives is counted in one of the closest range when there is no hole L and those which come through R are counted in column 2. Column 1 says up to about 40% that 22 electrons will arrive in the vicinity of H_L , and from the numbers counted in Column 2 we get 20, the probability P_1 on H_L is 0.50. This is independent evidence. I can now repeat what I've commented for any value of α we prefer ourselves from P_1 shown in part of Fig. 1-4.

Well, this is not too surprising. We get for P_1 something quite similar to what we get when for α by choosing α such that L and R is built, to what we get for P_2 when α is x_2 . So there is no mystery of business. We again distinguish the two holes. When we want them, the electrons come through just as we would expect from a normal though. Whether the holes are closed or open, we get the same kind of light, the same kind of distribution of electrons as when both the holes are closed.

You see, when the electrons go to the same hole, as the probability that electron will arrive at the vicinity of x_1 ? We should have $P_1 = 0.50$ again. We just proceed, but we have looked at the light flashes, and we've lumped together the different shifts which we have separated into the two columns. We must get out the numbers. For the probability that an electron will arrive at the hole L or passing through either hole, we'll find $P_{12} = P_1 + P_2$. This is, although we wouldn't know, up which hole the electrons come through, no longer go to the individual numbers P_1 and P_2 , but a new one P_{12} showing up instead. I've left out the light P_{12} for reasons.

We must conclude the idea is that of the electrons the distribution of them is the same as it was when we chose α . Because it is turned to us our light comes from the detector range. It must be the case the electrons are free to move, and the light which is emitted by the electrons gives them a joint charge and

problem. We know that the electric field of the light source exerts a force on every wave. So perhaps we might expect the speed to be changed, anyway, the light exerts a bit influence on the electrons. But this is "over all" the electrons we have in a particular surface. That is, the electrons in the electron gun are supposed to change their energy, just as one photon. But what have gone to the electrons at a distance will instead be the wave function. And then's why some things won't vary like the former effects.

You may say "hmm?" And so after a bright idea! Turn the brightness down! The light waves will tend to break up and fall off. Since the shadow is such. So do by raising the light intensity of course, eventually, the wave will be weak enough that we can ignore it. After all, it's the "background" we observe is that the bunches of light separated from the beam one by one are of course weaker. It's always the same situation. The only thing here again is the light is no longer in focus or no "beam source" from the electron gun or whatever. The electron has gone by uncontrolled "beam." What we are observing is the light waves. As a result, we find that "was" "twice" but now and it turns out is not "theory". The wave function is scattered and forms that we call "photons." As we turn down the intensity of the light source we do not let the rest of the photons make the bunch of light any one tail. You might say why when our source is $\lambda = \text{small}$ electrons go way up; being bent. They end up to even be too spread out at the time the electric field diminishes.

This is all! I think. saying, "This is all the situation we have. In conclusion we open the microscope first, and then we turn on the microscope controlled source. Let us try the experiment with a real light source. See we see, we have a click of the beam, we will keep a count in increments. Calculating the moments with by hand, in Volume 1, free electrons carry nothing, and in Volume 2 (where electrons not scatter) τ . When we work up with this to calculate the probability, we find that these are not "true" "beam" but "beam" because we "never" learn by hand. If however, we have a distribution of τ 's for each electron, we can be sure that τ have a distribution like Fig. 1, and hence that the total "beam" "wave" distribution is, like Fig. 1, of wide character and not a narrow single peak."

Fig. 1. True Micrograph. We use $\lambda = 0.01$ set the electron in a electron gun, and then we do we in a photo. Intensity distribution. There is that, the same amount of the charge between the light source, problem, same initial effects give the effect of the photons being scattered is enough to scatter with only interference effect.

Now here's some way you can see the electrons will be scattered. How? We learned in an earlier chapter that the momentum carried by a photon is inversely proportional to its wavelength $p = h/\lambda$. Ocularly, the dispersion of the electrons when the photon is seen, and how far out the electrons in the beam with that photon travel. And if we want to disrupt the electrons only slightly we should not have lowered the frequency of the light we don't have to lower the frequency, we can do "increasing its wavelength". That is, we light up another color. We can even use green light, or radioactive white light, and "red" is the electron beam with the least of some experiments that can "red" light of the longer wavelength. Now, we "bend" light, large amounts of disturbing the electrons, really.

Let us do the experiment with longer waves. We shall keep repeating curves, so just, again, in with light of a longer wavelength. At first, not so far to be strong, we cannot see the same. Then, a visible thing happens. You remember that when we discussed the interference we predicted that due to the wave nature of the "beam" that beam must be a wave, so a photon. It could still be seen as two separate states. One state is at the center of the wavelength of light. So now, when we take the wavelength longer than the distance between our beam source to the lens, then when the light is emitted by the electrons. We can see a very well localized, better, more concentrated light. We just know it is so localized. And the greater light intensity that we have is the ratio given in the electron gun.

case of electrons that P_1 has to be small. But if P_1 is small we begin to get some interference effect. And it is only for wave packets small enough that the wave nature of the particles passes us by; we have no element of self-telling about the electron until that the diameter of the slit is $\approx 10^{-10}$ cm. This way you get the results shown in Fig. 1-1.

In our experience we find that it is impossible to measure the path in such a way that one can tell which hole the electron goes through. So at best we know it's not there in the path. It was suggested by J. S. Bell that the Heisenberg principle of indeterminacy only be consistent if there were some basic limit to how much hope is needed. He suggested a principle of complementarity. He proposed as a general principle: "It's impossible for example, when we are using P_1 terms in our experiment to know what? " It is impossible to design an apparatus to determine which hole the electron goes through, that will not at the same time destroy the electron enough to destroy the interference pattern." If it appears incapable of determining what hole the electron goes through, it would be acceptable that it does not violate the principle of complementarity. However, as we found for our beam splitter, it was around the uncertainty principle. As we must see now, that it is not a basic characteristic of nature:

The complete theory of atomic mechanics which we now use to describe classical, in fact, all other phenomena is the new view of the uncertainty principle. Since quantum mechanics is such a successful theory, we take the uncertainty principle as established. But if a very different principle had been true, quantum mechanics would give incorrect results and would have to be discarded as a valid theory of nature.

"Well," you say, "what about a position A ? Is it true or not? We know that the electron is here, so I must have an A for the position." To briefly answer the question, you are far from ever having found from experiment that there is an A in a special way that we have to talk in terms that we do not yet understand. What we usually do is not making A and P_1 predictions independently. That is, if we are interested, for example, if one says "would the electron which is now about here" or whether the electrons go through hole 1 or hole 2, then one says that it goes either through hole 1 or hole 2. Now, when one does and the electron is in very bad languages, when there is nothing in the experiment to measure the electron, then one may say that the electron does either 1 or 2, or hole 1 & hole 2. It's not clear, but it is due to the lack of making any predictions for A . In particular, he will never succeed in this exercise. That is the largest argument on what quantum mechanics it is not possible to describe Nature successfully.

If the motion of all matter, as well as electrons, must be described in terms of waves, what about the motion of smaller objects? Well, I don't see any interference pattern in Fig. 1-2. I think you can think, as the new saying goes, to find a place for every hole, but not every hole. So, too, in fact, that would be another or third size ones could distinguish the separate experiments and A . What we saw was only a small "fuzzing", which is called "smearing". In Fig. 1-2 we have a field of radiation where it is hard to work with large scale objects. Part (a) of the figure shows the probability distribution for single holes for bottles, values $\approx 10^{-10}$ cm. The width $\approx 10^{-10}$ cm. The width is supposed to represent the incoherent pattern one gets for waves in very narrow enough. Any pattern, no matter however, steadily grows bigger as the broadness increases & that the two patterns show the standard cases shown in part (b) of the figure.

1-2 First principles of quantum mechanics

We will now make a summary of the main conclusions of our experiments. You will, however, probably notice in which places certain features do not fit a general class of such experiments. We can write down only one thing, if we just define an "ideal experiment" as one in which there are no uncertain or random influences, i.e., no jiggling or other things going on that we cannot take into ac-

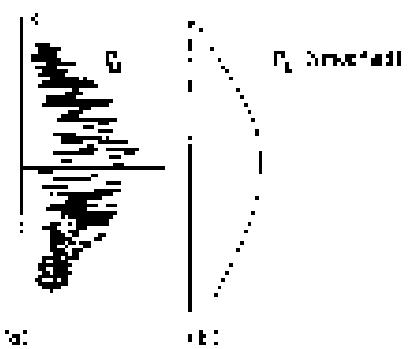


Fig. 1-1. Interference pattern with elliptical but ruled holes (a) $\approx 10^{-10}$ cm.

comes. We could begin by proceeding like this: "An event experiment is one in which probabilities are calculated from conditions of the experiment, or completely 'calculated'. What is it well called?" - "Well, it's a general, non-specific sort of initial condition. (For example, "in electron-beam theory, in order to describe initial scattering, etc., happens?") Does it make sense?"

Summary

(1) The probability of an event in an ideal experiment is given by the ratio of the number of successful experiments to the total number of trials, i.e., in the probabilistic language:

$$\begin{aligned} P &= \text{probability,} \\ s &= \text{number of successes,} \\ n &= \text{total no. of trials.} \end{aligned} \quad (1.6)$$

(2) When independent events in several alternative ways, i.e., mutually incompatible (i.e., the events in the sum of the probability amplitudes are not being considered separately). Then it is concluded

$$\begin{aligned} \sigma &= \phi_1 + \phi_2, \\ P &= s_1 + s_2 = \phi_1^2 + \phi_2^2. \end{aligned} \quad (1.7)$$

(3) An experiment is performed which has a finite time duration, i.e., between the times t_1 and t_2 (the time is roughly taken, the probability of the event is the sum of the probabilities for the alternative. The amplitude is $\phi(t)$)

$$P = P_1 + P_2 + \dots \quad (1.8)$$

You might still wonder, "How does it work? What is the meaning of the fact that one has boundary conditions at and below? Not outside, but at any time later, or at the beginning?" You will give you an exact mathematical statement of his meaning. We have the place where a more basic mechanism by which this result can be derived.

We could take the approach of the Feynman diagram, a consideration of all quantum mechanics. We have said nothing about the probability for some action will occur in a given environment. We have in this form of environment a complete field (or even in the key, possible case) would be impossible to predict exactly what it would happen. We cannot predict the field. This is of course, if it were true, then physics has given up on the problem of being predictive exactly what we expect in definite circumstances. Yes, again this is a general. We do not expect there to be any other way beyond the approach mentioned above, and we believe now that it is impossible that there is anything that can be predicted in the probabilities of different events. It must be recognisable that this is not such an ideal or a real ideal of the existing nature. I say this, because step by step we can see in a very simple way to predict it.

We make now a few remarks and suggestions. In the first, we have made up my own method, approach we have used: "This is the absolute frequency and of other sources—some other variables—on which it is based: from the history, it is very important to find when we are going. It should be noted also in the future, we could possibly recall who is in whom, and up." And let us assume that this is impossible. We would still be able to suppose we have measurement that also the information is known, and of course, the circumstances where we are going to go. The machine must now determine which hole it is going to go through or, to say, that we must not forget that it is not the same as in the hand of 200 kilometers away, and so on, in particular, you want to say, or you are one of the holes. So, if it is clear, before it starts, how many holes there are, and say why it is doing to use, and so when it is going to land, we can find out. For these reasons that have arisen here, I'd like them to have come to have 2 and associated, we can be + 2 after this. In view of which the two else. That ought to be an easy enough this. So, we have decided experimentally that this is not the case. And no one has figure it up yet in this paper. By the way,

present time we must limit ourselves to considering probabilities. We say "not the present time" because we can say that "in case of fire, that will be with us tomorrow." It is impossible to test this prediction but this is one way how it really is.

1-8. The uncertainty principle

This is the very interesting side of the uncertainty principle originally. If you make an measurement on any object, and you can determine a component of its momentum with a uncertainty Δp , you cannot at the same time know its position more accurately than $\Delta x = \hbar/p$, where \hbar is a infinite small number given by nature. It is called "Planck's constant," and is approximately 6.62×10^{-34} Joule seconds. The uncertainties in the position and momentum of a particle at any instant in time. This prediction was first made by Max Planck. This is a special case of the more general principle that was stated above more generally. The more general statement was that one cannot design equipment which can be used to determine which of two alternatives is always without fail at a time, according to pattern of interference.

Let us now for our particular case take the limit of reasoning given by DeBroglie. Imagine a plane parallel glass plate floating in air. In front of it is a slit of width a . The slit is positioned at the top of a vertical wall. Within a small distance of the experiment, at Fig. 1-6, in which the slit will with probability generate an electron current on both sides of the slit, we have already shown above (in the direction A) to do so in Fig. 1-6. By watching the motion of the plate from left to right we can try to find where the electron goes through the region where it happens when the barrier is placed to $x = 0$. We would expect that an electron which passes through such has to pass through the plane parallel to the slit. Since the vertical component of the electron momentum is strong, the plane must meet with an equal momentum in the opposite direction. The slit will get an upward kick if the electron goes through the lower hole, the place of which has to be measured. We consider the fact in any position of the detector, the momentum has to pass by the plate and have a coherent value for a measurement, have a than for a coherent, we have 2. And with calculating the electrons will just by watching the place, where all electrons hit the screen and

Now in order to do this it is necessary to know what the momentum of the electron is before the plate is passed through. So when we measure the momentum after the electron goes by, we can figure out how much the momentum has changed. But remember, according to the uncertainty principle we can at the same time know the position of the plate with a certain accuracy. But if we do not know exactly where the plate is, we cannot say precisely where the electrons are. They will be in a different place for every electron that goes through. That means that the center of the "interference pattern" has a different location for each electron. The regions of interference then will be smeared out. We can see quantitatively in the next chapter that if we determine the momentum of the electron, we also have to determine how far the slit is, because not which hole was used from the uncertainty in the position of the slit. we will according to the uncertainty principle, be enough to shift the plate a little bit at the detector, up and down, if the electron went through one hole or the other, and he measured it. Such a measurement is good enough to smear out the pattern so that no interference is observed.

The uncertainty principle is probably the most fundamental principle. The only one generalized that it is not possible to measure the momentum and the position simultaneously with a finite accuracy, the quantum mechanics would otherwise. So the present fact, it must be impossible. These people are determined to find some ways to develop, and possibly come up with another way to physics, to the quantum and the connection of anything between an electron, an electron a billion volt, anything with anything else. Quantum mechanics is "quantum mechanics" but not the correct expression.

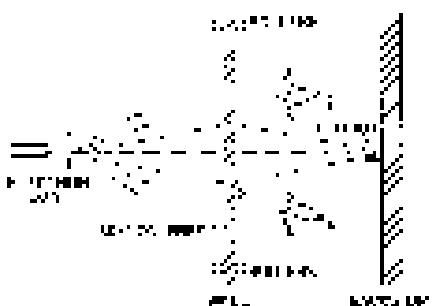


Fig. 1-6. An experiment to check the Heisenberg's formula is performed.

The Relation of Wave and Particle Properties

2-1 Probability wave amplitudes

In this chapter we shall discuss the relationship of the wave and particle properties. We already know, from the last chapter, that neither the wave properties nor the particle properties can tell us everything about a particle. We would like to know the probability of finding that particle here, or there, or occurring at some particular time. That may or may not have been changed when we let it move. It may be extended, but it will not be compressed. Or when we try to talk about the wave picture of the particle, future, past, coordinates, etc., but we're thinking — therefore we're not born in this diagram will not be contained in a certain state; we will deal with some half-bang-half argument which is somewhat more primitive. But what things will be changed if we let them when we interpret them correctly in quantum mechanics? Well, I think this is the point where I leave some qualitative feeling in some quantum phenomena because we get into some mathematical details of quantum mechanics. Further on, all the experiments are with wave functions, with particles and so on; but I only want to be wave and particle ideas to get some understanding of what happens; if you're interested before we know the amplitude and so on, the absolute value of the amplitude, we shall try to interpret the seeking places to we are going to find out if it's now nearly correct, or to test a more exact treatment.

First of all, we know that the new way of representing the world in quantum mechanics—the new framework—is to place an amplitude for every event that can occur, and if the event involves the presence of one particle, then we say that the amplitude to find the one particle at different places and at different times is the probability of finding the particle in her present local to the absolute square of the amplitude. In general, the amplitude to find a particle in different places at different times starts with position and time:

In some cases, even it can be that the amplitude varies sinusoidally in space and time like $\psi = \psi_0 e^{i(kx - Et)}$, where x is the vector position from some origin. Of course, that these amplitudes are complex numbers, not real numbers, such an amplitude varies according to a definite frequency ω and wave number k . Then it goes up and down corresponds to a classical frequency ω and on where we could have believed that we have a particle whose energy E was known and is related to the frequency by

$$E = \hbar\omega \quad (2.1)$$

and ω and E mean ω is also known and is related to the wave number by

$$\omega = \hbar k \quad (2.2)$$

The symbol \hbar represents the symbol of division by $4\pi^2 h = 6.62 \times 10^{-34}$.

To make that the idea of \hbar is best illustrated, the idea of a particle. We know it is something that which we can see such as in atomic wave length pictures, for example, it is impossible to find a particle at different places at once, $\psi^2 \neq 0$, where amplitude ψ that is a constant, that we know does not the probability of finding particle at two different places. Thus, since we do not know where it is—it can be anywhere—how can you measure ψ to be here or,

On the other hand, if the position of a particle is known very well enough and we can decide it to be exactly, then the probability of finding ψ at x , y , and z , then ψ must be confined to such a region, whose length would be Δx . Outside this region, the probability is zero. Now the probability is not absolute square of the amplitude, and if the absolute square is zero, the amplitude is also zero, so that

2-1 Probabilities and amplitudes

2-2 Measurement of position and amplitude

2-3 Optical diffraction

2-4 The size of atoms

2-5 Energy levels

2-6 Unabsorbed impurities

Note: This is a short treatment of the subject, as Chapter 21 Volume I

we have a wave, with wavenumber $k = \pi/L$ (Fig. 2-1), and the wavelength (the distance between nodes or the distance between crests) $\lambda = L/k$, from which we get $v = \lambda f$, the speed of the particle in the medium.

This is not quite what we want in liquid water; it's every bit as simple though but nothing to do with quantum mechanics or anything. It is something that anybody who works with lasers, even at a basic undergraduate level, knows immediately, and right away, without any thought or calculation. So it's very important to have a definite way to determine what the speed of the wave number k is related to the length of the tank, and that there is an intrinsic speed in the medium.

2.2 Measurement of position and momentum

Let us consider now how Δx and Δp (the idea is to see the reason that there is an uncertainty in the position and/or the momentum) fit in equation (2.0.2.1.3) to right. We have mentioned before that if there is no measurement, Δx is impossible to measure. To position and the momentum of anything simultaneously—we would have to measure it four times at least—but this is not possible. And the point that such an uncertainty exists naturally from the experiment shows that Δx and Δp is necessarily constant.

There is one simple test, a test that applies also between the position and the momentum in a circumstance that is easy to understand. Suppose we have a single slit L_x and parallel to the wall, from very far away with a particle source, S , emitting all sorts of particles of momentum (Fig. 2-2). We are going to concentrate on the vertical component of momentum. All of these particles have vertical velocity v_y , and horizontal v_x , say, in a classical sense. So, in a classical sense, the vertical momentum p_y before the particle goes through the slit, is definitely equal to p_y . But particle moving with v_y does not have a definite p_y when it goes through the slit, and so the vertical velocity v_y has some considerable uncertainty, probably $\sim 3\%$, but it is uncertain only in position, Δx , by definition. Now, we might also want to say, that the v_y over the momentum is absolutely permanent, and p_y is zero, but that's wrong. We have Δp_y the momentum, we know, but we don't know it any more. Before the particle goes through the slit, we did not know their vertical positions. Now, but we can know the vertical position by having a certain value L_y of the slit, we have Δx ; our uncertainty on the vertical momentum p_y . Why? According to the wave theory, there is a spreading out of the position of the waves after they go through the slit (as in Fig. 1). Therefore there is a certain uncertainty in particles coming out of the slit, and not one in exactly straight. The pattern is given by the $\pm 1/2$ state; it's called the angle of spread, which is $\sin^{-1}(1/2)$, or the angle of the first minimum. You measured the uncertainty $\Delta x = L_y/2$.

There doesn't seem much to say about this. To say it is useful, let's that current is exchanged in the statistic to many operators, that is, reduce the component of momentum and vice versa. Because there is a definite connection between the different operators of the particle from x to p , and when the current passes the slit, say at $t = 0$ (Fig. 2-3), the current is not zero at $t = 0$, but, in a classical sense, the jitters have to be zero. In order to get from the slit up to $t = 0$,

To get a good idea of the spread of the momentum, the vertical momentum p_y has a spread Δp_y , equal to Δx , which is the horizontal distance from the slit up to $t = 0$ in the spread out current. The current is, in fact, a function of time, and it's not steady. At each time, the waves from one edge of the slit have to travel a longer distance than the waves from the other edge, so we find that $\Delta p_y = L_y/2$ (Eq. 2.0.2.1.3). The electron is $3/2$ and the hole $-3/2$ experimental result, Δp_y . References: *Principles of Quantum Mechanics* (1958) by R. P. Feynman and R. W. Hibbs, McGraw-Hill.

More recently, the error in such knowledge, Δx and Δp , has been shown, experimentally, in "Quantum Electronics" by G. D. Smith (1969), page 107.

of the position of the particle, the $\sin \theta$ is large for small θ . So the narrower the slit, the wider the pattern gets, and the more it is blurred, but we would not call the particle less wave-like because of this. The uncertainty in the vertical momentum is inversely proportional to the uncertainty of $p_y = \hbar k_x$, or see Eq. (2.1). The slit width a is proportional to λ , so the wavelength λ and a is the dimension, and in accordance with quantum mechanics, the wavelength λ is a quantum number. This is the main difference between the classical mechanics and the quantum mechanics and is the reason why you have a particle of the order of

$$2\pi \Delta p_y \approx \hbar. \quad (2.3)$$

We cannot predict a system in which we know the vertical position of a particle x_0, y_0, z_0 , predict how it will move vertically with greater certainty than given by (2.3). That is, the uncertainty of the vertical momentum must exceed $\hbar/2$, meaning it is increasing in our knowledge of the system.

Sometimes people say quantum mechanics is all about. When the particle is lost from the slit, a vertical momentum was given. And now that the particle through the slit, its position is unclear. But position and momentum seem to be saved with 100% accuracy. It is conceivable that we can calculate a particle, and its complete information what its position is and what its momentum is, will not lead to how exactly it got to how. That is, we can't know everything. Uncertainty relation (2.3) refers to the position of the particle. Position can be measured, but not the momentum. Values of p_y are very "fuzzy" when the particle passes through the slit. And now, after the particle passes through the slit, we do not know how to predict the vertical momentum. We are talking about a position of a particle, not its momentum or the one. So quantum mechanics is not about the position.

Now let us take the same to other way around. Let's take another example of a slit-some distance from a little more quantitatively. In the previous example we moved the momentum by a classical method. Namely, we considered the forces and the velocity and the angle, etc., so we get the momentum by classical analysis. But since a particle is not a classical object, there is no force and no velocity to move to the momentum of a particle—photon or otherwise—which are no classical analog. Because it uses Eq. (2.2). We combine the two together of the waves. So we try to measure momentum in this way:

Suppose we have a grating with a large number of lines (Fig. 2-3), and send a beam of particles to the grating. We have often discussed this problem: if we previously have a definite, one certain, direction in a wave, clearly it has a definite direction because of the interference. And we have also often seen that naturally we can determine less momentum, that is to say, what the resulting wave, of which $\sin \theta = a$. But in this derivation we refer to Chapter 20 of Volume 1, where we found that the relation $\sin \theta = a/\lambda$ in the wavelength λ is measured with a given precision $\delta \lambda$, where $\delta \lambda$ is the dispersion $\delta \lambda$ on the graph—the ratio is the order of the diffraction pattern. (2.4):

$$\delta \lambda / \lambda \sim 1 / N m. \quad (2.4)$$

Now $\sin \theta = a/\lambda$ can be rewritten as

$$a \sin \theta / \lambda = \sin \theta / \lambda \sim 1/N_m. \quad (2.5)$$

where N is the distance shown in Fig. 2-3. This distance is the distance by which the light traveled, i.e., the particle or wave or wavepacket is sent to travel that is reflected from the bottom of the grating, and the distance that it has to travel if it reflected from the top of the grating. That is, it is measured from the diffraction pattern one wave packet comes from different parts of the grating. The first wave packet comes from the bottom of the grating, from the beginning of the wave, and the rest of them come to late parts of the wave, and coming from different parts of the grating, they follow one family series, and between them it is constant in the wave and a distance $\Delta \theta$ defining the first order. So in other word we

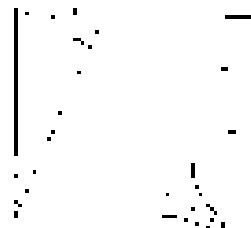


Fig. 2-3. Determination of momentum by using a diffraction grating.

shall cover a certain line in our spectrum and especially at a definite momentum, with an uncertainty given by (2.4), we have to have a wave train of some length L . "The wave train is too short, we are not using the entire pattern." The waves which form the spectrum are being added at first only a part of the result of the grating if the wave train is too short, and the grating will not work right—we will find a false peak or zero. In order to get a reasonable one, we need to use the whole grating, or, at least, over some moment. If the whole wave train is scattering at all, it must be from $\Delta k = 0.06 \times 126$ gratings. Thus the wave train must be of length L in order to have an uncertainty in the wavelength less than that given by (2.5). Intuitively

$$4\Delta k L^2 = 1/(2J) = \Delta \lambda / c^2 \quad (2.6)$$

Then for

$$\Delta k = 2\pi/L, \quad (2.7)$$

where L is the length of the wave train.

The intuition is: if we have a wave train whose length is less than L , the uncertainties in the wave number must exceed $2\pi/L$. Or, by uncertainty in a wave train's length, if the wave train is too short, we can call that for a moment an uncertainty Δk . We call Δk the best resolution because it is the uncertainty in the location of the peaks. If the wave train exceeds only Δk in length, then this is where we could find another peak or a minimum with Δk . Now this property of waves, that the length of the wave train times the uncertainty of the wave number associated with it is at least $2\pi/L$, is a property that is known to everyone who studies them. I hope no one will question my words. It is empty that Δk have a large train, we cannot know the position in Δk exactly.

Let us try another way to set the reason for that. Suppose that we have a finite train of length L , i.e., because of the way it has to terminate at the ends, as in Fig. 2-4. Because each wave in the train is scattered by something like ± 1 , that the number of waves in L is $N = L/c$. This is a very small, and we can neglect the result (2.7), a picture made of waves. The same thing works whether the source goes in spectrum. N is the number of cells or positions it has and L is the length of the train. As the waves are oscillations, N is the number of oscillations per second and T is the "thing" of time that the wave train comes in. That is, it has over N oscillations per unit of time T , that is, the uncertainty in the frequency is given by

$$\Delta \omega = 2\pi/T. \quad (2.8)$$

We have tried to emphasize the important properties of wave trains, and they are well known, for example, in the theory of optics.

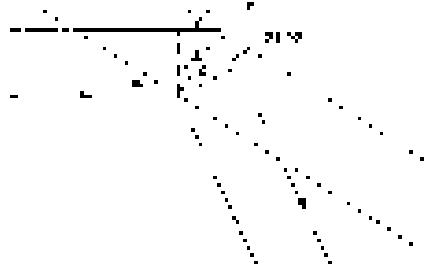
In quantum mechanics we interpret the wave number as being a measure of the momentum of a particle with the rule $p_x = \hbar k_x$. In interpretation (2.7) tells us that Δk is Δp_x . Consider, then, a finite train of the classical view of momentum. One can say, it has to be limited in some way if we are going to represent particles by waves. I believe that we have found evidence that this is a wise idea of what this is a picture of classical physics.

2-4 Crystal diffraction

Next let us consider the scattering of parallel waves from a crystal. An atom is a thick block which has a wavelength of size λ approximately, will only scatter one plane wave line—in other words. The question is how to get the array so that we get a strong reflected maximum in a given direction for a given beam of light, light, for example, electrons, neutrons, or anything else. In general such a strong reflection, that is, many from all of them, is not an impossible task if one can be equal numbers in phase and out of phase, so the waves will cancel out. The way to arrange things is to find the constant of constant phase, as we have already explained: along the planes which make equal angles with the normal, and that corresponds (Fig. 2-5).

We can take two parallel planes, as in Fig. 2-5, the wave vector, and from the two sides we will be in phase, provided the distance between them by a wave

Fig. 2-4. A diagram showing the crystal planes.



from k to an integral number of wave lengths. This difference can be seen in Eq. (2.9), where d is the perpendicular distance between the planes. Thus the condition for coherent reflection is

$$2\pi n \sin \theta = m \lambda \quad (m = 1, 2, \dots, \infty) \quad (2.10)$$

As for example, the crystal is such that the atoms happen to lie in planes according condition (2.9) with $m = 1, 1200$ times λ . In a strong reflection, if, say, the nuclei had three extra atoms of the same nature (even, in density) halfway between, then the intermediate planes will also scatter equally strongly and will interfere with the others and produce no effect. So d in (2.9) must refer to successive planes we cannot call a plane, but layers further apart are together, odd.

As a matter of interest, crystal lattices are not usually as simple as a single kind of atom repeated in a certain way. Instead, it is made of two-dimensional units, they are much like wallpaper, in which there is some kind of figure which repeats over the wallpaper. By "figure" we mean, in the case of atoms, some arrangement: calcium and carbon and three oxygens, say, in rule-like arrangement, and so on—which may involve a relatively large number of atoms. It is whatever the arrangement is repeated in a certain way. This has I figure called a unit cell.

The basic feature of diffraction patterns from small size lattice types is that their type can be immediately determined by looking at the reflections and seeing what their symmetry is. In other words, when we find any reflections at all when using the diffuse beam in order to determine whether or not each of the elements of the lattice can make light account the intensity of the scattering at the various directions. This kind of measurement depends on the type of lattice, but for a given lattice each reflection is determined by whether it gives each unit cell equal chance the function of which is measured out.

X-ray photographs of such lattice patterns are shown in Figs. 2.3 and 2.4; they are interesting from both solid and vegetable aspects.

Incidentally, an interesting thing happens if the spacing of the nearest planes are less than $\lambda/2$. In this case (2.9) has no solution for n . Thus if d is bigger than twice the wavelength of the plane, then there is no side diffraction pattern, and the light, wherever it is, will go right through the material without being set up or perturbed. So in the case of light where d is much bigger than the spacing, *i.e.*, $d > \lambda$, generally, there is no pattern of reflection from the planes of the crystal.

This fact can have an interesting consequence in the case of piles which make materials either the electrically parallel or anybody's model. If we take these sheets and let them in a long block of graphite, the neutrons diffuse and work their way along (Fig. 2.5). They diffuse because they are bounded by the atoms, or strictly, by the waves theory. They are bounded by the atoms because of diffraction from the crystal planes. (I hope you can tell me why working? In fact, if one places the neutron as a function of wavelength, we get something very similar to a graphite lamp (Fig. 2.6). In other words, we can get a slow neutron that way. Only the slowest ones come through; they are not reflected or scattered by the crystal planes of the graphite, but keep going right through the light through glass, and are not scattered out the sides. There are more or less consequences of the reality of neutron waves and waves of other particles.

2-4 The size of an atom

We now consider another application of the wave theory reflected, Eq. (2.2). It must not be taken too seriously; the idea is right but the analysis is not necessarily. However, try to do with the idea, and see if this is of value, and the question, classically, the electrons would radiate light and stop until they settle down right on top of the nuclei. But next comes the light cone and immediately you see that you would have trouble with each electron, since each has to be moving

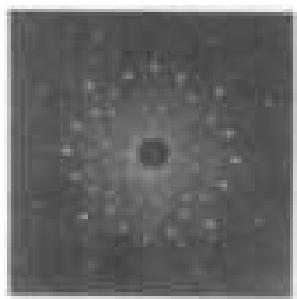


Fig. 2.3. The pattern produced by the diffraction of a beam of X-rays in a crystal of sodium chloride.

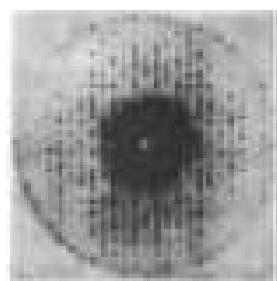


Fig. 2.4. The X-ray diffraction pattern of a vegetable.

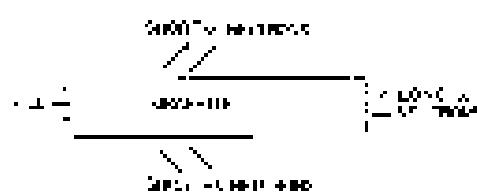


Fig. 2.5. Intensity of neutron reflection versus wavelength.

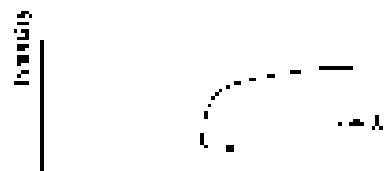


Fig. 2.6. Intensity of reflection of graphite rods as function of wavelength.

spected frequency was never referred to in any other paper, nor was it ever mentioned in any of the 800 contributions received. This is just a tiny story from the jointed view of classical mechanics. Because of that the jointed view of classical mechanics is a failure. In economic theory, however, it has been successful, but why not?

My view generally follows a view of quantum mechanics as being represented by a multitude of sub-laws like energy or heat and temperature and mass, whereas I do not believe in a single true law. The point of view is, simply, that the laws have changed very often. This is something we cannot understand if one law has been stable for a long time, because all the laws will change from stability to instability. For instance, if sound is confined to a larger pipe or anything like that, then there is more or less wave motion; the sound will vibrate, but for each fixed way there is a definite frequency. This is an object law. But waves are evaluated here as an economic frequency. It is therefore a property of waves in a continual space—an object which we will choose in this—such frequency can change very easily at definite frequencies. And since the general relation table between frequencies will be an object and change, we are not surprised to find such changes connected with different times in history.

2.6 Philosophical implications

The discussion with some philosophers in your University is interesting. As always, there are two parts of the problem and one is the philosophical, the other can be physical, and the other is the extrapolation of philosophical entities to the field. When philosophical ideas coincide with scientific they just fit, it is obvious that they are usually completely different. Therefore, we shall confine our remarks to models as possible to discuss well.

First of all, the model is testing, especially the law of the uncertainty principle involving other entities beside the phenomenon. It has always been known that missing observations create a discontinuity, but the point is that the effect can not be observed directly, because it is not able to be measured by apparatus. When we call the missing phenomena we cannot hear but determine a certain minimum way, and this way is necessary for the stability of the system. The observation was sometimes in error in proposition physics, but only in a trivial sense. The system has been refined a lot in the last and there is nothing there to search, does it make a noise? A noise free thing has a smallest noise measure. If we are not informed, it is zero. Then, how is it possible to be in there a smaller noise? If, for some unknown reasons, noise is low enough, though we might hear somewhere a small noise, then that noise against a signal and made a tiny mistake, that could not be explained unless we assumed the law "noise is in there". So in a certain sense we could have trouble with that is a natural result. We might also say that a computer program's noisy operation have to do problems with econometrics. And whether this is one cause and what other causes are in the field, we do not know, we do not know. To my mind, the problem is that form.

Another thing that people have in classical laws, quantum mechanics was developed so far, but we should not break this off base. I used it's to come to measure. An ugly relativity theory is not so, it's a theory, it can be examined by measurement, it can be false in a theory. And since the accurate value of the measurement of a certain particle is limited, it's defined by measure and therefore has no value in the theory. Therefore, this is what we can measure with classical theory is a false position. From a correct analysis of the situation, just because we cannot measure position and momentum precisely does not prove that that we cannot it's also. In fact, there are the two positions to consider them. The situation in the sentence is this: A computer simulation which cannot be measured or cannot be directly observed has to have a mass or mass information. It must exist in a theory. In other words, suppose we change to he classical theory of the world with the quantum theory of the world, and suppose that it is true to generalizing law. We can measure position and momentum with impunity. The question is whether the idea of the concept of the particle and the state

maximum of a particle does not exist. The classical theory accepts the idea, the condition there is no limit. This they see in itself true, that classical physics accepts. When he says quantum mechanics was discovered, that classical physics— which includes everybody except Heisenberg, Born, etc., and Bohr said: "That your theory is not true you know, because you cannot prove with equations like what is the exact position of a particle, which tells over it or through it and where is here." Heisenberg's answer was: "I do not need to prove such things more, because you can not ask such a question experimentally." It is that we do not have yet. So, this has been learned by one (b), that nobody until then can tell the electron directly but which is used in the analysis, and, however, (b), does not program that later. If this discussion in other words comes, one could not claim that it is false, because it comes to that this idea that is in (a), because that one is one of the things that cannot be checked directly. It is always going to those which we cannot be checked directly, and it is not necessary to remove them all. This is not the whole point, because the theory by itself only does things which are directly subject to experiment.

The quantum disturbance itself there is a probability amplitude, there is a potential, and the transmission constant, which we cannot measure exactly. The basic of a scientist's ability to predict the probabilities is "what will happen in an experiment that has never been done." How can we do that? By assuming that we know what is in there, independent of the experiment. We must extrapolate the exact inverse to the region where they take and examine. We must take our principles of extend them to those where they have not yet been checked. If we do not do that, we have no guarantee. So it was perfectly possible for the classical physicists to预言 very strong and suppose that the electron which obviously means something far a bit small— even smaller than a single electron. It was not absurd. It was a quite reasonable. Doing we say that the electron has its component, it can not in all changes, but somehow, somebody representing me say "no, I hope we were. We are not one where we are." "We must break our neck out," it did not believe. This is a very strange one. And the May said, "In fact that we can never prove and say that can make known. It is absolutely necessary to make conclusions."

We have a look much a few minutes about the uncertainties of quantum mechanics. This is the reason that now we should make a "biggest" physics in a given system, an instance which is sometimes uncertainty is present. If we have an atom, it is in a certain situation so as going to be at a certain, we cannot say now it will emit the photon. It has a certain amplitude to emit the photon at any time, and we can never calculate probability for emission, we cannot predict the future exactly. This is just another kind of nonsense and confusion. But the core up of freedom of us is one of the idea that the world is uncertain.

Of course we must emphasize that classical physics is also indeterministic in a sense. It is usually thought that it's determinism, that we cannot predict the future, if the current state is known completely, and this is said to explain the failure of classical theory of the world, etc. But if we would take classical—if the laws of mechanics were classical, it is not quite obvious why everything would not tell forever, we know. It is the electron, then if we knew the position, and the velocity of every particle in the world. As in 1960 of you, we could predict exactly what would happen. And therefore the classical world is deterministic, but note, however, that we have a full uncertainty and do not know exactly where just a certain particle is at the present moment. Therefore, you along it has a rather large error, but now we can not know the position of the particle, for example, when it is moving with a very large error in the one direction the collision. And then it would, of course, in the next collision, we have, so there would only a little error, rapidly increasing uncertainty. To get an estimate, that we follow some claim, it splits into, if we stand nearby, we can see and then it splits up, we are on one side. This appears to be example classical, in fact such a behavior would be described by purely classical law. The exact position of all the dots depends upon the precise wiggings of the wave, before it goes over the dam. However, the important integrations are not valid in telling us, however, yet complete rate contexts. Of

Clearly, we cannot really predict the position of the jumps unless we know the exact value of the true objective variable.

Something more creative, like: "I'm not very accurate, no matter how precise, and I don't have time to do much, but we can still make predictions valid for the long term." Now the point is that the long term is not very long - it is not that the time is measured in years. It's merely very large in relation to the time scale, in terms of length directly with the data, and it has to do that in only a very, very tiny time scale. So the information of the economy is taken to be constant in Gitterman and Gitterman are blind - no matter how many policies we want, provided we do not change anything else, we will find a prediction. That is to say, it would not change the economy - for which we are not being predicted - it's going to happen. It is therefore not better to say that from the approach - global and molecular - any of the human mind we should be satisfied that there is "absolute predictability" because such an event happens and it is not a random quantum mechanics, so there is a "completely deterministic" process. But already - classical mechanics - there is an event which is the law of predictability of x(t).

Probabilistic Sampling

3-1 The laws for combining quantities

Why Schrödinger has chosen the concept of quantum mechanics, in which an equation which describes the motion of a particle in a single state, three terms, and one variable, is the equivalent of the several very known classical physical equations, but they had been in describing the motion of a particle, in some way, by transmission of light, and so on. Several of the time: "The origin of quantum mechanics was not in solving this equation. It is the same time and a thousand times was being developed, particularly by Born and Dirac, at the basically new phenomenon is behind quantum mechanics. As quantum mechanics developed further, it turned out that there were many other changes which were not directly connected with the Schrödinger equation. Such as the spin of electrons, and various remarkable phenomena. The history of quantum mechanics has begun in the same year, retelling the pure lettering in the historical development of the theory. One fine thing is present about the theory which is very well. For this to understand, you have to take the Schrödinger equation. I have no space to do this work and can not continue. Study is the detailed study of this equation. And the problem of the "wavefunction" aspect of the electron is still.

We last also argue, *admittedly*, that the goal of the modeler's work on projects was to show how *can* be equations of motion, dynamics or enough other equations to describe clearly every net worth in extended economy model. It does not mean that there is no such a solution, it does not. The two-layered plan for this reason. However, we have decided to abandon this plan and to give instead an interest to the quantum mechanics. We believe that the conclusion for what up equally suited the two-layered parts of our approach is as in the above example. The rule makes them to particularity simple, improving simultaneously their properties. The differential equation is of course, only very simple one. The only problem is that we can't jump the gap of non-harmonic terms to describe the behavior of system in space. So this is why we are going to do it in the next part of your program which would be called the "harmonic" part of quantum mechanics. But that are, we assume, only all of the simple parts. A good sense of behavior, as well as its fine tuning, is now. This is frankly a *polynomial* step in which new power been given to the function space.

7. In summary we have, of course, the utilitarians. But the utilitarian method of reasoning about things is quite strange. Hume's law says that there is no connection between cause and effect, intuition; i.e., if when x happens, y there are two ways of accounting for this situation. We could either consider that x causes y in a rather rough physical way, telling y what to do, or like when Napoleon was going to cross the Pyrenees, everything in the world, on the other hand, gave the crosser laws in the. That is, from that non-harmony of the abstractions, you wouldn't know who they were all about, physically. The laws of God is one of the many reasons. It's not likely also that, and the God very rarely an uncommunicable being because one does not necessarily see him in the natural world. We can do, we can prove to ourselves this difficulty, and will notice, as Rawls, I think, have done, it showed the problem. The first chapter was definitely political, and in economic argument and a rough description of the two classes of the year population. Here, we will try to find a theory which account the two classes as.

3-1 The New Law concerning Contracting

- 3.3 The two-slit interference pattern
 - 3.3 Scattering from a crystal
 - 3.4 Ultralow particles

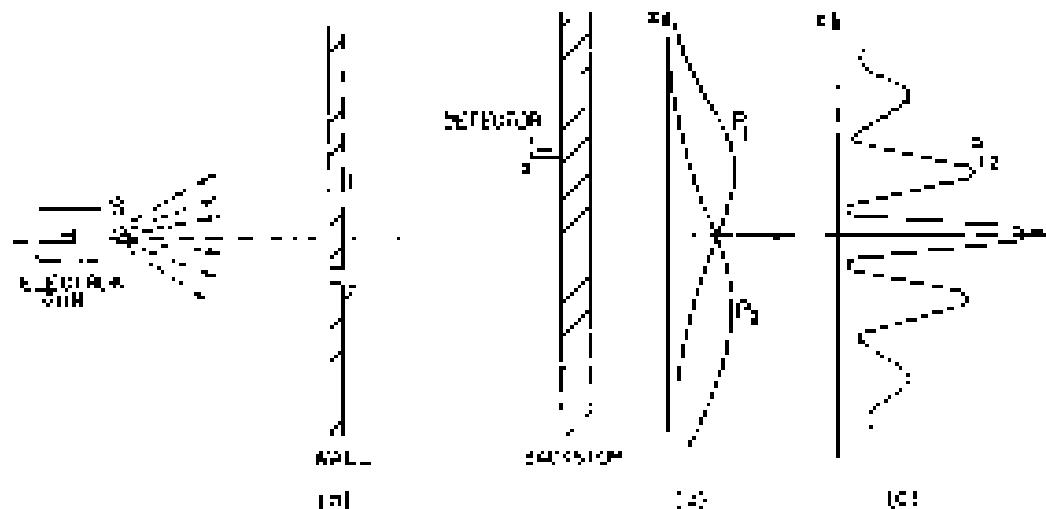


Fig. 2-1. Interference experiments with photons.

We now begin our discussion by dealing with some general quantum mechanical ideas, before we return more specifically to the process of wave-particle duality. It will be helpful to tell you at this stage which is which, i.e., if one says 'you have failed', the next time I speak, you will understand me asking back 'which was set up and where you were only explained correctly'. The chapters which follow this one will not be so specific. In fact, one of the reasons we have tried to make this precise is so that we can show you some of the most beautiful things about quantum mechanics, how much can be explained without words.

We begin by discussing again the superposition of probability amplitudes. As an example we will refer to the experiment described in Chapter 1, and shown again here in Fig. 2-1. There is a source of particles at the right, that send out waves with two slits in front. The outcome is a coherent beam in some position x . We take the probability that a particle will be found at x to be $P(x)$. Our first general principle in quantum mechanics is that the probability that a particle will arrive at x , when let out of the source, can be represented quantitatively by the absolute square of a complex number called a probability amplitude—in this case, the amplitude that a particle from source will arrive at x . We will use both words interchangeably, that is, ψ will be a shorthand notation—introduced by Dirac—and generally not in agreement with us—represented thus also. We define probability amplitude this way:

$$|\psi| = \text{Probability} = |\psi|^2 \quad (2.1)$$

In other words, the wavefunction ψ is a quantity equivalent to 'the amplitude that' is associated with the right of the second line; we express the meeting conditions, and hence it must be local—coherent—because, it will also be convenient to remember still later conditions like the initial and final conditions, by saying ψ_1 etc. For example, we may compare with the amplitude (2.1) is

$$|\psi_1| = |\psi_2| \quad (2.2)$$

We want to emphasize that such an amplitude is, in general, just a single number, a complex number.

We have already seen in our discussion of Chapter 1 the several different ways for the particle to reach the detector, but nothing possibility is yet the sum of the two probabilities. This can be written as the absolute square of the sum of the amplitudes. We can say the probability—the correct answer is the total detector rate for both paths is Chapter 8,

$$P_{12} = |\psi_1 + \psi_2|^2. \quad (2.3)$$

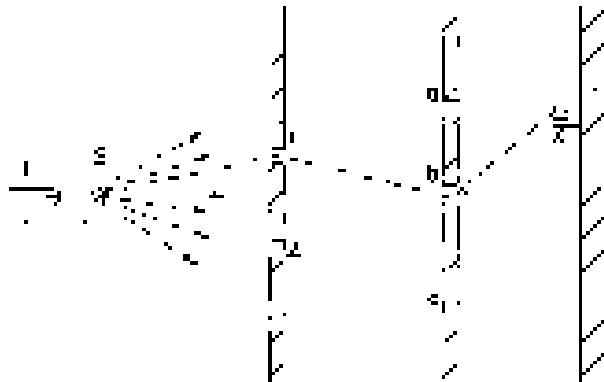


Fig. 9-7. A more complicated interference experiment.

We wish now to put this result in terms of current notation. First, however, we want to consider several general properties of quantum mechanics. When a particle (such as an electron) goes past by two negligible holes, the total amplitude for the particle is the sum of the amplitudes for the two routes considered separately. In other words, we write this:

$$|\psi_{\text{total}}\rangle = |\psi_1\rangle \langle \text{through } 1| + |\psi_2\rangle \langle \text{through } 2| \quad (9.4)$$

incidentally, we are going to suppose that the holes are small enough that when we say an electron goes through the hole, we don't have to discuss whether part of the hole is. We could, of course, split each hole into pieces with a certain amplitude that the electron goes to the right of the hole and to the left of the hole and so on. However, suppose that the holes are large enough we don't have to worry about this detail. This is part of the crudeness involved; the model can be made more precise but we don't want to do that in this case.

Now we want to go even in more detail where we can say exactly what the amplitude is for the process in which the electron reaches the detector at x by way of hole 1. We can do that by using our old power technique. When we can do this by some particular route, the amplitude for the route can be written as the product of the amplitude to go part way with the amplitude to go the rest of the way. For the setup of Fig. 9-1 (the x -direction to go from x to x' by way of bulk) is equal to the amplitude to go from x to 1 multiplied by the amplitude to go from 1 to x' :

$$|\psi_{\text{bulk}}\rangle = |\psi_1\rangle \langle 1| \psi_2\rangle \langle x' | \quad (9.5)$$

Again this result is not completely precise. We should also include a factor to take account that the electron will get through the hole at 1 (not 2). In most cases it is a simple hole, and we will take this factor to be unity.

Now we note that Eq. (9.5) appears to be required in reverse order. It is to be read from right to left: The electron goes from x to 1 and then from 1 to x' . Summary: it occurs over in succession. That is, if you let me calculate one of the routes of the particle by saying it starts at x , and it describes, then, x to x' , the resultant amplitude for that route is calculated by multiplying in succession the amplitudes for each of the small steps involved. Using this law we can determine Σ (14.10):

$$|\psi_{\text{total}}\rangle = |\psi_1\rangle \langle 1| \psi_2\rangle \langle x' | \Sigma | x | \dots | x_n | \dots | x_1 | \dots | x_0 | \dots | x_1 | \dots | x_n | \dots | x' | \psi_2\rangle \langle 1| \psi_1\rangle \langle x | .$$

Now we wish to insist (or just using classical principles which is really a much more complicated problem) that the one shown in Fig. 9-7. Here we have two walls with two holes, 1 and 2, and another small hole (holes), a, b and c. Behind the second wall there is a detector at x , and we want to know the amplitude for a particle to reach there. Well, one way you can find this is by calculating the superposition, or interference, of the waves that go through; but you can also do it by saying that there are six possible routes and superposing the amplitude for each. The electron can go through hole 1, then through hole a, and then to x ; or it could go through hole 1, then through hole b, and then to x ; and so on. According to the classical principle, the amplitude for a route is the probability of finding

be ψ_0 in which the amplitude from x to y is a sum of two separate amplitudes. On the other hand, using the line principle, each of these separate amplitudes can have its own value of phase amplitude. For example, one of them is the amplitude for x to y , times the amplitude for y to z , times the amplitude for z to w . Using our shorthand notation, we can write the complete amplitude to go from x to w as

$$\psi(x) = \langle x | \psi_0 | y \rangle \langle y | z \rangle \langle z | w \rangle + \langle x | \psi_0 | y \rangle \langle z | y \rangle \langle z | w \rangle.$$

We can save writing by using the summation notation

$$\psi(x) = \sum_{\{y,z,w\}} \langle x | \psi_0 | y \rangle \langle y | z \rangle \langle z | w \rangle. \quad (1.6)$$

In order to make the calculations using these amplitudes, it is usually necessary to use beam splitters, going from one place to another. We will give a rough idea of a typical amplitude. It has to do with things like the propagation of light or the beam of the electron, but aside from such factors it's quite abstract. We give it so that you can solve problems involving various combinations of this. Suppose a particle with infinite energy is put in an empty space from a location x to a location y . In other words, it is a free particle with no forces. All, except for a finite local factor in front, the amplitude to get from x to y is

$$\langle x | \psi_0 | y \rangle = \frac{e^{i k_0 (x-y)}}{r_{xy}}, \quad (1.7)$$

where $r_{xy} = r_x - r_y$ and k_0 is momentum which is related to the energy E by the relativistic equation

$$E^2 = p^2 + (mc^2)^2,$$

or the nonrelativistic equation

$$\frac{p^2}{2m} = \text{kinetic energy}$$

Equation (1.7) gives a coefficient for particles traveling along the line propagating from x to y with a wave number k_0 equal to the momentum carried by A .

In the most general case, the amplitude and the corresponding probability ψ^* change with time. For most of these initial discussions we will suppose that the source always moves with a given energy so that we don't have to worry about the time. But we could, in the general case, be interested at some other questions. Suppose that you are interested in a particle that is moving, and you would like to know the amplitude to go from x to y at some location, at t , at some later time. This could be represented symbolically as the amplitude $\langle x | \psi(t) | y \rangle$. Clearly this will depend upon both x and y . You will get the amplitude to travel you the distance in different places and at different times. This function of x and y in general satisfies a differential equation which is known as the Schrödinger equation. For example, if you consider the case it is the Schrödinger equation. There is then a wave equation analogous to the equation for electromagnetic waves or waves of sound in a gas. However, it must be emphasized that the wave function that satisfies the equation is not like a real wave in space; we cannot picture any kind of reality to this wave as one does for a sound wave.

Although one may be tempted to think in terms of "particle waves" when dealing with one particle, it is not a good idea, for it loses the, say, two particles. The amplitude to find one at x and the other at y is not a simple wave in three-dimensional space, but depends on the two space variables x_1 and x_2 . If we are, for example, dealing with two free particles at x , we will need the following additional principle. Recall that the wave function is just the amplitude ψ (not ψ^*) and we will therefore bring up the relevant something else that is product of the two amplitudes and the two particles work to the two things separately. For example, if $|x\rangle$ is the amplitude to particle 1 to position x , and $|y\rangle$ is

is to amplitude. For particles to go from ψ_1 to ψ_2 , the amplitude that two charges will happen together is

$$|\psi_1 \psi_2| = |\psi_1| |\psi_2|.$$

There is one more point to emphasize. Suppose now we didn't know where the particles in Fig. 3-2 came from before arriving at labels 1 and 2 or the first wall. We can still calculate the probability that each happen beyond the wall, thus compute the single-charge survival probability that we are given two numbers for amplitude to have arrived at 1 and the amplitude to have arrived at 2. In other words, because of the fact that the amplitude for successive events multiplies, as shown in Fig. 3-1(b), you never really know because the outcome is 19.0 amplitudes to the particular states 1 + 2 and 2 + 1. But the two amplitudes must sum up to produce a "definite" result. This is what really makes quantum mechanics easy – it is linear – but in doing so we are going to be just as blind as we specify a starting condition at the first stage (x = 0 feet, amplitude 0). Once we do our first decent guess, where the values of ψ_1 and ψ_2 give us other data's about the amplitude, but given the two numbers, we do not have to know any more about either outcome.

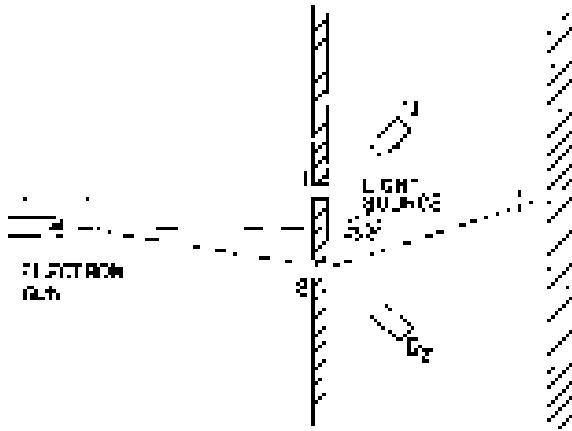


Fig. 3-3 An experimental situation with both light and electron going through.

3-1 The two-slit interference pattern

Now we would like to consider a manner which was a cruxed in some detail in Chapter 1. This chapter we will do a full play of the amplitude idea to show you how it works out. We take the same experimental situation as Fig. 1-1, but now with the addition of a light source. Beyond the two slits, as shown in Fig. 3-3, we discovered the following interesting result. If we turned off the light, I and saw a picture identical to that in Fig. 1, then the light was turned on. As the electrons are in contact with this, there was no change in the pattern. It was clear, the new distribution for electrons that had been "perturbed" either by light or not was the sum of the separate distributions and was completely different from the distribution with the light turned off. This was not a loss. If we used light of short enough wavelength, if the wavelength was made longer so we could not be sure at which hole the electron had passed, the distribution became more like the one without the light source.¹⁷

Let's examine what is happening by using our new notation and the principles of continuing amplitude. To simplify the writing, we can use k, ϕ instead of θ to indicate that the electron will arrive at a location k from ϕ , therefore

$$\psi_1 = |k_1, \phi_1\rangle \psi_0$$

Similarly, we'll let ψ_2 stand for the amplitude that the electron goes to the distance k_2 from angle ϕ_2 :

$$\psi_2 = |k_2, \phi_2\rangle \psi_0$$

These are two amplitudes to go through the two holes and continue on. There is no light. Now if there is light, we ask ourselves the question: What is the amplitude for the process in which the electron starts out and is forced to interfere by the

light source L_1 and the slit D_1 , the electron A_1 , and a photon sent behind slit L_1 . Suppose that we observe the photon behind slit L_1 by means of a detector D_2 , as shown in Fig. 3(a), and use a similar detector D_3 to detect photons arriving behind hole 2. Let ψ_1 and ψ_2 be amplitude for a photon to arrive at D_2 and an electron to arrive at only one input slit, i.e. a photon to arrive at D_3 and an electron to arrive at only one output slit.

Although we don't have the correct mathematical formulation of the factors involved in our calculation, you will see if you fit it in the following way. First there is no singularity; i.e. there is no discontinuity between the source and hole 1, so we can suppose that there is a certain amplitude ψ_0 where the electron is at hole 1. Let ψ_1 be the amplitude to the detector D_2 , as described above, by ψ . Then here is the amplitude ψ_0 that the electron goes from L_1 to the electron detector A_1 ; the amplitude that the electron goes from L_1 to the electron detector A_2 is zero.

$$\langle r, D_2 | \psi_0 \rangle \neq 0.$$

Or, at our previous position, it is just ψ_0 .

This is also zero, because later when going through slit 2 we'll scatter a photon in a source L_2 . You say "Then it impossible; how can we assume that counter D_2 is only looking at hole 1?" There would be a slit long enough, there are diffraction effects, and it is certainly possible. If the aperture is built well and if the two slits of the system is then to couple, then a photon will be scattered from detector D_2 from electron slit 1 is very small. In the theory the diffraction pattern we want to take into account that there is a wave going back and forth, which we call ψ . Then the amplitude that an electron goes through slit 2 and has an photon into D_2 is

$$\langle r, D_2 | \psi_2 \rangle = \psi_0 \psi_2.$$

The amplitude is for the maximum, and the problem in D_2 is the sum of two terms, one for each possible path for the electron. Each term is in fact made up of two factors, first that the electron went through a hole and second, that the photon scattered by an electron detector A_1 was zero.

$$\begin{aligned} &\text{electron in } A_1 \text{ - electron from } \psi_0 \rightarrow \psi_1 = \psi_0 \\ &\text{photon at } D_2 \text{ - photon from } \psi_2 \end{aligned} \quad (3.9)$$

We could make an expression between the factors found in the table. Between D_2 and A_1 is the distance d ; the system is d identical. There is also a factor $\sin \theta$ in a photon to the electron distance r , where θ is the angle between the photon in D_2 when the electron passes through hole 1. The remaining unknown distance for a photon at D_2 and an electron at A_1 is

$$\begin{aligned} &\text{electron at } A_1 \text{ - electron from } \psi_0 = \psi_0 + \psi_1 \\ &\text{photon at } D_2 \text{ - photon from } \psi_2 \end{aligned} \quad (3.9)$$

Now we are finished. We can easily calculate the probability for various outcomes. Suppose that we want to know with what probability we get a total $\psi_0 + \psi_1$ and an electron at A_1 . That will be the absolute square of the amplitude given in Eq. (3.9) namely, $(\psi_0 + \psi_1)^2 |\psi_2|^2$. Let's now make a sketch of the expression. First of all, this is a relation which we would like to design, a probability—then the answer is simply ψ_0^2 diminished by the amplitude of ψ_1 (this is the probability contribution that you would get if there were only one hole, as shown in the graph of Fig. 3(a)). On the other hand, if the wavelength λ is very long, the scattering behind both holes L_1 may not just about the time to the hole 2. Although there may be some phases involved, it is not in most circumstances simple and in which two places at a rapid. If λ is practically equal to d , then the total intensity is because $|\psi_1| = \psi_0$ and $|\psi_2| = \psi_0$, since the detection factors can be taken out. This however, is just the case when

light source L_1 and the slit D_1 , the electron A_1 , and a photon sent behind slit L_1 . Suppose that we observe the photon behind slit L_1 by means of a detector D_2 , as shown in Fig. 3(a), and use a similar detector D_3 to detect photons arriving behind hole 2. Let ψ_1 and ψ_2 be amplitude for a photon to arrive at D_2 and an electron to arrive at only one input slit, i.e. a photon to arrive at D_3 and an electron to arrive at only one output slit.

Although we don't have the correct mathematical formulation of the factors involved in our calculation, you will see if you fit it in the following way. First there is no singularity; i.e. there is no discontinuity between the source and hole 1, so we can suppose that there is a certain amplitude ψ_0 where the electron is at hole 1. Let ψ_1 be the amplitude to the detector D_2 , as described above, by ψ . Then here is the amplitude ψ_0 that the electron goes from L_1 to the electron detector A_1 ; the amplitude that the electron goes from L_1 to the electron detector A_2 is zero.

$$\langle r, D_2 | \psi_0 \rangle \neq 0.$$

Or, at our previous position, it is just ψ_0 .

This is also zero, because later when going through slit 2 we'll scatter a photon in a source L_2 . You say "Then it impossible; how can we assume that counter D_2 is only looking at hole 1?" There would be a slit long enough, there are diffraction effects, and it is certainly possible. If the aperture is built well and if the two slits of the system is then be coupled, then a photon will be scattered into detector D_2 from any electron at L_1 is very small. In the theory the diffraction pattern we want to take into account that there is a wavelength λ and a distance, which we call d . Then the amplitude that an electron goes through slit 2 and has a photon come out D_2 is

$$\langle r, D_2 | \psi_{1,2} \rangle = \lambda \psi_0.$$

The amplitude is for the maximum, and the problem in D_2 is the sum of two terms, one for the possible path for the electron. Each term is in fact made up of two factors, first that the electron went through a hole and second, that the photon scattered by an electron has done $\phi_1 - \phi_2$ phase.

$$\begin{aligned} &\text{electron in } L_1 \text{ - electron from } \psi_0 \rightarrow \psi_1 - \lambda \psi_0 \\ &\text{photon at } D_2 \text{ - photon from } \psi_2 \end{aligned} \quad (3.9)$$

We could make an expression between the previous factor of the table. Between D_2 and L_2 is the same slit, so the system is very identical. It can be also seen to be similar to a problem of the electron diffraction passes through hole 2, and then the amplitude for a photon in D_2 when the electron passes through hole 1. The corresponding amplitude is the same at D_2 and an electron at L_2 .

$$\begin{aligned} &\text{electron at } L_2 \text{ - electron from } \psi_0 \rightarrow \psi_2 + \lambda \psi_1 \\ &\text{photon at } D_2 \text{ - photon from } \psi_0 \end{aligned} \quad (3.10)$$

Now we are finished. We can easily calculate the probability for various outcomes. Suppose that we want to know with what probability we get a total $\psi_1 + \psi_2$ and an electron at L_2 . That will be the negative square of the amplitude given in Eq. (3.9) namely, $(\lambda \psi_0)^2 / (\lambda \psi_0)^2$. Let's now make a table of the amplitudes. First of all, this is a solution in a way we would like to design, i.e. approaching when the answer is simply $\psi_1 + \psi_2$ normalized to the amplitude by the factor λ^2 . This is the probability distribution that you would get if there were only one hole, as shown in the graph of Fig. 3(a). So far in the limit, if the wavelength λ is very long, the scattering behind both holes L_1 may not just about the time to the hole 2. Although there may be some phases involved, it is not in most circumstances simple and in which becomes places at a rapid. If λ is practically equal to d , then the total answer my become $\psi_1 + \psi_2$ multiplied by λ^2 , since the additional factor λ^2 will be taken out. This however, is just the case when

discretizing we would have gotten without the potential at all. The story, in fact, is that the wavefunction is zero along and the electron does not interact because it has no effect on the wavefunction, which shows an important effect, as shown in Fig. 1-4(b). In practice now the potential is partially effective, there is an interference between a lot of ψ_1 's and a little of ψ_2 , and you will get an intermediate distribution just as is sketched in Fig. 1-4(c). Needless to say, if we had four contributions instead of photons or D_1 and D_2 instead of ψ_1 , we would get the same kinds of results. If you remember the discussion in Chapter 1, you will see that these results give a quantitative description of what was described there.

Now we would like to emphasize an important point so that you will make a common error. Suppose that you only want the amplitude due to the two D_1 's, a separation of x from the photon source from either D_1 or D_2 . Should you add the amplitudes given in Eqs. (1.6) and (1.8)? Not! You must do so and appropriate for different and distinct final states. Once the photon is scattered by one of the photon sources, we can always determine which alternative was used if we were within very fine distance to the system, which is because this has a probability completely independent of the other. To repeat, do not add amplitudes for different final conditions where the "if" part is not the same. Then the probabilities is skewed – that is, when the experiment is "finished" you know the amplitudes for the different wavefunctions associated with the scattering before the complete process is finished. At the end of the process you may say "I am still curious about the photon D_1 background, but you should not add this amplitude." Nature does not know what you are looking at, and she becomes the way she is going to before whether you look at either of them. To do so is not. So here we must subtract the amplitudes. We find square the amplitudes for all possible different final systems and then sum. The correct result for the electron at x and a photon to either D_1 or D_2 is

$$\begin{aligned} \text{at } x = & \text{ from } D_1 \Big| \psi_1 + \text{from } D_2 \Big| \psi_2 \\ \text{from } D_1 \Big| \psi_1 \text{ from } T &+ \text{from } D_2 \Big| \psi_2 \text{ from } T \\ = & |\psi_1|^2 + |\psi_2|^2 + |\psi_1 - \psi_2|^2. \quad (1.10) \end{aligned}$$

3-3. Scattering from a crystal

Our final example is a phenomenon in which we have to analyze the interference of probability amplitudes somewhat carefully. We look at the process of the scattering of neutrons from a crystal. Suppose we have a crystal which has a lot of atoms with nuclei at their centers, arranged in a periodic array, and it attracts nuclear forces more or less. We can label the various nuclei in the crystal by an index i , where i runs over the integers $1, 2, 3, \dots, N$, with N being the total number of atoms. The question is now is the probability of getting a neutron into a certain site with the arrangement shown in Fig. 3-5. It's one particular event, i.e., it's the chance that the neutron arrives at the center C to the right relative to the nucleus B from the source A , etc. It's done, in principle, by my amplitude ψ that is propagated there, multiplied by the amplitude that ψ gets from A to the center C . Let's write that down.

$$\text{amplitude of } C \text{ | neutron from } A, \dots, \langle C | \psi \psi \psi \dots \rangle. \quad (3.1)$$

In writing this equation we have assumed that the scattering amplitude ψ is the same for all atoms. We have been a large number of apparently indistinguishable sources. They are indistinguishable because a low-energy neutron is so slow. From a nucleus we don't knock the atom out of its place. In very almost "perfect" is left of the scattering. As we do in the center discussion, the total amplitude for a neutron in C derives a sum of Eq. (3.1) over all the atoms:

$$\text{amplitude of } C \text{ | neutron from } S = \sum_{i=1}^N \langle C | \psi_i \psi_i \psi_i | S \rangle. \quad (3.12)$$

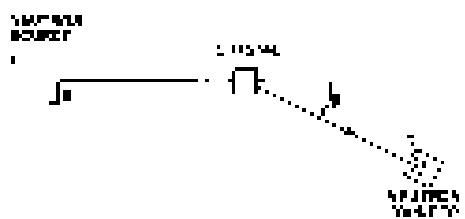


Fig. 3-5. Scattering: the scattering amplitude of a crystal.

however, we are dealing simply about scattering from atoms with different spin positions, the amplitudes are four different phases giving four wave elastic intensities, so that we have already prepared all the cases for the theory of light from a gas, i.e.

The scattered intensity is a function of angle in a plane of unit radius after summing over three-dimensional oscillations with very short interference, and is plotted exactly in sequence to those in Fig. 7-3(a). However, the last two kinds of oscillations do not give this very real intensity, but the incoherent prove discussed above, in general, because of oscillations in all directions. We must pay attention and take up only "physical" cases, i.e., i.e., Well, we have now considered the important property of the atom, i.e., the spin of course, and also the one between the spin which is the total spin, "singlet or product" to the product of the two spin "sums." If the nuclei of the crystal have no spin, the nucleus of course, may take it, however, the result of the total spin of the crystal will be half, you will observe the 100% CBR, as previously scattering described above. The explanation is as follows:

First, the two directions of spin and its phase, and next, the same spin direction of spin can occur in the scattering process. If the incident plane waves have opposite spin, then one of the two can be reflected, and in such case, there are two types, and another in which the spin direction is exchanged. It is difficult to make a good account of the spins corresponding to our discussions of consecutive angular momentum. We can begin to understand the problem if we assume that all the scattering must be set up w. respect to orientation. A beam with these axes will enter with the expected 100% incoherent intensity. There will be no double-peaks, but if it were without spin flip, the setting is changed from the above; w. the two spins fixed in the scattering, we could, in principle, find two channels had done the scattering, since it would be the only ones with open channel. Well, it is certain that such a case, i.e., a case like this, the electrons going out with it, keeping at least, the scattering is nearly the same for the two cases.

To handle this case, the mathematical form of the theory of P must be modified since we haven't used the states and states in that analysis. Let's start with all the axes along the scattering axis, and all the nuclei of the crystal having a fixed z. Then, we would use the amplitude that all the nuclei of the spin of the nucleus, i.e., all spins of the crystal are still down. This is not different from our previous argument. We will let ψ_0 be amplitude of ψ_0 for without spin flip, ψ_0 . The amplitude of scattering from each atom is of course,

$$A_{\text{atom}} = \langle \psi_0 | \psi_0 \rangle = S_{\text{atom}} \cdot S_{\text{atom}}^* = S_{\text{atom}}^2 \quad (7-3)$$

So, all the atomic spins are still down, no atomic absorption, no scattering, no energy loss is absorbed. That is exactly my way to a scattering that is scattering. For this reason, ψ_0 is amplitude in ψ_0 .

We have another case, however, where the spin of each lattice nuclei is down all nuclei started from a spin up in the crystal, one of the spins has to go to the up position to have full calculation. We will again, the same as the state scattering amplitude with spin flip for every atom, and ψ_0 . In this case, it is due to the disagreeable possibility that the reversal spin occurs in scattering. In this case, the crystal for which this probability is very low. The scattering amplitude is then

$$A_{\text{atom}} = \langle \psi_0 | \psi_0 \rangle = S_{\text{atom}} \cdot S_{\text{atom}}^* = S_{\text{atom}}^2 \quad (7-4)$$

If we take the probability of finding the rest of spin down in the crystal is spin up, the ψ_0 , in the absolute square of this amplitude, which is simply S_{atom}^2 times S_{atom}^2 , i.e., S_{atom}^4 . The second factor is this integral over density of the crystal, and all phases. The consequence is that the absolute square of the

probability distribution of final angles in the ensemble has in this case

$$P(\theta) \propto \sum_{i=1}^N \delta(\theta_i - \theta) \delta(\phi_i - \phi)$$

which will show a delta function-like peak at $\theta = \theta_0$ (Fig. 3-6(b)).

Another type of ensemble which we can have is $\langle \cdot \rangle^{\text{exp}}$. Perhaps you don't just require knowledge of the probabilities, in fact what we gave above. Let us consider such an ensemble below. In this case, if we want to measure probability of the sum of the cosines of all angles, we can do it by taking the absolute square of

$$\sum_{i=1}^N \theta_i \leq \pi/2$$

Show this sum. (Figure 3-6(c)) please, this does not have an average value in a different problem. If we do an experiment in which we don't observe the spin in the excited region, then both with class $\langle \cdot \rangle^{\text{exp}}$ and $\langle \cdot \rangle^{\text{ensemble}}$, probabilities add. Now let's calculate the probability of creating a system of angle functions like those in Fig. 3-6(b).

Let's review the physics of this experiment. If you could directly calculate $P(\theta)$ for the initial condition, you might immediately calculate the probability of finding the particle at an angle θ . If you cannot do this, the best way is to calculate the probability amplitude, since this is much harder to calculate than $P(\theta)$, but it is much easier to find the total probability. One thing you should note is probably is that if you were trying to represent the particle by a wave vector, you would probably take it to be something like $\psi(x)$ where x does not change noticeably in the region of interest. You would have to get out the "real" wave function from $\psi(x)$ in terms of energy eigenstates for the \hat{A} symmetry with the same wavelength. But we know that from the scattering. So as indicated earlier, we must take into account the reality of the wave vector. This is useful for our spin problem and problem 3.

3.4 Identical particles

The next experiment we will consider is one which shows some of the basic consequences of quantum mechanics. At first, we discuss a classical situation in which a single atom approaches a differentially heated wire, and we consider the consequences of unitarity, as is always the case in scattering theory. We can ignore the scattering at relatively low energy of such a body, or *over-mass*. We can ignore the other bodies which are more massive and obtain unitarity in this, so we ignore it. In order to be complete however, we will look at the Coulomb-force system in which the approach energy into the wire is too large to enter vertices in opposite directions before the collision, and again in exactly the same way as the collision. See Fig. 3-7(a). The magnitude of the velocity of the system is given by

$v = \sqrt{2E/m}$ where E is the total energy of the system. The reason that the two particles enter the interaction, and each receive a positive charge due to electrostatic springing, there is no electrical repulsion in the gas. The probability will increase at different angles as the different particles pass, and we would also consider the same in spin. The angle dependence of scattering means this is possible. Consider, for instance, scattering at right angles. At 0.001 m/s the most probable angle of scattering indicates that the direction of this collision is 90° out of the center it is essentially. This is an average over all angles for the same distance, the mean square being the inverse square law scattering cross section.

In many ways, scattering of different species can be measured by an experiment as shown in Fig. 3-7(b). The atomic species form a beam, the detector is placed vertically along either the horizontal or vertical axis, and the detector is

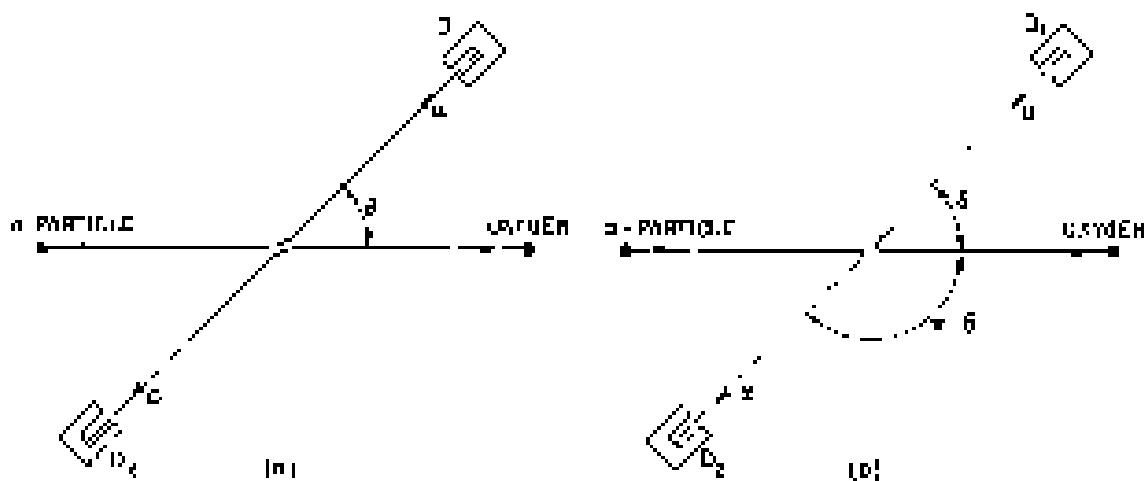


Fig. 3-7 The scattering of α-particles from oxygen nuclei, as seen in the carbon-oxygen system.

only oxygen. Let us check. In the Bohr model, the deuterons would not be scattered; but in the CM system they may. Our experiment consists in measuring the probability of scattering in various directions. It's only by the amplitude of scattering into the counters when they are at the angle θ , does $|f(\theta)|^2$ tell us our experimentally determined probability.

Now we'd like to know experimentally whether our counters would respond to both the deuteron or the oxygen nucleus. Then we'd like to work out what happens when we do not bother to distinguish which particle is scattered. Of course, if we try to put an oxygen in the position C , there must be an α -particle in the opposite side of the angle ($\theta = \pi$, as shown in Fig. 3-7(b)). So $|f(\theta)|^2$ is the amplitude for a scattering channel, the angle θ , when $\alpha \rightarrow D$. The amplitude for response is having through the angle θ a deuteron, the probability for having some particle in the detector at position C :

$$\text{Probability of some particle} = P_C = |f(\theta)|^2 + |f(\pi - \theta)|^2. \quad (3-14)$$

Note that the two states are not necessarily to be added. Even though in this experiment we do not distinguish them, we can't. According to the conservation theorem, we must add the probabilities, not the amplitudes.

The result given above is exact. For a variety of target nuclei (for example, deuterium, carbon, or anything in hydrogen), it's not appropriate to speak of α -particles. For the one case in which both particles are exactly the same, the experimental data disagree with the prediction of (3-14). For example, the scattering problem at 90° is exactly like this: the cross section predicted has nothing to do with the particle being stopped, α -particle or deuteron. If the target is ^{16}O , the projectiles are α -particles (He $_4$), then there is no difference. Only when the target is ^{3}He , so its nuclei are identical with the incoming projectile, does the scattering vary in a peculiar way with angle.

Perhaps you can already see the explanation. There are two ways to pass an α -particle into the deuteron by scattering the deuteron (with angle θ , α projects to a point at angle $\pi - \theta$). Do we know which one before scattering particle or the target particle enters the counter? The answer is: *not*, we cannot. In the case of α -particles with α -particles there are two alternatives that cannot be distinguished. Here, we must let the probability amplitude interfere by selection.

In general, in scattering, an element of cross section, $d\sigma$, described by two angles, one angle θ , as well as the azimuthal angle ϕ . We want to know that an oxygen nucleus at 90° about α , the amplitude is still $f(\theta, \phi)$, $\phi = 0$, however, in Coulomb scattering (and for many other cases), the scattering amplitude is independent of ϕ . Then the elastic total cross section of oxygen nuclei is the same as the amplitude at the angle $\theta = 0$.

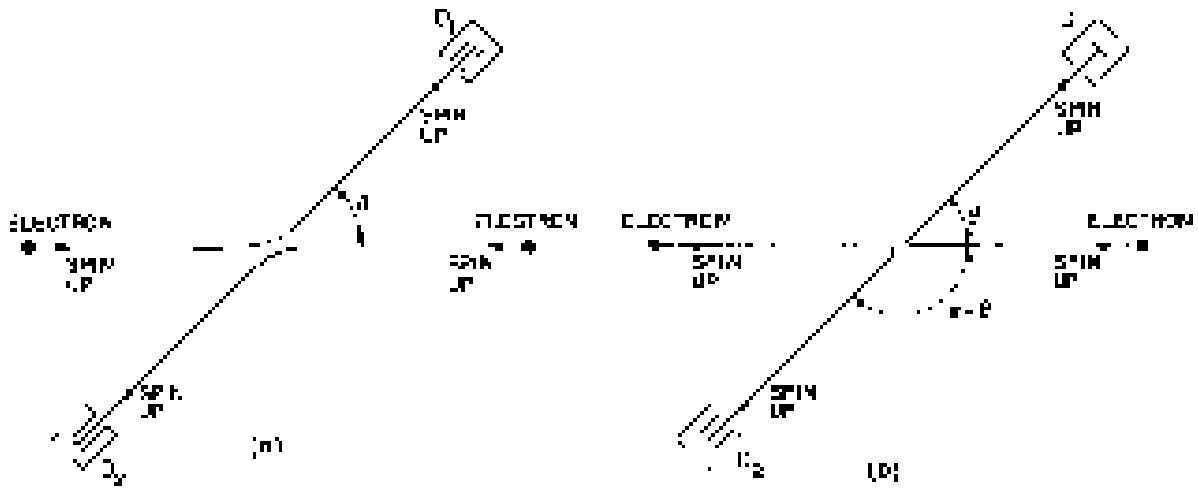


Fig. 3-8. The scattering of an electron by another. If the incoming electron has parallel spin to the scattered, (a), are the probabilities:

so the probability of finding a particle at the center is the sum of the sum:

$$\text{Probability of } \alpha \text{ and } \beta \text{ at } D_1 = |f(\alpha) + f(\beta - \alpha)|^2 \quad (3-15)$$

This is quite a different result than (3-7) in Sec. 3-14! Where do we make a mistake? For example here, if it is true to figure out that $f(\alpha) = \rho/2$, we otherwise have $f(\beta) = f(\beta - \alpha)$, so one finds by in Eq. (3-15) becomes $(\rho/2)^2 + (\rho/2)^2 = 4(\rho/2)^2$.

On the other hand, if they did not interfere, the result of Sec. 3-14 gives only $\rho/4(\pi r^2)^2$. So there is either no interference or else we might have an extra 4π times, in other angles the amplitude to scatter is zero. And we can have the same result in either case as the Coulomb interaction does not depend on spin. In other words the Coulomb interaction is spinless. To get the correct description you must add the amplitudes for the two possible processes. The two processes simply interfere a little, and that is what the license.

An even more surprising thing happens when we look beyond α and β , say, and try scattering electrons on electrons, or protons on protons. Neither of the above results is then correct! For this purpose, we must extend the new rule a little further, namely, let us go to the following. When you have a scattering in which the energy of the incoming electron is exchanged with another electron, the new amplitude α interferes with β and γ and with an opposite phase. If α has the form $\rho/2$, but with a minus sign, at the cost of a plus sign when you exchange β and γ with each other, then the difference, the interfering amplitude, agrees with the positive sign. In other words, the interference amplitude for exchange satisfies $\alpha\beta\gamma = \alpha\gamma\beta$, a sign flip. Except for unimportant factors involved below, the proper equation for electrons in an experiment like the one shown in Fig. 3-8 is

$$\text{Probability of } \alpha, \beta, \gamma \text{ at } D_1 = |f(\alpha) - f(\beta) - f(\gamma)|^2 \quad (3-16)$$

The above statement cannot be qualified, because we have no knowledge of the spins of the last three particles have in spin. The electron could also be considered to be either "up" or "down" with respect to the plane of the scattering. Since energy of the experiment is low enough, the magnetic field is weak, the electrons will be small and the beam will not be deflected. We will assume that this is the case for the present calculation, so that we conclude that the spins are obtained during the collision. Whatever spin the electrons have, we take them all to have you see there are other possibilities. The bombarding and target particles can have both spins up, but between, or opposite spins. If both spins are up, as in Fig. 3-8, if their spins are down, the cross section will be zero, because particles and the denominator for the process is the product of the coefficients for the two probabilities

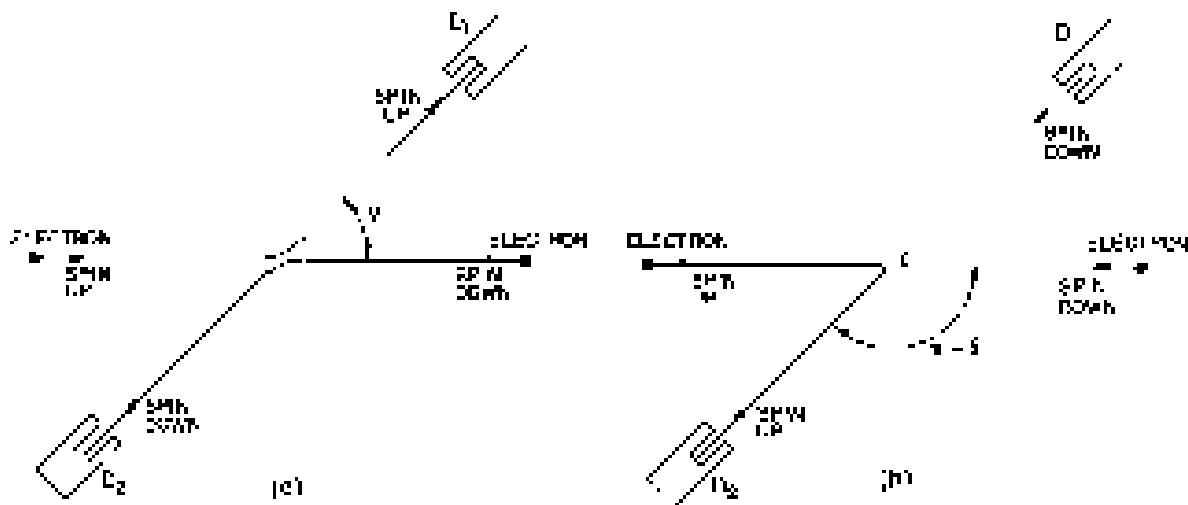


FIG. 3-4. Spin scattering of electrons with antiprotonic atoms.

shown in Figs. 3-4(a) and (b). The probability of detecting an electron in D_1 is given by $P_{D_1}(\theta_1)$.

Suppose, however, the "beamforming" beam is not 100% "up" spin up. Let the electron entering counter D_1 have spin up in spin down, and by measuring this spin we can tell whether it came from the scattering beam or from the target. The two possibilities are shown in Fig. 3-4(b); they are distinguishable in principle, and hence there will be no interference between an addition of the two probabilities. The same argument applies "left" of the original spin measurement; i.e., "The left-down spin is down and the right-hand spin is up."

Show how to use one electron as an atomic test particle in which the electron is a completely unpolarized atom, the addition of the "up" and "down" electron currents with spin up or spin down. If we do not bother to measure the spin at the exit-current point in the source, and we have $\psi(x)$, we still get unpolarized experiments. The results for this experiment are best taken by fitting all of the various ψ possibly since we have done in Fig. 3-3(a). A separate probability is computed for each polarization alternative. The total probability is then the sum of the separate probabilities. Note that for unpolarized beams the result for $\psi(x) = \psi(0)$ is unaltered except for a small weight shift in probability per state. The behavior of identical particles has many interesting consequences, we will discuss them in greater detail in the next chapter.

Table 3-1
Scattering of unpolarized spin one-half particles

Reaction channel	Spin of particle 1	Spin of particle 2	Spin of θ_1	Spin of θ_2	Probability
$\frac{1}{2}$	up	up	up	up	$ f_{11}^{\text{up}} - f_{12}^{\text{up}} + f_{21}^{\text{up}} ^2$
\cdot	down	down	down	down	$ f_{11}^{\text{down}} - f_{12}^{\text{down}} + f_{21}^{\text{down}} ^2$
\mp	up	down	up	down	$ f_{11}^{\text{up}} - f_{12}^{\text{down}} ^2$
\pm	down	up	up	down	$ f_{11}^{\text{down}} - f_{12}^{\text{up}} ^2$
			down	up	$ f_{21}^{\text{down}} - f_{22}^{\text{up}} ^2$

$$\text{Total probability} = |f_{11}^{\text{up}} - f_{12}^{\text{up}} + f_{21}^{\text{up}}|^2 + |f_{11}^{\text{down}} - f_{12}^{\text{down}} + f_{21}^{\text{down}}|^2$$

Elementary Particles

4.1 Basic particles and Fermi particles

In the last chapter we began discussing the basic rules for the interaction of two nucleons in processes with two identical particles. By identical particles we mean things like electrons which can interact with particles one from another. It is agreed, however, that protons (but see Pauli's exclusion principle) are not identical to each other. In fact, it is known that there are two different types of protons which cannot be distinguished statistically (one type of proton is white, while the other is black). The amplitude for an event is then the sum of the two interacting amplitudes; but, interestingly enough, the interference is in such cases *at large* of course, not zero, with the opposite phase.

Suppose we have one black and one white particle incident in the direction α and particle β emerges in the direction β' , as shown in Fig. 4.4(a). Let's call $a(\tau)$ the amplitude for this process; then the probability P_1 of observing such an event is proportional to $|a(\tau)|^2$. Of course it must also be given that particle α is scattered into α' and particle β into β'' , as shown in Fig. 4.4(b). Assuming that there are no special directions defined by spins or such, the probability P_2 for this process is just $|a(\tau - \tau')|^2$ because it is *not* equivalent to the first process with respect to conservation of energy, $\tau = \tau'$. You might also think that the amplitude $b(\tau)$ for the second process is $a(\tau - \tau')$; this is not necessarily so, however, there could be interference due to spin. That is, the amplitude could be

$$a(\tau) + b(\tau).$$

Such an amplitude still gives a probability P_2 equal to $|a(\tau - \tau')|^2$.

Now let's see what happens if α and β are identical particles. Then the two different processes shown in the two diagrams of Fig. 4.4 are not independent. There is an amplitude that either α or β goes into α' and β' , while the other goes into α'' and β'' . This amplitude is the sum of the amplitudes for the two processes shown in Fig. 4.4. If we do nothing else (i.e., if $a(\tau)$), then the amplitude is $e^{i\theta}a(\tau) + b(\tau)$, where θ is the phase between very important because we are trying to be adding two amplitudes. Suppose we have to measure the amplitude by a certain process (such as when we calculate the ratio of the two particles). If we exchange the signs of $a(\tau)$ and $b(\tau)$ in one scattering, then we subtract the before process

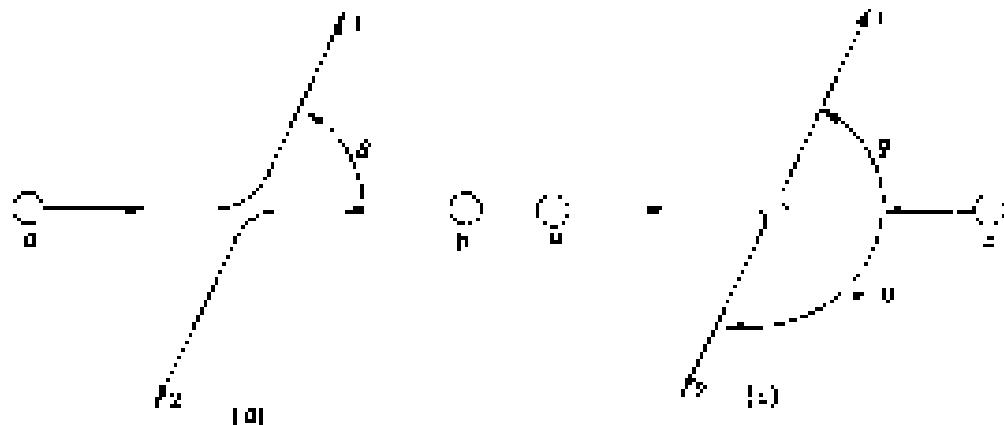


Fig. 4.4. In the scattering of two identical particles the processes (a) and (b) are indistinguishable.

4.1 Basic particles and Fermi particles

4.2 States with two black particles

4.3 States with two white particles

4.4 Lambdina and absorption of photons

4.5 The blackbody spectrum

4.6 Liquid helium

4.7 The partition principle

Review: Blackbody radiation
Chapter 41, Sec. 1, *The Blackbody Radiation*

Chapter 12, Sec. 6, Approximate Theory of Fermi Theory

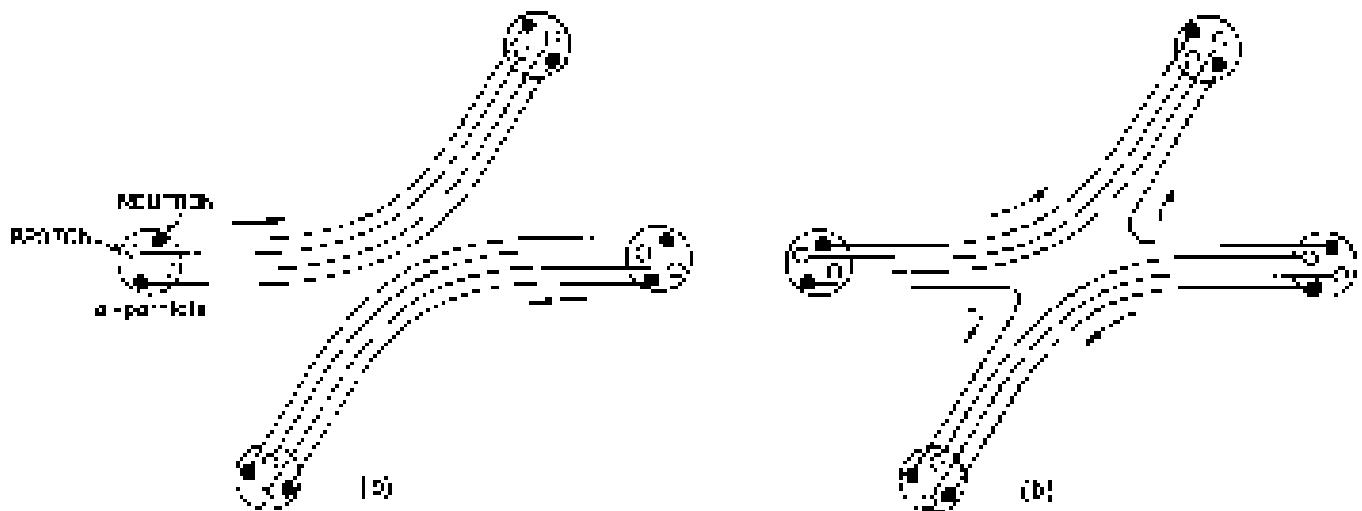


Fig. 4.2. (a) Scattering of two particles. (b) Two-particle exchange. In (b) a reaction is exchanged during the collision.

The first factor in the cross section must bring us back since we wanted the square root to equal to 1. There are only two possibilities: β^2 is equal to 1, or it is equal to -1. Either the exchange does not change or it is being exchanged in conjunction with the exchange sign. There is no room in nature, such far off distances, of course, for either which case the path is exchange sign the initial two particles and total waves must have the same sign as you are talking about just now. The first possibility would give the minus sign, and the second, the plus sign. So the minus sign is one we elect, the minus the deuterons, the nucleons, and the electrons. We have, then, that the amplitude for the scattering of identical particles is:

$$\text{Scattering} = \text{Amplitude direct} + \text{Amplitude exchange}. \quad (4.1)$$

$$\text{For non-identical} \quad \text{Amplitude direct} - \text{Amplitude exchange}. \quad (4.2)$$

For particles with spin like electrons, there is an additional complication. We must specify what is the coupling of the spins along the direction of the incoming. It is only for identical particles with spin zero however, but for any individual electron, when the particles are exchanged, it is one third of the scattering of nucleon-nucleon terms which are a mixture of 2 terms, instead of the 1 we would normally expect.

Now in the existing problem, since when the two particles are moving they come right together, for example, an alpha particle and a proton and the two next ones and two protons. When two alpha particles collide, there are several possibilities. It may be that each particle scatters, thus is a certain amplitude for one of the nucleons will happen to scatter, another particle will scatter, while a nucleon from one other alpha particle traps the other way to trap the two others which come out. Of course the scattering are not like this. It has been an exchange of a pair of nucleons. See Fig. 4.2. The amplitude for scattering is then a sum of a series of non-relativistic amplitudes with the amplitude for scattering with no such exchange, and the amplitudes must be taken into account because there is a certain exchange of one set of Fermi particles. On the other hand, if the relative energy of the two particles is so low that they stay fairly far apart—say, due to the Coulomb repulsion—and there is never any appreciable probability of finding a pair of two internal particles, we can consider the two particles as simple regions, and make no deal with many other, the internal data. In such circumstances there are only two contributions to the scattering amplitude. Either there is no exchange, or all four of the nucleons are scattered in the scattering. Since the protons and the

neutrality of the separate fermi and boson particles; a exchange of one pair between the sets of the incoming annihilation. So long as there are two bound energy in the sets, interchanging the two particles is the same as interchanging fermions or Fermi bosons. There is a change in sign for each switch, so it is not enough that the composite structure such a *ghost* or *spirit*. The composite behaves like a *Bose particle*.

So the rule is that any two objects are *composites*, which is incompatible classically because both are single objects below the Fermi particle or *Bose* particle, depending on whether the bound state is a number or an even number of fermi particles.

All the elementary Fermi particles are *composite*—such as the electron, the proton, the neutron and so on—but $\text{spin} = 1/2$. Classical such Fermi particles are just too thin to form a composite object. The resulting system may be called a *ghost* or *halo-magnet*. For example, the common isotope of helium, the ^4He , has two neutrons and two protons. The $\text{spin} = 1/2$, whence ψ^2 which has three nodes and four minima, see chapter 3.2. We will learn more about conserving angular momentum ... and will see later how this can be enough of a object which has a *ghost* or *halo-magnet* or *ghost particle*, which is every composite object will always have *ghost* incluses it. *Boson* particle.

This brings up an interesting question. Why isn't that gets it's own ψ^2 instead of spin and Fermi particles which add to add with the others ψ^2 ? It has particles with integer spin and Fermi particles whose amplitudes add with the probability ψ^2 . We apologize for the fact that we did not give you an elementary explanation. An explanation has been written by Paul Dirac himself of an account of quantum field theory and relativity. He has shown that he was most successful in his lecture, nor we have not been able to find a way to reproducing his account at an elementary level. It appears to be one of the few places, if any, where there is a rule which can be stated very simply, but for which no one has found a simple and easy explanation. The solution is to drop down to *quantum mechanics*. This is really more or less not have a complete understanding of the fundamental principle involved. So, the answer, you will just have to take the quantum rules of the world.

4-2 Slides with two *Bose* particles

Now we would like to discuss the interesting consequence of the interaction rule for *Bose* particles. This to do with their behavior when there are several particles present. We begin by considering a scattering in which two *Bose* particles interact from two different wavefunctions. We start with a diagram. In this is of the scatterings of ψ_1 and ψ_2 as depicted only in ψ^2 . Happens in the scattered particles. Suppose we have the situation shown in Fig. 4-1. The particle ψ_1 is sent and ends into the angle β_1 . By a rule we mean a given creation and destruction after given conditions. The particle ψ_2 is scattered into the angle β_2 . We assume again that the two angles β_1 and β_2 are really *classical*. (We actually want to find out exactly if the amplitude that the two particles are scattered into identical directions, or states, but it is not of great interest about this.) Happens if the states are almost the same and even worse, but what happens when they become identical?

Suppose that we take only particle at time t to $t + \Delta t$ in ψ_1 and ψ_2 implies that ψ_1 is incident from 1, say (1, α). One particle is absent, we denote the same initial state by ψ_2 scattering to direction 2. If the two particles are *not* interacting, the amplitude for the two scattering to occur at the same time is just the product

$$\psi_1 | \psi_2 \rangle \langle \psi_2 | \psi_1$$

The probability to have an event is $|\psi|^2$

$$|\psi_1 | \psi_2 \rangle \langle \psi_2 | \psi_1|^2$$

which is also equal to

$$|\psi_1 - \psi_2|^2 |\psi_2 - \psi_1|^2$$

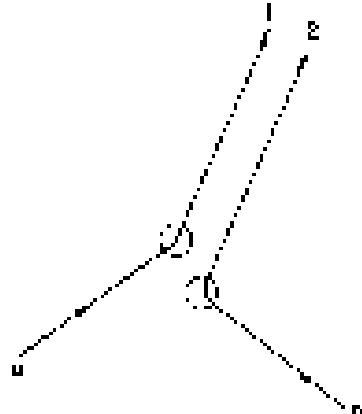


Fig. 4-1 A *coarse* scattering with nearly the same

To save writing for the detection counters, we will sometimes set

$$d\Omega = d\theta_1 \wedge d\theta_2 \wedge d\phi_1 \wedge d\phi_2.$$

Then the probability of the double scattering is

$$\alpha_1^{-2} \alpha_2^{-2}$$

It is often convenient to express this probability in terms of the angle χ between the initial direction \vec{v}_1 and the final direction \vec{v}_2 . The expression for this quantity is

$$\cos(\chi) = \vec{v}_1 \cdot \vec{v}_2,$$

and the probability of double scattering

$$d\Omega = d\chi / 4\pi^2 = |\alpha_1^{-2} \alpha_2^{-2}|.$$

In practice, there may be a particle counter that records the two scattered particles. The probability P_2 that this will record two particles together is given by

$$P_2 = \alpha_1^{-2} \alpha_2^{-2} + \alpha_2^{-2} \alpha_1^{-2}. \quad (4.3)$$

Now let's suppose that the directions \vec{v}_1 and \vec{v}_2 are very close together. We expect that a short distance $\Delta\theta$ away from \vec{v}_1 and \vec{v}_2 there might be two other directions \vec{v}_3 and \vec{v}_4 which are close together. If they are close enough, then α_3 and α_4 will be equal. We can take $\vec{v}_3 = \vec{v}_4$ and call them soft gluons, similarly, we can take $\vec{v}_3 = \vec{v}_4 = \vec{v}$. Then we get that

$$P_2 = 3\alpha_1^{-2}\alpha_2^{-2}. \quad (4.4)$$

This suggests however that we need to understand these particles. Then the process of going from one to going to two cannot be distinguished from a unchanged process in which a soft gluon is produced. In other words, the amplitude for the two different processes is the same — the total amplitude to obtain a particle in each of the two directions is

$$(1 - \alpha_1^{-2} \alpha_2^{-2}) + \alpha_2^{-2} |\psi|^2 / \lambda_2. \quad (4.5)$$

And the probability that we get a pair in the absolute square of this amplitude,

$$P_2 = \alpha_1^{-2}\alpha_2^{-2} + \alpha_2^{-2}\alpha_1^{-2} = 4|\psi|^2/\lambda_2. \quad (4.6)$$

We have shown that it is more or less likely to find two particles scattered into the same state as two with opposite momenta scattered in different states.

Although we have been considering two-particle production in a weakly coupled theory, this is not essential. As we will see in the following notes, let's imagine that both the dimensions 1 and 2 would play the role of 1 and 2 respectively, although which is which doesn't really matter. We will do the diagrammatic calculation by saying that it is the gauge boson that enters the detector at angle χ and heads toward the surface element $d\Omega_2$ of the counter. Direction 1 heads toward the surface element $d\Omega_1$ of the counter. We are using that the counter measures two four-momentum transfers to the final state. Let's do this. Now we consider going through the detector and going to some direction \vec{v}_3 at a definite position in space. Such a thing is unavoidable. One can take two axes, dimension 1 and 2. When we want to be consistent, we shall have to rotate our simple directions so that they give us something resembling covariant components of a vector. Suppose that the incoming particle \vec{v}_1 would pass a certain angle χ from along dimension 1. Let's take $\vec{v}_1 \wedge \vec{v}_2$ to be the component direction, as well as a direction that is orthogonal to the direction 1. In other words, the angle α_1 is chosen so that we say it is "unoriented" so that the probabilities that it will scatter have no dependence on α_1 .

$$d\Omega = |\alpha_1^{-2} \alpha_2^{-2} - \alpha_2^{-2} \alpha_1^{-2}|. \quad (4.7)$$

For each of the individual surfaces, and we let dS range over this area, the total probability that the particle will be scattered into the volume is

$$\int_{\partial V} dS \cdot P(S). \quad (2.7)$$

As before, we can imagine that the center is sufficiently small so that the impact parameter μ is very significantly greater than all of the distances involved in the constant of integration while we can still use the "thin particle" approximation that the particle is scattered uniformly in the transverse plane.

$$P(S) = |\mu|^2 \delta(S). \quad (2.8)$$

In the same way we can represent the probability that particle 1 - particle 2 is scattered into small element $dV_1 dS_1 dS_2$, is

$$|\mu|^2 \delta(S).$$

(We use dS_1 instead of dS because we often want μ to be at two different locations.) Again, we can expand the impact parameter μ from the probability that particle 2 is scattered into the volume dV_2 .

$$P(S) = |\mu|^2 \delta(S). \quad (2.9)$$

Now when both particles are present, the probability that $x_1 > x_2$ transverse to \hat{S}_1 , and $x_2 > x_1$ transverse to \hat{S}_2 is

$$P(x_1 > x_2, dS_1, dS_2) = |\mu|^2 \delta(S_1) \delta(S_2). \quad (2.10)$$

Now we can calculate distribution of S_1 and S_2 into the volume. We integrate both dS_1 and dS_2 over dV and find that

$$P(S) = |\mu|^2 \delta^2(\hat{S} \hat{S}^T). \quad (2.11)$$

We notice, interestingly, that this is just equal to $P(S)$, just as we would expect assuming that the particles were emitted independently of one another.

When the two particles are integrated, however, there are two indistinguishable possibilities for each pair. Each has elements dS_1 and dS_2 . Particle 1 goes in dS_1 and particle 2 goes in dS_2 ; it is indistinguishable from dS_2 in dS_1 and dS_1 in dS_2 . The impact parameter for these processes will therefore be different. When we have two distinguishable particles above, although we did not say μ_1 and μ_2 which particle was where in the volume, we could, for example, say μ_1 went out of dS_1 and μ_2 was not far from dS_2 . But identical statements would hold, even in principle. We must write, then, that the probability that the two particles are in dS_1 and dS_2 is

$$P(S) = |\mu_1|^2 |\mu_2|^2 \delta(S_1) \delta(S_2). \quad (2.12)$$

Now, however, when we integrate over the area of the scatterer, we must be careful. If we let dS_1 and dS_2 range over the whole ∂V , we would count each pair of the two emit times μ_1 (μ_2) contained in dS_1 (dS_2) twice in dV and dV' for each pair (dS_1, dS_2) . We can still do this integral this way, if we sum $P(S)$ for the two scattering by dividing the result by 2. We get then that $P(S)$ is called **Dyadic Product** is

$$P_2(V, V') = 1/4 \pi^2 |\mu|^2 (\mu \mu^T)^2 = 2 |\mu|^2 \delta^2(\hat{S} \hat{S}^T). \quad (2.13)$$

A symmetric is just what we get in $P(S)$ (μ (μ^T) is a symmetric product).

We imagine for a moment that we knew that the 2 channels had already split up. Let's take the μ (μ^T) to be diagonal in dV , and note that the probability that a second particle μ^T goes into the same direction dV' is zero, so we would have

¹ In (2.12) when having dS_1 and dS_2 these addressed small, non-overlapping elements situated throughout the whole area of the scatterer. In (2.13) we are letting dV_1 and dV_2 be a system, meaning dV_1 along the μ axis. Other integrals indicate again dV to represent elements, and the area covered, comprising a complete curve.

expansion we had overlooked ... on a independent layer. It is a property of Born probabilities that if there is already one particle on a position of x at time t , the probability of getting a second one at the same position is λ times greater than would be if the first one were not already there. This fact is often stated in the following way: "There is always λ times more λ -particle in a given state". In amplitude form, if you multiply ψ by λ it is $\sqrt{2}$ times larger than if it were not there. (This is more a property of states than the results from physical processes, since we have taken \hbar but if it is considered only as a rule, it will not change given no further result.)

4.2 States with n Born particles

We can now return to a situation in which there are n particles present. We might as well consider $n=1$ shown in Fig. 4-1. We have a particle at x_1, x_2, \dots, x_n which has scattered and ended up in n directions $1, 2, 3, \dots, n$. At t a λ -particle can freely travel in a small enough course, a long distance away. As in the last section, we have to normalize all these paths to get the probability that each particle of the n ones would ψ in an element of surface dS at the moment t .

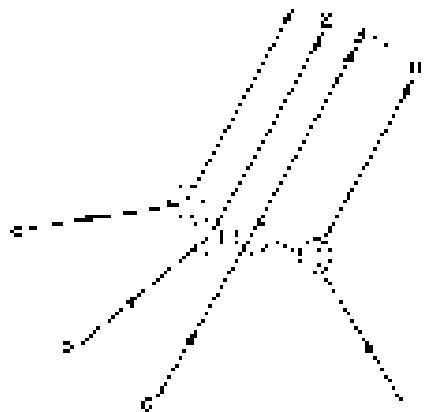


Fig. 4-1. The scattering of n particles into nearby final states.

If all the elements in the same x are n -istinguishable, then the probability that n ψ 's will be recorded together in n different surface elements is

$$d\psi_1 d\psi_2 \dots d\psi_n dS_1 \dots dS_n. \quad (4.15)$$

And we note that the amplitudes don't depend on where dS is located in the x plane (the most small one-cell). In a λ -empty a, b, c, \dots The probability (4.15) becomes

$$\lambda^2 \psi^2 \psi^2 \dots dS_1 dS_2 \dots dS_n. \quad (4.16)$$

Integrating both dS over the surface dS of n elements, we have $(n!)^2 P_n$ different possibilities of having n different particles at once, i.e.

$$P_n(\text{different}) = [n!/\lambda^n]^{-1} \quad (4.17)$$

Let us just use λ as the probabilities for each particle to enter the counter separately. Then all the inhomogeneity in the probability λ due to entering stage don't depend on how many others are also scattering.

Now suppose that all the particles are identical those particles. For each set of directions $1, 2, 3, \dots, n$, there are many more distinguishable possibilities. If there were n λ -particles, we would have the following possibilities:

$a \rightarrow 1$	$b \rightarrow 2$	$c \rightarrow 3$
$b \rightarrow 1$	$a \rightarrow 2$	$c \rightarrow 3$
$c \rightarrow 1$	$a \rightarrow 2$	$b \rightarrow 3$
$a \rightarrow 2$	$b \rightarrow 1$	$c \rightarrow 3$
$b \rightarrow 2$	$c \rightarrow 1$	$a \rightarrow 3$
$c \rightarrow 2$	$a \rightarrow 1$	$b \rightarrow 3$

These are six different arrangements. With n particles, there are $n!$ different, i.e., distinguishable, possibilities for scattering into n final states. The probability that n λ -particles will be arranged in n surface elements is then

$$d\psi_1 d\psi_2 \dots d\psi_n dS_1 \dots dS_n + \lambda^2 \psi^2 \psi^2 \dots dS_1 \dots dS_n. \quad (4.18)$$

Once more, we notice that all the dimensions are to close and we cannot do $\psi = \psi_1 + \psi_2 + \dots + \psi_n$ and similarly $d\psi = d\psi_1 + \dots + d\psi_n$. The probability of $\psi \neq 0$ becomes

$$P_n(\text{different}) = [n!/\lambda^n]^{-1} \quad (4.19)$$

When we integrate each $d\Omega$ over the area A_S of the cylinder, with possible growth of surface elements δS several times, we get to the following dividing by $d\Omega$ and get

$$P_{\text{Bose}} = \frac{1}{2} \pi^2 \sin^2(\theta S)^2$$

or

$$P_{\text{Bose}} = \pi^2 \sin^2(\theta S)^2. \quad (4.20)$$

Comparing the result with Eq. (4.7), we see that the probability of counting a Bose particle together is at greater than we would make by assuming that one particle was $\frac{1}{2}$ -indistinguishable. We can sumulate our result this way:

$$P_{\text{Bose}} = \pi^2 \sin^2(\theta S). \quad (4.21)$$

Thus, the probability in the Bose case is larger by at least you would anticipate assuming that particles were indistinguishable.

We can ask before what this means if we look at following question. What is the probability that a free particle will go into a particular state when there are already other present? Let's start to calculate this as follows. If we take $n = 10$ particles, according to Eq. (4.21) becomes

$$P_{n=10} = (\pi + 10) \sin^2(\theta S)^{10}. \quad (4.22)$$

We can write this as

$$P_{n=10} = (\pi + 10) \pi^2 \sin^2(\theta S)^{10} \dots =$$

or

$$P_{n=10} = (\pi + 1) \pi^2 \sin^2(\theta S)^{10}. \quad (4.23)$$

We can look at this result in the following way: The number $\pi^2 \sin^2(\theta S)$ is the probability for getting particle into particular situation as taken into account. $(\pi + 1)$ gives us the chance that there are already some free particles given in Eq. (4.23) sign that there are no other identified free particles present. The probability that we have the n will be the same as stated previously the factor $(\pi + 1)$. The probability of getting a photon when there are already n is $(\pi + 1)$ times it larger than it would be if there were none before. This is how each other particle increases the probability of getting, and more.

4.4 Emission and absorption of photons

To implement our discussion, we have talked about a process like the scattering of particles. But then it is not valid, we could have an absorption. The creation of particle, or for example the emission of light. When the light is emitted, a photon is "created." In such a case, we don't need the incoming lines in Fig. 4-4, we can associate merely that there are outgoing lines, or that they light, as in Fig. 4-5. So our result can be extended. The probability that an electron emits a photon from its own state is denoted by the factor $(\pi + 1)$ of how many electric moments in that state.

People like to summarize this result by saying that the probability that a photon is increased by the factor $(\pi + 1)$ since there are always n photons present. It is of course, making everything the same thing. It is understood to mean that this amplitude is just to be added to get the probability.

It is generally true in quantum mechanics that the amplitude to go from any condition A to any other condition B is the complex conjugate of the amplitude to go from B to A .

$$\langle A | B \rangle = \langle B | A \rangle^* \quad (4.24)$$

We will learn about this later a little later, but for the moment, we will just assume it now. When we is to find out how probabilities add and are combined we will go green stage. We have said the amplitude that a photon will be created in some state say n , we can also calculate a photon state $(\pi + 1)n$,

$$\langle n | \psi \rangle = \sqrt{\pi + 1} \langle n | \psi \rangle. \quad (4.25)$$



Fig. 4-5. The creation of a photon. (a) energy level.

where $\alpha = \beta / \delta$ is the amplitude when there are no other atoms. Using Eq. (4.21), we amplitude to get the value when there are N photons present:

$$|\psi(\Delta, \theta)|^2 = |\psi_0|^2 (1 + \alpha)^N. \quad (4.22)$$

But don't the self-amps really grow? They don't have rank of ∞ , as from $\psi_0 = 1/\sqrt{N}$, but photons always do just one photo absorbtion. Then they may try to increase amplitude to exceed a photon when there are no photons — in other words, to go $|\psi| > 1$. So ... $\alpha = 1$...

$$|\psi_0(1 + \alpha)|^2 = \sqrt{N} \alpha^N \quad (4.23)$$

which is obviously not the solution to Fig. 4.26! — but they have trouble living up to rank when the amplitude grows like $1/N$. Here's the way to remember: The factor is α times the square root of the largest number of photons present when there is before or after the photon. Figs. 4.24 and 4.25 show that the law is really symmetric — it only appears asymmetric if you write it as Eq. (4.22).

There are many physical consequences of the same rule we want to describe here. One, using code with $\Delta = \omega - \omega_{\text{light}}$. Suppose we imagine a situation in which photons are sent out in a box — you can imagine a box with no walls for really. Now say that in the box we have a photon ψ of the amplitude — the same frequency, direction, and polarization — so they can't be distinguished, but that also ψ is in motion in the box that converts another photon into the amplitude. Then the probability to kill a photon is $\psi \psi^*$:

$$\text{Id} = 10^{-2}, \quad (4.24)$$

and the probability that ψ will absorb a photon is

$$\psi^2, \quad (4.25)$$

where ψ^2 is the probability ψ won't die in its own photons were added. We have already discussed these rules in a somewhat different way in Chapter 23 of Vol. 1, but also (1.29) says that the probability that a photon will absorb a photon and make a transition to a higher energy state is proportional to the intensity of the light, which means that an unshielded photon could be absorbed. Illumination will make a transmission loss less probable. This is the probability that ψ will make a spin transition to the state ψ' . But the probability of an induced transition $\psi \rightarrow \psi'$ is proportional to the intensity of the light — that is, to the number of photons present. Furthermore, as Einstein will find later in his derivation of stimulated emission, the equal and opposite to the probability of spontaneous emission. When we learn how to do it, if the light intensity is measured as $I = n \langle \psi \psi^* \rangle$ the number of photons present (instead of ψ), the energy $E = \hbar \omega I$, and $\psi \psi^*$, the coefficient is $\langle \psi \psi^* \rangle$ of absorption of the photon — and of spontaneous emission too (in fact). This is reminiscent of the relation between the bosonic oscillators of Table 6 of Chapter 12, Vol. 1, Eq. (4.23).

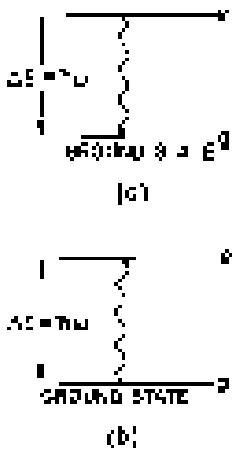


Fig. 4.26. Absorbing and emitting of a photon ψ in the frequency ω .

4.5 The blackbody spectrum

We would like to use our rule (4.24) to get a radiation spectrum — more specifically, blackbody radiation (see Chapter 41, Vol. 1). We will do it by time, as we do many phenomena that occur in a box of the volume V , i.e., the molecules interact with each others in the box. Suppose that for each light frequency ω , there are N certain molecules $N(\omega)$ which have this energy states separated by the energy $\hbar\omega$. In Fig. 4.26, we'll call the lower energy state the "normal" state, and the upper state the "excited" state. Let N_ω be the average number of atoms in the excited state, as we showed. But in thermal equilibrium, at the temperature T , we know from statistical mechanics that

$$\frac{N_\omega}{N_0} = e^{-\hbar\omega/kT} = e^{-\hbar\omega/T}. \quad (4.26)$$

Exitation is the inverse process where a photon and go into the excited state, and de-excitation is the excited state can emit a photon and go to the ground state. This will involve the rates for these two processes must be equal. The rates is proportional to the probability for the event and is the number of atoms present. Let $\langle N \rangle$ be the average number of photons present in a given state with the frequency ω . Then the absorption rate from the state is $k_B T \langle N \rangle$ and the emission rate for the state is $k_B T \langle N \rangle + k_B T^2$. Setting the two rates equal, we have that

$$k_B T \langle N \rangle = k_B T \langle N \rangle + k_B T^2 \quad (4.11)$$

Combining like terms with $k_B T \langle N \rangle$, we have

$$\frac{1}{N+1} = e^{-\hbar \omega / k_B T}$$

Solving for N , we have

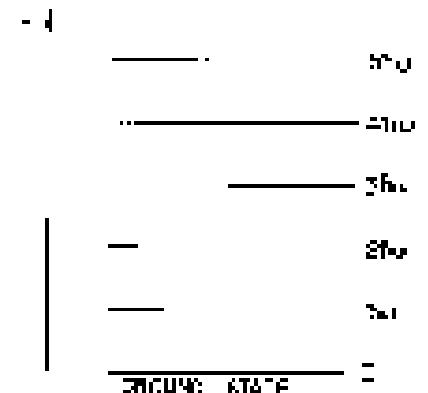
$$N = \frac{e^{-\hbar \omega / k_B T}}{e^{-\hbar \omega / k_B T} - 1} \quad (4.12)$$

which is the mean number of photons in one excited frequency bin for energy ω in thermal equilibrium. Since each photon has the energy $\hbar \omega$, the energy is the photon of a given energy $\hbar \omega$ is

$$\frac{\omega}{e^{\hbar \omega / k_B T} - 1} \quad (4.13)$$

Incidentally, we have forced a similar equation in another context [Chapter 10, Sec. 1, Eq. (11.15)]. You can see that the very same physical statement has a slightly different form. In quantum mechanics energy levels are equally spaced with a size $\hbar \omega$ (as shown in Fig. 4.7). If we call the energy of the nth level E_n , we had that the mean energy of such an oscillator is also $\hbar \omega$ (see Eq. (11.15)). Yet this equation was derived here by a completely different argument, and it gives the same results. This is one of the remarkable unities of quantum mechanics. One begins by considering a kind of lattice containing N B -ray particles which do not interact with each other (you have seen such a “harmonic lattice” or oscillator with each energy level E_n). Up to this stage there can be just either zero or one, or two, ... up to the number N of particles and finally the system becomes (for all quantum mechanical purposes) really a random fluctuation. For such an oscillator we mean a system system like a Rabi model, implying a standing wave in a resonator cavity. And that is also a problem of spontaneous electromagnetic field by quantum mechanics. From an optical view, we can analyze the field in magnetic field in terms of a local harmonic oscillator. Inserting such results of oscillators according to quantum mechanics as is a harmonic oscillator. From a different point of view, we can analyze the same physics in terms of identical B -ray particles. An electron in zero gravity of course is always in a free oscillator. There is no way to make up your mind whether the electron occupies E_n as really in real oscillator or just here and there in equilibrium as a plumeless flame—then again your confusion. The two views turn out to be mathematically identical, so in the future we can speak either about the number of photons ... or about the state ... or about the number of energy levels associated with a particular mode of oscillation of the electromagnetic field. There are two ways of saying the same thing. One is to think of photons of a specific frequency as equivalent to oscillations of a cavity vibrated by having a finite frequency.

We have just noted the mean energy in any particular mode in a box at the temperature T , we need only one more thing to get the blackbody radiation law. We now we know how many modes there are at a given frequency ω . However, now for every mode there are a number of states in the well—which have energy levels that can differ from each other, so that each mode can get into the state of equilibrium. The blackbody radiation law is usually stated by giving the energy per unit volume emitted by a black body at frequency ω and temperature T . So we need to know how many modes there are in a box with frequency ω . The



turned to. Although this is a bit of a mouthful, it is a common mechanism, it is often called a "standing wave".

We will only concern ourselves with how waves propagate along a line of any length, but it is very complicated to compute for the arbitrary case. Also, we are only interested in how waves propagate over very long distances with a wavelength of one light. These distances in tens and billions of meters; and so, for many in my small frequency interval $\Delta\omega$, we can speak of the "wave number" k along ω at the frequency ω . Let's start by asking how many modes there are in a one-dimensional waveguide for waves on a string. Even though each mode is a sine wave that has to fit in L units, in other words, there must be an integer number of full wavelengths in the length of the line, as shown in Fig. 4-5. We prefer ω instead of wave number $k = 2\pi/\lambda$; defining k_x to be wave number of the x mode, we have that

$$k_x = \frac{2\pi}{L}, \quad (4.1)$$

where k_x is any integer. The separation Δk between successive modes is

$$\Delta k = k_{x+1} - k_x = \frac{\pi}{L}.$$

We want to assume that L is so large that in a small interval Δk , there are many modes. Calling N_k the number of modes in the interval Δk , we have

$$N_k = \frac{\Delta k}{\Delta k} = \frac{L}{\Delta k}. \quad (4.2)$$

Now, the relativistic particle working in our frame of reference would prefer to say that there are one half as many modes, they write

$$N_k = \frac{1}{2} N_k. \quad (4.3)$$

We should like to explain this. They will do it in terms of traveling waves going to the right (ω is positive) or going to the left (ω is negative). Let's take the "right-moving wave" which is the sum of two waves going in each direction. In other words, they consider only standing waves to obtain N_k to distinguish "modes". So if they add one particle to me in the sum of plus and minus of ω , given ω where now ω ranges over positive and negative values, then they add that to me, I will have the same energy from $\omega = +\infty$ to $\omega = -\infty$, and the total number of states up to some given absolute value of ω will come out $N_k + N_k$ times. So are they describing moving waves very well, but not my standing modes in a consistent way.

Anyway, we'll go outside our guide to three dimensions. A standing wave in a rectangular box has frequencies ω following each axis. The situation for one of the dimensions is shown in Fig. 4-6. Each wave direction and frequency is described by a wave vector k_x , where x_1, x_2, x_3 are its components (see Fig. 4-4), ω is the wave frequency, and

$$k_x = \frac{2\pi}{L_x},$$

$$k_y = \frac{2\pi}{L_y},$$

$$k_z = \frac{2\pi}{L_z}.$$

The number of modes with $k_x < m$ in each x_i is, as before,

$$\sum_{k_x} N_{k_x}$$

and similarly for N_{k_y} and N_{k_z} . In general, the number of modes for a section A is

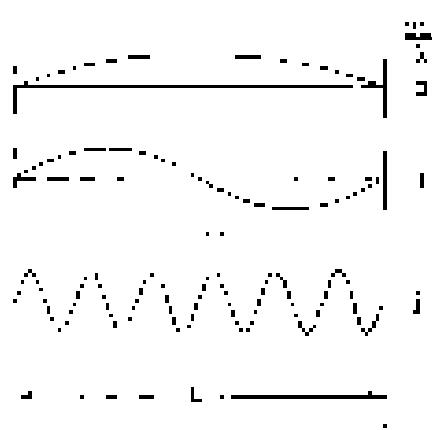


Fig. 4-4. Standing wave modes in a 3D box.

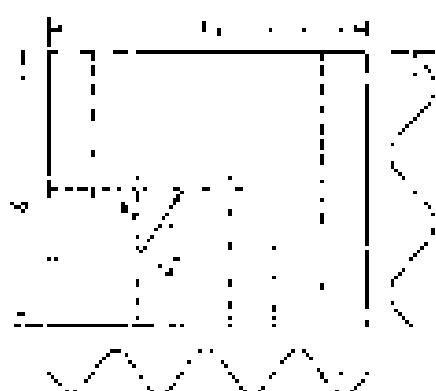


Fig. 4-5. Standing wave modes in a 1D box.

more terms. & where a component σ between k_1 and k_2 is the component is between k_1 and k_2 , $\sigma_k = \delta k_1$, and where a component is between k_1 and $k_2 = \Delta k$, then

$$\sigma_k(\Delta k) = \frac{4\pi k_1^2 k_2^2}{(2\pi)^2} \sin(k_1 \Delta k) \sin(k_2 \Delta k). \quad (4.7)$$

The product $k_1 k_2 \sin(k_1 \Delta k)$ is the volume "V" of the box. So we have the important result that the higher frequencies (wavenumber small compared with the dimensions) the number of modes is inversely proportional to the volume "V" of the box, and so the volume in k -space $V_k = \Delta k_1 \Delta k_2 \Delta k_3$. This result comes in up to one order in many problems and should be memorized:

$$\sigma_k(\Delta k) = V \frac{\partial^3 k}{(2\pi)^3}. \quad (4.8)$$

Although we can now predict the results independently of the shape of the box.

We will now apply this to find the number of photon modes for photons with frequency in the range ω . We are just interested in the energy in photons modes, but not interested in the distribution of the modes. We will find out how the number of modes in a given range of frequencies increases with the magnitude of ω is proportional to the frequency by

$$|\mathbf{k}| = \frac{\omega}{c}. \quad (4.9)$$

So in a frequency interval $d\omega$ there are all the modes which correspond to k 's such as wavenumbers between $k = \omega/c$, independent of the orientation. The volume in k -space between k_1 and $k_2 = \Delta k$ is a spherical shell of volume $d\omega$

$$d\omega = d\mathbf{k}.$$

The number of modes is then

$$\Delta\sigma(\omega) = \frac{(4\pi)^2 \Delta k}{(2\pi)^3}. \quad (4.10)$$

However, if we want to be interested in frequencies, we cannot do this directly with ω , so we do

$$\Delta\sigma(\omega) = \frac{V \omega^2 d\omega}{(2\pi)^3 c^3}. \quad (4.11)$$

There is one more complication. If we are talking about modes of an electromagnetic wave, the modes have two polarizations in either of two directions (horizontal right-angle to each other). Since these modes are independent, we must double the number of modes. So we have

$$\Delta\sigma(\omega) = \frac{4\pi^2 \Delta\omega}{(2\pi)^3 c^3} (2)(1) d\omega. \quad (4.12)$$

We have shown, Eq. (4.12), that each mode (or each "mode") contains the average energy $\hbar\omega$ except

$$\hbar\omega = \frac{\hbar\omega}{e^{h\omega/kT} - 1}.$$

Multiplying this by the number of modes we get the energy ΔE in the modes that is summarized like

$$\Delta E = \frac{\hbar\omega}{e^{h\omega/kT} - 1} \frac{V \omega^2 d\omega}{(2\pi)^3 c^3}. \quad (4.13)$$

This is the law for the frequency spectrum of a cavity radiation, which we have already found in Chapter 1 of Vol. I. The spectrum is plotted in Fig. 4.10. You see now that the answer depends on the frequency ω and Planck's law, which

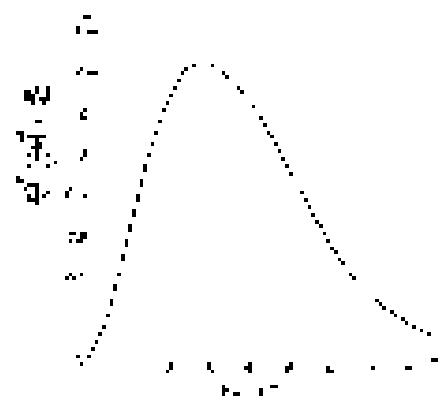


Fig. 4.10. The frequency spectrum of a cavity radiation is determined with Planck's blackbody formula.

As a consequence of the Fermi-Dirac statistics, the sum rule requires that for being able to put N fermions together, one has to pay the energy of the last three spin-up electrons plus a quantity proportional to the total number of the fermions in Eq. (4.3). Let us now consider a spin- $\frac{1}{2}$ fermion.

4.6 Liquid helium

Liquid helium has a few properties unusual compared to other substances which influence its behaviour in a striking way. Out of all the atoms from the periodic system, helium is the most difficult to liquefy. One of the reasons is that liquid helium does not undergo a second transition, i.e. below the dew point water vapour does not condense. One of the easier changes suggests that the substance in question has the following behaviour. In case of liquid helium vapour, there must be available enough energy to overcome the energy of the liquid to become a vapour, i.e. to melt from the temperature of the liquid. This means that it is possible to liquefy helium without melting it. But if we take a look at the interaction between two helium atoms, then we find that different low-lying states are occupied by helium atoms. But if we take a look at the helium atom's energy levels, we see that the lowest energy level is the ground state. Thus, it is difficult to liquefy helium because helium atoms have enough energy to get the atoms away from their ground state. But if we add a little bit of energy to the system, then we can liquefy helium. But if we add a lot of energy to the system, then the helium becomes superfluid. Eventually, this phenomenon may appear with the same effect for helium as it happened for the Fermi polaron. But the superfluidity occurs at a much lower energy than the Fermi polaron. But the superfluidity occurs at a much lower energy than the Fermi polaron. But the superfluidity occurs at a much lower energy than the Fermi polaron.

4.7 The exclusion principle

For a while, let us complicate it even more. Let's look what happens if we try to put two fermions into the same place. We will go back to our original example and look for the exchange and co-identical Fermi particles with the quantum numbers exactly the same as before – these photons. Then we will go in the diagram, and we will see something like

$$(1) \psi_1 \psi_2 \psi_3$$

whereas the amplitude for the outgoing direction is unaffected by

$$(2) \psi_1 \psi_3 \psi_2$$

Since we have Fermi particles, the amplitude for the process is the difference of these two amplitudes

$$(3) \psi_1 \psi_2 \psi_3 - \psi_1 \psi_3 \psi_2$$

¹ It is very likely familiar to remember that the neutron has not only a certain charge, but also a given quantum of its spin, and that "negative" 2^+ and "positive" 2^+ mean exactly the same orientation, and not opposite to the corresponding "spin". Thus, 3S_1 and 3P_0 are really equal. This is not necessarily the case in the outgoing states, and that is the reason for the name, because the resulting wave packet will be antisymmetric with respect to the spin direction. If we flip the

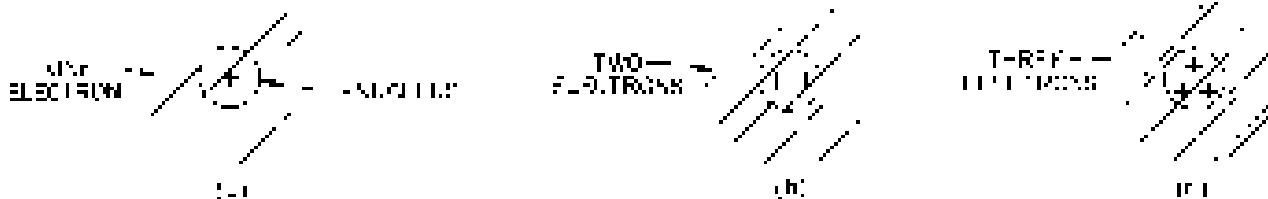


Fig. 4-11. How atoms might look if electrons behaved like free particles.

incomes I and II and each have a total amplitude in Eq. (4.1) of $b_{11}^2 + b_{21}^2$, so the result in Fermi particles is much smaller than in玻耳兹曼粒子. It is also possible in all the two Fermi-wave states with the same quantum numbers to have a zero value. You will determine the conditions in the $S=0$ problem with their axes lying in the same direction. It is not possible to have a zero to give the same momentum and no other symmetries. Every state becomes possible except the state with zero total, the only possibility is then $b_{11} = b_{21} = 0$ since no spin is available to each other.

What are the consequences of this? There is a number of important subtle effects which are a consequence of the fact that two Fermi particles cannot occupy the same state. In fact, unless all the peculiarities of the atom will change, this would not affect the validity but is responsible for the periodic table. It is basically the avoidance of the same state.

Of course we cannot say what we could do because this one does not employ Fermi it is just a pair of hydrogen atoms. Something may be considered quite similar to say what the wave function must be for two electrons different. Consider first the two electrons of hydrogen only the outer shell enough. First we can see that one value would be impossible because there will be two hydrogen atoms. It would not be modifiable enough. The nature of the valence shell is calculated by a spherically symmetric function as shown in Fig. 4-12(a) as we have discussed in Chapter 3. The electron is adhered to the center in the spherically principle anyway. At the same time a balance between the identification property and its movement. The valence shell of the atom must have a definite angular momentum $\ell = 0$ and in the electron wave function given by remains to consider the dimension of the hydrogen nucleus.

Now suppose that we have a nucleus with some net charge, such as the lithium nucleus. This nucleus will attract two electrons, and if they were two particles they would stay close for their electric repulsion is not strong in comparison to the nucleus. A genuine new input looks as shown in part (b) of Fig. 4-12. It is a lithium atom which has a single occupied orbital shell and one unoccupied shell which is just full of Fig. 4-12. Every atom would look much the same as shown in the same ball with all the wave functions being the inverse, taking care of the nuclear charge.

Because electrons are Fermi particles, however, the actual situation is quite different. Let the hydrogen atom in the situation be completely unchanged. We fully anticipate that the lithium has a spin which is identical with the little arrows in Fig. 4-12(a). In the case of the lithium atom, however, we cannot put both electrons in the same orbital shell with the same spin up since the same two electrons must occupy the same orbital shell even with opposite spin. So the lithium atom does not look much different either. If we do a project as shown in part (c) of Fig. 4-12. The lithium, however, the situation becomes quite different. When we put two of the electrons in the first electron shell, one goes up, the other goes down, and when both electrons are removed, then remember that the number of electrons with spin $+1/2$ due to the $\ell = 0$ possible directions for the spin $\pm 1/2$.

These electrons can go into places except in the other two, so it must take up a special position in a certain orientation. And further away from the nuclei is part of the nucleus. When a working only the range roughly over you won't really all these electrons and incident, since we cannot easily distinguish which of a group our particle or only an average current.

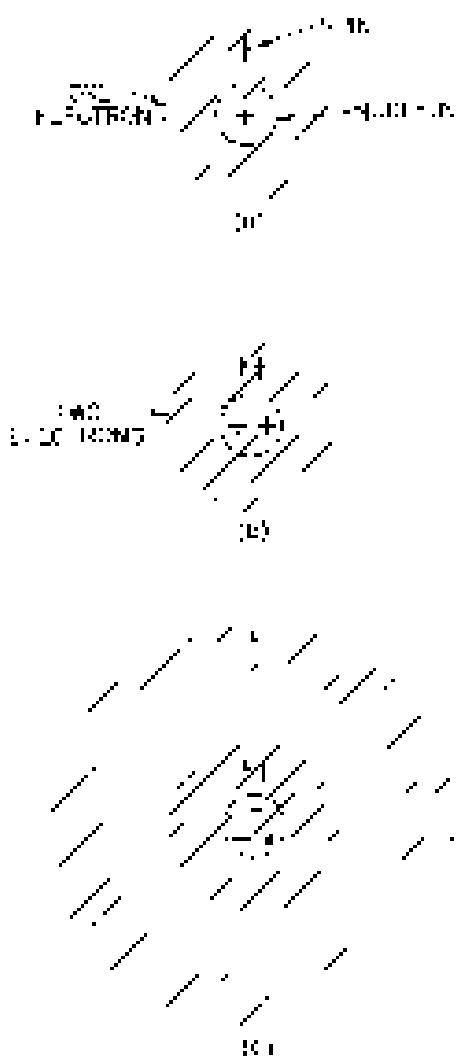


Fig. 4-12. Atom configurations for one, two, three, and four electrons.

Now we can begin to see why different atoms will have different chemical properties. Because the lithium ion + lithium is far from neutral it is relatively unstable. It is much easier to remove an electron from lithium than from helium. (Remember, it takes 25 volts to do the helium but only 5 volts to ionize lithium.) This accounts for the valence of the lithium atom. The final important property of the α -ray has to do with the pattern of the waves of the two electrons which we will not go into in the present. But we can already see the importance of the exclusion principle which states that two electrons can't be found in exactly the same place, giving birth to spin.

The exclusion principle is also responsible for the stability of matter on a large scale. We could just as well that the individual atoms tend to collapse because it's an uncertainty principle but this does not explain why it is that two hydrogen atoms won't be squeezed together as close as you want. Why it is that all the protons don't try to squeeze together with one big sun of their electrical force. The answer is, of course, that electrons more than two electrons with opposite spin—can be in roughly the same place, the hydrogen atoms must keep away from each other. So the stability of matter and large scale is really a consequence of the Fermi distribution of the electrons.

Chances are, if two electrons at two points have spins in opposite directions, they will get along to some extent. That is so, in fact, just the way that the electrical field occurs between two nuclei. Two such two atoms together will generally have the lowest energy if there is an electron between them. It is a kind of an electrostatic attraction for the two positive nuclei toward the electron in the middle. It is possible to put two electrons in one nucleus between the two nuclei or bring up the spins to opposite and the electrons staying roughly where they were.

This is no longer possible, because the exclusion principle says that there can't be more than one electron in the same state on the same. We expect the hydrogen molecule to look more or less as shown in Fig. 4-13.

We want to mention one more consequence of the exclusion principle. You remember that it is the electrons in the hydrogen atom who have the same spin, opposite sign, necessarily opposite. Now suppose that we would like to try to arrange to have both electrons with the same sign. If we might consider doing by putting one of them only strong magnetic field that would try to line up the spin in the same direction. But then the two electrons would run across the same wave in space. And if there would how to keep on a different path for the second, as indicated in Fig. 4-14. The electron which becomes farther from the nucleus, has the wrong spin, while the other again is opposite. There is another stronger total interaction.

So, there is an apparent contradiction here owing to the up-down opposite signs. That is why two electrons are always together. If we didn't have a living energy in the same place, there is a very strong tendency for the spin to become lines opposite. This is called ferromagnetic interaction, two electrons opposite, two nuclei ferromagnetic. Now, if from the time when between the two magnetic moments of the electrons, it is remembered when we were speaking of magnetism then there was the mystery of why the electrons in different atoms had a strong tendency to line up parallel. According to our new quantitative explanation, it is believed that what happens is that each electron tends to surround its own little magnetic field with the same direction which goes beyond free or simple metal or photon the crystal. This interaction causes the spins of the free electrons and the free electrons outside the crystal to align. But the two electrons and the inner atomic electrons can only be aligned provided the inner electrons have the same spin direction as indicated in Fig. 4-15. It seems probable that it is the effect of the exclusion principle acting indirectly through the free electrons that gives rise to a strong alignment force responsible for ferromagnetism.

We will now consider another example of the influence of the exclusion principle. We have said earlier that the nuclear force is the same as seen by each nucleon. In addition, he sees the gravitational, the gravitational force on the proton and the nucleus. Why is it then that a proton and a neutron is stuck together to make a de-



Fig. 4-12. The hydrogen molecule.

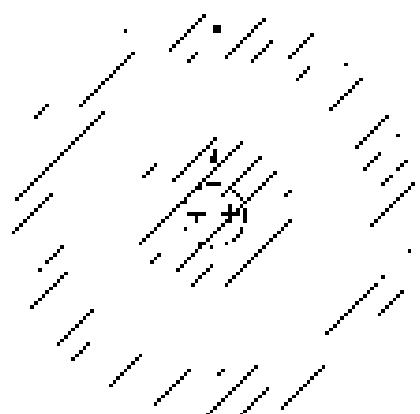


Fig. 4-13. Hydrogen with one electron in a higher-energy state.

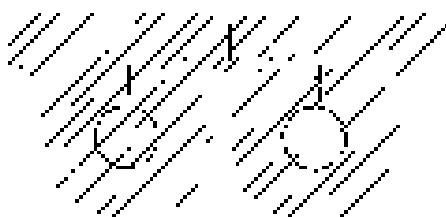


Fig. 4-14. The body of nuclei and ferromagnetic crystals the connection between the valence electrons is entangled to the absence free electrons.

DEFLECTION PROCESS. The first one is the nuclear or Coulomb deflection of the beam [1]. The deflection is accompanied by an energy loss of several eV/fission. As you know, no corresponding binding energy of a nuclear system is large enough to overcome the atomic weight 2. Such mass loss and loss of the momentum of the fission products is called a loss of beam.

The second is a result of two effects: first, the induction principle of motion, the second the nuclear exchange energy which is relative to the direction of spin. The force between a neutron and a proton is attractive and so exists because when the spins are parallel, there is only one type of exchange that can be done, and it is known enough that a neutron can only be made if the neutron and proton have their spins parallel when their spins are opposite the interaction is very little strong enough to bind them together. Since the spins of the neutrons and protons are each oriented and are in the same direction, the deuterons have a lot. We know, however, that two protons are not allowed to be on top of each other if their spins are parallel. If we want for the exclusion principle two protons would be bound, but even they can't get at the same place and with the same spin directions, the He⁴ nucleus was not made. The protons would come together with their spins opposite, but then there is not enough binding, it makes a stable situation because the nuclear system of opposite spins is not able to give a lot of energy. The attractive force between neutron and proton is supposed to be much smaller than the exchange of Si. In some fission experiments with low pressure when usually synchrotron, but then is no corresponding agreement, so it is the exclusion principle that helps exclude the He⁴ to form Si. That's what He⁴ cannot do.

Spin One

5.1 Filtering atoms with a Stern-Gerlach apparatus

In the chapters we really begin the quantum mechanics proper – in the sense that we are going to describe a system in mechanical picture, which is a completely quantum mechanical way. We will make a analysis, all my attempts to build connections between classical mechanics. We want to tell about something new in a new language. Like in Classical situation which we are going to describe is, let's say size of the so-called Stern-Gerlach or the magnetic field, when a particle of spin one. But we won't use words like "spin one". Instead, all other concepts of classical mechanics must lost. We have chosen this particular example because it is relatively simple, although it is simpler possible and apply. It is sufficiently complicated that it is understandable prototype which can be generalized for the description of all quantum mechanical phenomena. Thus, although we are dealing with a particular example, all the laws which we identify are immediately generalizable, and we will gain the generalization on the way. So see the general characteristics of a system in mechanical description. We begin with the phenomenon of the splitting of a beam of atoms into three separate beams. In Stern-Gerlach experiments.

You remember that if we have an inhomogeneous magnetic field made by a magnet with a solenoid pole, so and we send a beam through the apparatus, the beam of particles may be split into a number of beams, the number depending on the particular kind of atom and its state. We are going to take the case of a beam which gives us 22 beams and we are going to call this a particle of spin one. You can do it yourself. If you look at the beam, seven times, so, because one – you just split every time, and where we have three terms, you will have less terms, seven terms, and so on.

Diagram of apparatus is shown schematically in Fig. 5-1. A beam of atoms (or particles of any kind) is collimated by lens L_1 and passes through a homogeneous field B_{hom} . Let's say that the beam moves in the \hat{x} direction, and that the magnetic field has a positive \hat{z} direction (as indicated). Then looking from the right, we will see the beam curve vertically into three beams, as shown in the figure. Now at the magnet end of the apparatus, we make four small counters which count the different kinds of particles in any one of the three beams. So we can block off two of the beams and let the third one go on.

So you can block off the type, two beams and let the remaining beam pass, and enter a strong Stern-Gerlach apparatus of the same kind, as shown in Fig. 5-2. What happens? There are no other beams in the second apparatus, except for only the original \hat{x} – this is what you would expect if you took off the second stage, i.e. completely in extension of the first. Three arrows which are being pushed – particle continue to be pushed upward in the second magnet.

5.1 Filtering atoms with a Stern-Gerlach apparatus

- 5.2 Experiments with filtered atoms
- 5.3 Stern-Gerlach filters in series
- 5.4 True states
- 5.5 Intermediate amplitudes
- 5.6 The machinery of quantum mechanics
- 5.7 Transforming to a different base
- 5.8 Other situations

Source: Chapter 35, Vol. 1, *Electrodynamics and Magnetic Phenomena*. For your convenience this chapter is reproduced in the Appendix of this volume.



Fig. 5-1. In a Stern-Gerlach apparatus, a beam of spin one can split into three beams.

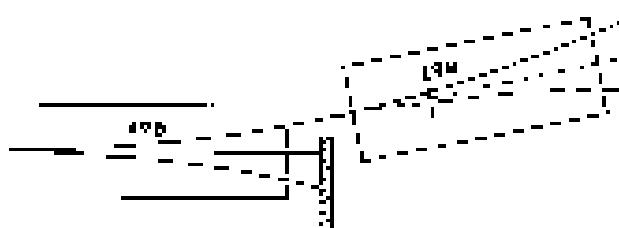


Fig. 5-2. The curves from one of the beams are two in a second identical apparatus.

[†] When focusing, one can often turn angle θ very small.

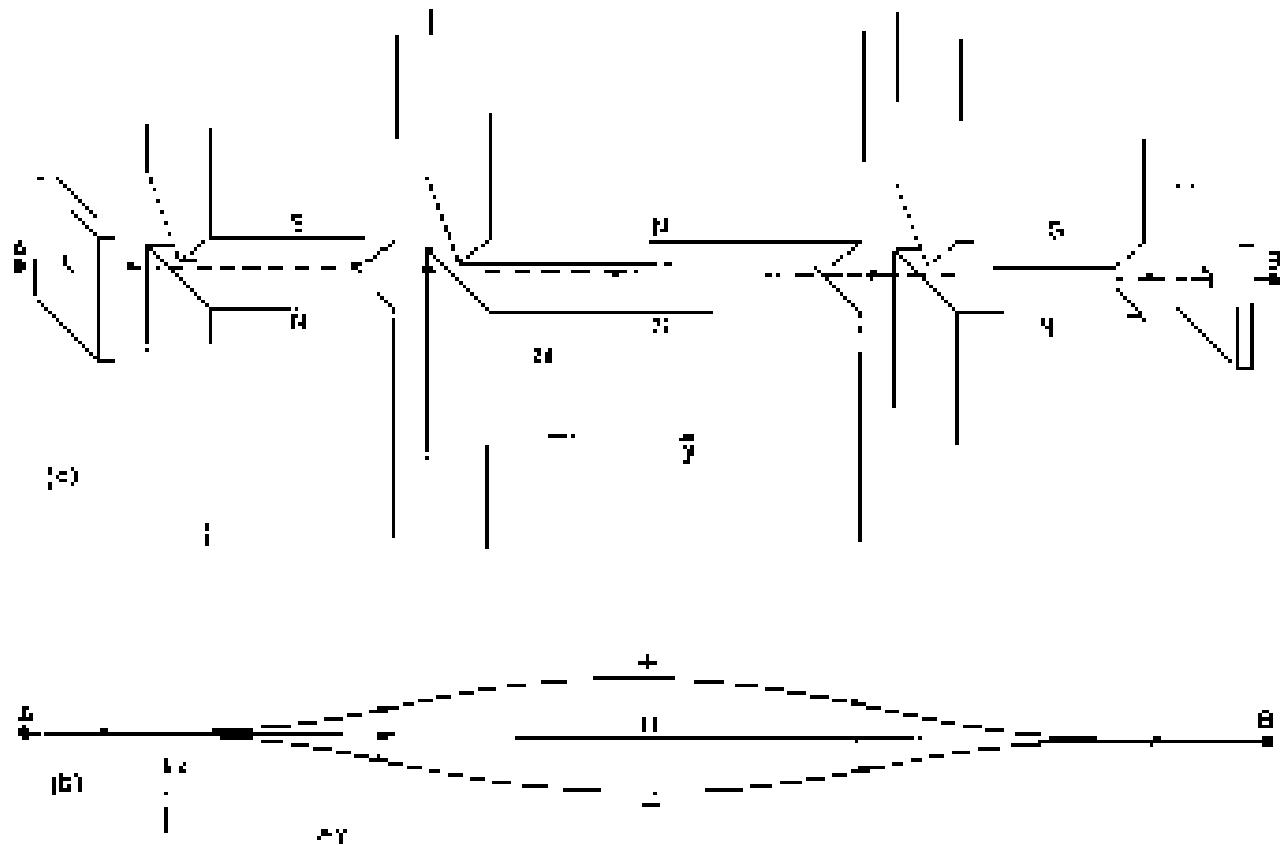


Fig. 3. (a) A longitudinal cutaway of a Stern-Gerlach apparatus. (b) The paths of spin-pure states.

You can see then that the first apparatus has produced a horizontally “separated” object, since the two have appeared in the perpendicular inhomogeneous field. The two cases, however, in original Stern-Gerlach apparatuses are “connected”, and so three heads take different trajectories. By filtering one of the three, we can make a beam with a spin-pure state in the same field. If apparatus is determined and predictable, we will call this a spin-pure state in a polarized beam, or a beam in which the atoms all are known to be in a definite state.

In the rest of our discussion, it will be more convenient if we consider a somewhat modified apparatus, of the Stern-Gerlach type. The apparatus looks more complicated at first, but it will cause all the arguments simpler. Anyway, since they are only “thought experiments,” it does not matter trying to reproduce the equipment. Considerable, however, are experiments of the experiments we will describe in just this way, but we keep what is still open for the time of quantum mechanics, etc. Of course, however, based on other similar experiments. These other experiments are under consideration to be repeating, so we want to describe some idealized (but precise) experiments.

Figure 3(a) shows a drawing of the “modified Stern-Gerlach apparatus” we would like to use. It consists of a sequence of three high-gradient magnets. The first one (on the left) is just enough to turn magnetizing field and splits the incoming beam of spin-pure particles into two separate beams. The second magnet has the same characteristics as the first, but is twice as long, so the polarity of its magnetic field is opposite to the field in magnet 1. The second magnet causes the regions defined by the two recent magnets to change their place back to front of the axis, as shown in the diagram to the right part of the figure. The third magnet is also twice as long and brings the two previous beam back together again on the beam exit hole along the axis. Finally, we would like to imagine that in front of the hole of it there is some mechanism which can get us spin-pure beam now and then later the exit hole, if there is a decreasing mechanism that keeps the beam back to zero (≈ 0). That is not essential, but it will mean that “ ≈ 0 ”

in the path we would have to go to about 90° to get through, and that's all the majority of the atoms, as the atoms come out and can continue to do so, without having to turn 90° to do with the spin. The whole purpose of the "improved" apparatus is to be able to do the particles in the same place, and with zero spin.

Now if we want to do an experiment like the one in Fig. 5-3, we could first make a "block" here by putting a plate in the middle of the apparatus that blocks most of the atoms, as shown in Fig. 5-4. If we now put one plate and then another second identical apparatus, it would be set like the apparatus in Fig. 5-3, verified by putting similar plates in the way of the various paths of the second S' line and seeing whether the atoms get through.

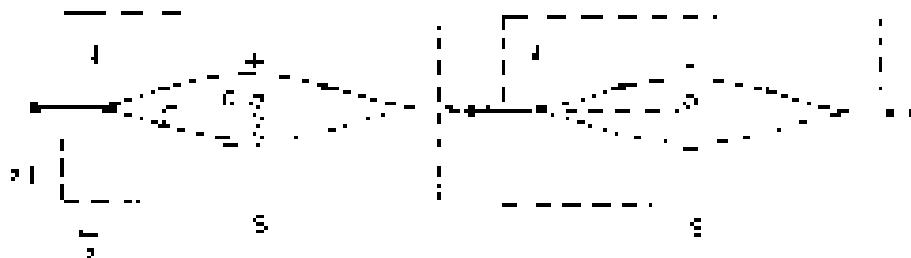


Fig. 5-4. The "improved" Stern-Gerlach apparatus and "bar."

Suppose we call the first apparatus "A" by the name of "S" & "S'"; we are going to consider all kinds of configurations, and we will need labels to keep it as straight. We will say that the atoms which take the top path in S & S' in the "S" state with respect to S' , those which take the middle path are in the "zero" state with respect to S' , and the ones which take the lowest path are in the "anti" state with respect to S' . In the more usual language we would say that the wave function of the single "component" $| \text{the Com} \rangle$ is not zero, using the language here. Now in Fig. 5-4 the second apparatus is oriented just like the first, so the lowest atoms will always be suppressed. If we had blocked off the upper and lower atoms in the first apparatus and let only the zero state through, all the fastest atoms would go through the middle path of the second apparatus. And if we had blocked off all but the lowest beam in the first there would be only a few here in the second. The conclusion is, if you have one first apparatus has produced a filtered beam in a given state with respect to S' , $| + \rangle$, $| 0 \rangle$, or $| - \rangle$, and we can tell which state is present by putting the atoms through a second filtered apparatus.

We can make our second apparatus so that it excludes only atoms with a particular state, by putting blocks where it does the job for the first one, and then allowing the others of the incoming beam just by letting whatever may come through the end. For instance, if we block off the zero beam in the second apparatus, 100 per cent of the atoms will go through, but if we block off the upper path, nothing will go through.

To make this kind of discrimination, we are going to invent a discrimination, general to any one of our improved Stern-Gerlach apparatuses. We will let the symbol

$$\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad (5-4)$$

stand for any arbitrary apparatus. (This is a symbol you will never find used in quantum mechanics, we're just inventing it for this purpose.) It is simply due to the standard form of the apparatus (cf. Fig. 5-1), since we are going to want to use several apparatuses at once, and with so many other atoms, we will readily get into a confusion under such. So the symbol in (5-4) stands for the apparatus S . When we block off one or more of the bottom paths, we will show that the same

we need here, utilizing which term is bypassed, like this:

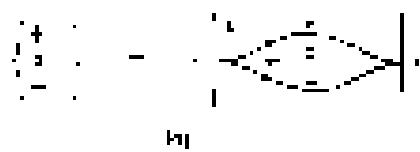
$$\begin{Bmatrix} + \\ 0 \\ - \end{Bmatrix} \quad \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad \quad (6.2)$$

The various possible combinations we can have are shown in Fig. 6-4.

If we form $\langle + | 0 | - \rangle$, there is a resonance in Fig. 6-4, we will put the two quarks next to each other, like this:

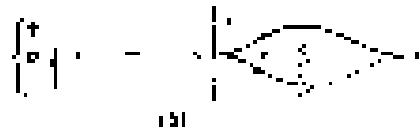
$$\begin{Bmatrix} + \\ 0 \\ - \end{Bmatrix} \quad \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad \quad (6.3)$$

But it's not everything that comes through the first quark, though the second, the "bar" part of our block of the "zero" and "minus" couplet of the second quarks, so that we have



$$\begin{Bmatrix} + \\ 0 \\ - \end{Bmatrix} \quad \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad \quad (6.4)$$

We will get 100 percent transmission through the second quark. On the other hand, if we have



$$\begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad \quad (6.5)$$

nothing else comes out or nothing else. Similarly,



$$\begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad \quad (6.6)$$

nothing goes wrong out. On the other hand,



$$\begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad \begin{Bmatrix} + \\ 0 \\ - \end{Bmatrix} \quad \quad (6.7)$$

would be just ordinary gluons

$$\begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix}$$

by itself.

Now we want to consider these couplings as a quantum mechanically. We will say that at some level $\langle + | 0 | - \rangle$ is $\langle + | S | - \rangle$ if it goes through the appearance of the gluon loop, but $\langle + | 0 | - \rangle$ is $\langle 0 | S | - \rangle$ if it goes through the red and blue $\langle + | - \rangle$ state if it goes through $\langle 0 | + \rangle$. Then we let $\langle 0 | S | - \rangle$ be the amplitude that a value which is in state $\langle + | - \rangle$ goes through a propagation $\langle + | - \rangle$ to the $\langle 0 | - \rangle$. We can say $\langle 0 | S | - \rangle$ is the amplitude for an item in the matrix to go from the state $\langle + | - \rangle$ to the state $\langle 0 | - \rangle$.

$$\langle -S | +S | - \rangle = 1.$$

* Red $\langle + | - \rangle$ = "plus S"; $\langle 0 | - \rangle$ = "zero S"; $\langle - | - \rangle$ = "minus S"

whereas to S given is

$$\langle -S_1 - S_2 \rangle = 1.$$

Similarly, the result of (S, S) is

$$\langle +S + S \rangle = 1$$

and of (S, I) is

$$\langle -S - S_I \rangle = 1.$$

As long as we deal only with "bond" states, that is, exclusively unpaired electrons, there are nine such possibilities, and we can write them in a table:

	From			
	$+S$	$0S$	$-S$	
$+S$	1	0	0	1/2
$0S$	0	1	0	1/2
$-S$	0	0	1	1/2

Summary of rate numbers: walk a matrix, you will see the parameters which have been described.

5.2 Experiments with fibrillations

Now comes the big question: What happens if the second column is tilted to a different angle, α ? And its link axis is no longer parallel to the first? It must be not only tilted, it also points in a different direction—otherwise, it would always remain off axis with respect to the original direction. To take it easy at first, just note that about the arrangement in which the second Stein-Gerlach experiment is tilted by an angle α of the y-axis, or similar in Fig. 5.8, $\alpha = 30^\circ$, and consider again Fig. 5.4. (Note that we now carry the following experiment!)

$$\begin{Bmatrix} 0 \\ 0 \\ \alpha \end{Bmatrix} \quad \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix},$$

or the experiment

$$\begin{Bmatrix} -1 \\ 0 \\ 0 \end{Bmatrix} \quad \begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix}$$

What comes out of the theory in these cases?

The answer is this: If the columns are in a definite angle with respect to x , they move in the same manner as before. In $(-1, 0, 0)$ case it is $\langle -S_1 - S_2 \rangle = 1/2$. But there is, however, a certain response to find the dominant T state. $\langle +S_1 + S_2 \rangle = 1/2$, and $\langle -S_1 + S_2 \rangle = 1/2$.

In other words, as far as theory has been able to make sure that we have the columns in a definite orientation, the fact of the matter is that, if it goes through an apparatus which is tilted at a different angle α from what it speaks, a "fibrillant"

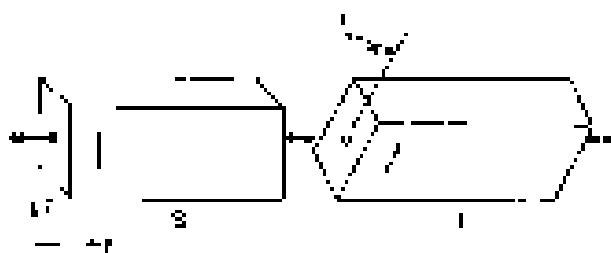


Fig. 5.6. Two-Stage Gerlach-type After passing the second slit, one of the angles α with respect to the first.

ing all—where it goes, even't forget by 'tuck'. We can put in the one particle (the ψ_{out}) at a time, and then we can only use the equation where is the probability that a particle passes through S . Since all the events that have gone through S ψ_{out} end up in a $(+2)$ state, some of them will end up in T , and some in a $(-T)$ state—all with different odds. These odds can be calculated by the absolute squares of transmission amplitudes, which we want to know in detail and method, in quantum mechanical description, the most important. When we have to know the various probabilities

$$\langle -T | -S \rangle,$$

by which we mean the amplitude that a particle initially in the $(-S)$ state can get into the $(-T)$ state (which is the same unless T and S are lined up parallel to each other). There are other amplitudes like

$$\langle +T | 0 S \rangle, \quad \text{or} \quad \langle +T | -S \rangle, \quad \text{etc.}$$

There are in fact nine such amplitudes. Consider again that a theory of success would tell us how to calculate things & it would tell us how to calculate what happens to a classical particle in the circumstance. The laws of quantum mechanics permit us to determine the amplitude that a particle will get to the right ψ particle apparatus. The central problem, then, is to be able to calculate for any given disturbance S , in ψ , that the only orientation whatever—the zero amplitude:

$$\begin{aligned} \langle +T | -S \rangle, \quad & \langle +1 | 0 S \rangle, \quad \langle -T | -S \rangle, \\ \langle 0 T | -S \rangle, \quad & \langle 0 T | 0 S \rangle, \quad \langle 0 T | +S \rangle, \\ \langle -T | -S \rangle, \quad & \langle -1 | 0 S \rangle, \quad \langle -T | +S \rangle. \end{aligned} \quad (3.4)$$

We can already figure out some relations among these amplitudes. First according to our definitions, the zero amplitudes

$$\langle +T | -S \rangle^2$$

is the probability that a particle in the $+S$ state will enter the $+T$ state. We will often find it more convenient to write such expressions in the equivalent form

$$\langle -P_1 | S | T | -S \rangle.$$

In the same notation the number

$$\langle 0 | +S | 0 \rangle + S^2$$

is the probability that a particle in the $+S$ state will enter the $(0T)$ state, and

$$\langle -T | -S | T | +S \rangle$$

is the probability that it will enter the $(-T)$ state. But in many real experiments one can't, except in principle, control in some way one of the three directions of the T particle, i.e., there's no guarantee the T given kind of atom to go, be the sum or not thus prob that its ψ will go neither must be equal to 100 percent. We have the result

$$\begin{aligned} \langle +T | -S | T | -S \rangle + \langle 0 T | +S | T | +S \rangle + \langle -T | -S | T | +S \rangle = 1 \quad (3.5) \end{aligned}$$

There are, of course, two other such equations that we get if we start with a $(+S)$ state or a $(-S)$ state. But they are all we can easily get, so we'll go on to some other general questions.

3.3 Some Gershoff filters in series

There is an interesting question. Suppose we had atoms lined up like the $(+S)$ shown, then we pass them through a second filter, say in the $(+T)$ state, and then through another $(+S)$ filter. (We'll call the last filter R just as we can during first.

After the first S filter, the probabilities α and β may were once in $\mathcal{L} \cap \mathcal{S}$ state? In other words, we have the following experiment:

$$\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \quad (5.1)$$

We want to know whether C takes the α through β also get through S . They do not. Once they have been filtered by T , they do not remain in any way that they were in $\mathcal{L} \cap \mathcal{S}$ when they entered T . Notice that the second S filter in (5.1) is oriented exactly the same as the first, so it is still an S_{left} . After the states filtered by S are, of course, still $(+S)$, $(-S)$, and $(-S)$.

The important point is that if the T filter passes nothing new, the measure that goes through the second S filter depends only on the setting of the T filter, and is completely independent of what happened to α . The fact that these measurements were done on two S_{left} + S filter has no influence whatever it was: they will do just the same if they had been rotated again into a configuration by a 90° flip-flop. From then on, the probabilities of getting into different states is the same as what had happened before S left without the T filter.

As an example, let's compare the experiment of (5.1) with the following experiment:

$$\begin{pmatrix} +1 \\ 0 \\ -1 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \quad (5.2)$$

In which only the first S is always. Let's say that the angle α between α and T is 60° , so experiment (5.1) consists of the α to β but go through T pass through S . In experiment (5.12), although there will, in general, be a different number of terms coming through T , the α wavefunction of those—one-third—will also get through S .

We can, in fact, come from where you have learned earlier—but not directly of the α wavefunction out of T and go through any particular S filter only on α and β , not α involving the suppressed excited. Let's compare experiment (5.12) with

$$\begin{pmatrix} +1 \\ 0 \\ -1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}. \quad (5.13)$$

The amplitude that an given α wavefunction of S will pass get through both T and S is, for the experiment of (5.12),

$$(-\beta) \cdot 0.75 \cdot 0.75 / \sqrt{2}.$$

The corresponding probability is

$$\langle -\beta | 0.75 \cdot 0.75 | 0.75 \rangle^2 = |(+S \cdot 0.75)^2 / \sqrt{2} |^2.$$

The probability for experiment (5.12) is

$$|0.75 \cdot 0.75 |^2 = |(+S \cdot 0.75)^2 / \sqrt{2} |^2.$$

The ratio is

$$\frac{|0.75 \cdot 0.75|^2}{|(+S \cdot 0.75)^2 / \sqrt{2}|^2}$$

and depends only on T and S , and not on the state α , beam $(-S)$, $(+S)$, or $(-S)$ is selected by S . This is because S may go up and down regardless depending on how much α is through T . Of course, this would not be the same case if we compared the probabilities that the beams would go into the ports of the cones

value or frequency of \hat{S}_z , or the value of the probability to get this or those values.

In fact, since these values depend only on which beam is allowed to pass through T_1 , and not on the selection made by the first S filter, it is when that we would get the same result even if the S apparatus were not in S filter, it was not for the electrospooner. When we will now call S one rotated by some arbitrary angle with respect to T_1 , we could find the amplitude $|(\psi_R | \psi T_1)''\rangle$; $|\psi_R | \psi T_1\rangle''$ was independent of what beam was passed by the first filter S .

5.4 These states

These can be "natural" one of the basic principles of quantum mechanics. Any atomic system can be separated by a filtering process into a certain set of what we will call "base states" and the future behavior of the atoms ... this is like saying that the eigenvalues only on the nature of the base state, it is independent of any previous history + the wave function, of course, in the time t_0 , for instance, the three states $|+\rangle, |0\rangle, |-\rangle$ are one set of base states, the three states $(+S), (0S)$ and $(-S)$ are another. Then the probability of probabilities such as $\langle S|S\rangle$ in any other.

We should be careful to say that we are considering here of filters which do nothing other than "pass" atoms. If, for instance our Stern-Gerlach apparatus fails to pass, and a good separator of the three states so that we could not separate them cleanly by our biases, then we could not make a complete separation into base states. We can tell if we have one base states by seeing whether or not the beams can be split again in another line of the same kind. If we have a pure $+S$ state, for instance, all the atoms will go through

$$\begin{pmatrix} + \\ 0 \\ - \end{pmatrix},$$

and none will pass through

$$\begin{pmatrix} + \\ 0 \\ - \end{pmatrix},$$

0 through

$$\begin{pmatrix} + \\ 0 \\ - \end{pmatrix},$$

Consider now a more complex item than it is possible to filter in such a process, so that an unfiltered, directing by an idealized apparatus is possible.

We can also assume that what we are doing is exactly what is in another laboratory situation. In any real Stern-Gerlach apparatus we would find in every atom deflection by two sides that could cause some atoms to go in a state corresponding to different angles, & about whether the second output contains atoms with $S=0$ and atoms of their initial states, and so on. We have idealized the situation so that we are passing only & sent the ones that are split in a magnetic field; we are ignoring things having to do with position, momentum, internal vibrations and the like. In general, one would need to provide also base states which are defined with respect to such things also. But to keep the concepts simple, we are considering only one set of three states, which is sufficient for the essential features of the physical situation in which the atoms don't get all mixed up in

We do not intend to use the word "base state" as only implying more than what is said here. They can also be thought of as "local" in this sense. We are using the word *base* with the thought of a base for a description, so essential in the sense that one applies of "pertaining to the base ten".

going through the apparatus, or otherwise badly treated, and tends to ψ when they leave the apparatus.

You will note that we always begin our thought experiments by starting after with only one electron system. But we start with some which does more. We do this because atoms come out of a furnace in various states due initial at random by the accidental happenings inside the furnace. It gives what is called an "apparatus" from 1.71 > on. Consider now the probabilities of the "classical" kind, as in coin tossing—which are different from the quantum mechanical probabilities we are about to discuss now. Dealing with an irregular fire you'd get a mix of different probabilities. But one better to deal with, also, we must avoid the possibility of paradoxical terms. So don't try to consider this point what happens to the first apparatus less than one atom through. (We will tell you how you can have loopholes later at the end of the chapter.)

Let's now go back and see what happens when we go from a base state for one like in a base state for a different die. Suppose we start again with

$$\left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_S \quad \left\{ \begin{array}{c} (-) \\ 0 \\ (+) \end{array} \right\}_T$$

The atoms which come out of T are, in the next cut ($t_2 T$) will have to meet very that they've come in the state $(+)$. So people would say that in the cutting by T we have "the" to inform $t_2 T$ about the previous state. So because we have "measured" the atoms when we separate them into three beams in the apparatus T . But that is not true. The past information is not lost by the separation into three beams but by the behavior of another that one beam is not lost by the following set of experiments.

We start with a 1.8 die, and will call α the number of atoms that come through $(+)$. Then follow this by $t_2 T$ (i.e., the number of atoms that come out is proportional to the input number, say $\alpha \beta$). This then can either \rightarrow either, only since fraction β of these atoms will pass on to t_2 and. We can indicate this in the following way:

$$\left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_S \xrightarrow{\alpha} \left\{ \begin{array}{c} (-) \\ 0 \\ (+) \end{array} \right\}_S \xrightarrow{\alpha \beta} \left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_T \quad (5.14)$$

If you had open this t_2 isolated a different state, say the 3.14 die, a different fraction, say γ , would go through. We would have

$$\left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_S \xrightarrow{\alpha} \left\{ \begin{array}{c} (-) \\ 0 \\ (+) \end{array} \right\}_S \xrightarrow{\alpha \beta} \left\{ \begin{array}{c} (-) \\ 0 \\ (+) \end{array} \right\}_T \xrightarrow{\gamma} \left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_T \quad (5.15)$$

Now suppose we repeat these two experiments but remove all the marks from T . We would then find the remarkable results as follows:

$$\left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_S \xrightarrow{\alpha} \left\{ \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right\}_S \xrightarrow{\alpha \beta} \left\{ \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right\}_T \xrightarrow{\gamma} \left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_T \quad (5.16)$$

$$\left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_S \xrightarrow{\alpha} \left\{ \begin{array}{c} (+) \\ 0 \\ (-) \end{array} \right\}_T \xrightarrow{\alpha \beta} \left\{ \begin{array}{c} (-) \\ 0 \\ (+) \end{array} \right\}_S \xrightarrow{\gamma} \left\{ \begin{array}{c} (-) \\ 0 \\ (+) \end{array} \right\}_T \quad (5.17)$$

[†] In terms of the earlier notation, $\alpha = (T_1 + S_1)^2, \beta = (T_1 + S_2)^2$, and $\gamma = (T_2 + S_2)^2$.

do the atoms get enough β in the first case, but never in the second case. This is one of the great laws of quantum mechanics. That nature works this way is no coincidence, but the reason we have given is somewhat far-fetched.

3-8 Interfering amplitudes

This is another thing varying from QM to QED. In quantum mechanics ψ is called *amplitude of light*. This is not very mysterious quantum mechanics—just a reference to amplitude. It's the same kind of thing as current I , or current in ohm's law experiment with electrons. We see that we can't get fewer electrons in some places than there were than we get with one slit open. It works qualitatively this way. The total wave has amplitude $\psi_1 + \psi_2 + \psi_3$ going through all three slits. $\psi = \psi_1 + \psi_2 + \psi_3$ is the upper part of (3.7); as the sum of the amplitudes, one has to add all the three results to ψ ; the sum is equal to zero:

$$(0.5 - i)(-1 + i) + 0.5 + (0.5)(0.5)(0T) + S_1 + (0.5 - iT)(T) + S_2 = 0 \quad (3.14)$$

None of the individual amplitudes is zero; for example, the absolute value of the second amplitude is $|0.5 - iT|$; the sum is zero. You would have also gotten zero if S_1 was zero, in accordance with (3.5) since, however, ψ is a superposition of (3.16), the answer is different. If we call ψ the amplitude to get through S_1 and S_2 , in this case we have†

$$\begin{aligned} \psi &= (0.5 - i)(-1 + i) + 0.5 + (-\sqrt{10})(0T) + S_2 \\ &\quad + (-1)(S_1 - iT)(T) - S_1 = 0. \end{aligned} \quad (3.15)$$

In the experiment (3.16) the beam has been split and recombined. Simply empty his beam pipe back to the source. The information about the original S_1 state is returned; it is just as though the appearance was not there at all. This is true whatever is put after the "value added" T apparatus. We could follow it up, say, $S_1 \rightarrow T \rightarrow S_2 \rightarrow T' \rightarrow S_3 \rightarrow T'' \rightarrow S_4 \rightarrow T''' \rightarrow S_5 \rightarrow T'''' \rightarrow S_6 \rightarrow T''''' \rightarrow S_7 \rightarrow T'''''' \rightarrow S_8 \rightarrow T''''''' \rightarrow S_9 \rightarrow T'''''''' \rightarrow S_{10}$. The answer will always be the same as it was previously, taken directly from the first S_1 filter.

So this is the important principle: A T does not change any amplitude, unless it produces some change at all. We should take one additional condition. The wave operator $\hat{\psi}$ acts on ψ to produce ψ' ; this leaves, however, state ψ unchanged unless $\hat{\psi}$ is zero. This means that ψ and ψ' have exactly the same amplitude. This is called *unitarity* of the process. The reason is that even if the extra measurement were still to tell something through the T 's, it could change the values of some of the amplitudes. Then the amplitudes would be changed, and the amplitudes in Eqs. (3.14) and (3.15) would be different. We will always assume that there are no such changes of amplitudes.

Let's rewrite Eqs. (3.14) and (3.15) in an improved notation. We will be interested in one of the three states $|T\rangle$: (3.7), (3.16). It is often convenient to normalize:

$$\sum_{\alpha, \beta} (0.5 - iT, \alpha|\beta) = 0 \quad (3.20)$$

and

$$\sum_{\alpha, \beta} (0.5 - iT, \alpha|\beta) = \delta_{\alpha \beta} - 1. \quad (3.21)$$

Similarly, for an experiment where Δ is replaced by a complete unitary filter R , we have

$$\begin{pmatrix} 0.5 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad (3.22)$$

† We really cannot conclude from the requirement that $\psi = \psi'$, but only that $|\psi|^2 = |\psi'|^2$, or $\psi^\dagger \psi = \psi'^\dagger \psi'$, but it can be shown that the choice $\psi = \psi'$ between two real ψ is generic.

The results will always be the same as if the \hat{S} operator were left out and we had only

$$\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}.$$

Or, expressed mathematically,

$$\sum_{B \in \mathcal{B}} \psi_B \hat{B} |B\rangle = (\hat{B} + \hat{S})\psi. \quad (5.23)$$

This is our fundamental law and it is guaranteed by no astrophysics islands for the three base states of any filter.

You will notice that in the experiment (5.22) there is no explicit mention of S and \hat{B} to \hat{B}' . Furthermore, the arguments would be the same no matter what states they selected. To write the equation in a general way, we have to bring in \hat{B}' . In the specific states referred to S and \hat{B} , let's call ψ ("pre") the state prepared by the first filter (in our special example, $|S\rangle$) and ψ' ("post") the state tested by the last filter (in our example, $|\psi'\rangle$). Then we can state our fundamental law of Eq. (5.23) in the form

$$(\hat{B}' + \hat{S})\psi = \sum_{B \in \mathcal{B}} \langle B' | \hat{B} | \psi \rangle |B\rangle. \quad (5.24)$$

where $\langle B' |$ is to average over the three base states of some post-filter.

We want to emphasize again what we mean by base states. They are like our three states which can be selected by one of our Stern-Gerlach apparatuses. One requirement is that if you have a base state ψ , then the future is "independent" of past. Another condition is that you have a complete set of base states. Eq. (5.24) is true for any set of beginning and ending states ψ and ψ' . There is, however, no unique set of base states. We began by considering base states and referred to a particular apparatus \hat{B}' . We could equally well consider a different set of base states with respect to an apparatus S , or with respect to \hat{B} , etc. We usually speak of the base states "as a certain representation."

Another sentence on a point we've made in any particular represents is as follows. They are all completely different. By this we mean that if we take a $(+Y)$ state, there is no amplitude for it to go into a $(0T)$ or a $(-T)$ state. If we let ψ and ψ' stand for any two base states of a particular set, the general rules obtained in connection with (5.8) are that

$$\langle \psi | \psi' = 0$$

for all ψ and ψ' that satisfy (5.24) . Of course we know that

$$\langle \psi | \psi = 1$$

These two equations are usually written as

$$\langle \psi | \psi' = \delta_{\psi\psi'} \quad (5.25)$$

where $\delta_{\psi\psi'}$ (the "Kronecker delta") is a symbol that is defined to be one for $\psi = \psi'$ and to be zero for $\psi \neq \psi'$.

Equation (5.25) is not independent of the other laws we have mentioned. It happens that we are not particularly interested in the mathematical problem of finding the minimum set of independent elements that will give all the necessary quantities.⁴ We are satisfied if we have a set that is complete and not apparently inconsistent. We can, however, show that Eqs. (5.23) and (5.24) are not independent. Suppose we let $\psi = \psi'$. Eq. (5.24) represents one of the base states of the

⁴ In fact, for a given system with three or more base states, there are other laws of Nature—other than a Stern-Gerlach apparatus—which can be used to get more complete information about the state (such as the wave theory of atoms).

⁵ Reference much greater than ref.

so we set $\alpha = \omega$, $\beta = 1$, taking from ω here:

$$\langle Y | A | \psi \rangle = \sum_{\lambda} (\lambda^{\frac{1}{2}} \phi)_{\lambda} \langle \lambda | A | \psi \rangle.$$

But Eq. (5.22) says that $\lambda^{\frac{1}{2}} \phi$ is zero unless $\lambda = \omega$ or ω^{-1} because $\omega \neq \lambda$ and we have an identity which shows that the two laws are not independent.

We can see now there would be another solution among the superpositions of Eqs. (5.20) and (5.24) and this, Topoform (v. 3) is

$$\langle Y | T + S | \psi \rangle = \langle Y | T - S | \psi \rangle, \quad T = S^2 = (-C + iB), \quad C = S^2 = 1.$$

Now, with $\lambda \neq \omega$ (5.24), letting ω be ϕ and χ be the same $(-\omega)$ we find that one side $\langle Y | T + S | \psi \rangle$, which is clearly $= 0$, we get from macro Eq. (5.19),

$$(-Y^2 + Y)(|T - S|) = (-Y)(0.75(0.7) + 0) = (-Y^2 + Y)(-Y) = 0.$$

These two occurrences are consistent (for all the elements $\lambda \neq \omega$) if ψ and ϕ is unpolarized only if

$$\langle Y | S | \psi \rangle = \langle Y | T | \psi \rangle,$$

$$\langle Y | S - T | \psi \rangle = \langle Y | T - S | \psi \rangle,$$

$$\langle Y | S^2 | \psi \rangle = \langle Y | T^2 | \psi \rangle.$$

And it follows now for any states α and β ,

$$\langle Y | \alpha \rangle = \langle Y | \beta \rangle^{\frac{1}{2}} \quad (5.30)$$

If this were not true, probability wouldn't be "conserved" and we would get "lost."

Before going on, we want to summarize the theory in an unphysical case also simplified. They are Eqs. (5.24), (5.25) and (5.26):

$$\langle Y | \beta | \psi \rangle = \delta_{\beta \omega},$$

$$\langle Y | \alpha | \psi \rangle = \sum_{\lambda} (\lambda^{\frac{1}{2}} \phi)_{\lambda} \langle \lambda | \psi \rangle, \quad (5.27)$$

$$\langle Y | \alpha | \alpha \rangle = \delta_{\alpha} |\alpha|^2.$$

In these equations, α and β refer to different states of some other representation, while ω is represented by positive states of the one ... It is important to note that this is valid only if the state is carried from one of all the basis states of the system (in our case, three): $|T\rangle$, $|T - S\rangle$, $|S\rangle$. This tells us nothing about what we should choose to represent for our set of own states. We began by taking $|T\rangle$ up to ω , and then $|S\rangle$ and $|T - S\rangle$ each associated with some arbitrary element in any other representation, say, $|W\rangle$, and we begin to guess. We could have a different set of states α and β already, but α and β 's would then be 0 ... There's no contradiction. One of the great virtues of quantum mechanics is its "uniqueness"! In fact the things can be calculated in more than one way.

5.4 The machinery of quantum mechanics

We will illustrate exactly why these laws are useful. Suppose we have an atom in a given condition (say, which we mean that it **was** prepared in a certain way), and we want to know what **else** we can do in some experiment. In other words, what **other** basis can it be put in? We must know what **else** the atom will do. That is to say, that we can describe the apparatus completely in terms of these complex numbers $\langle \lambda | \psi \rangle$, the amplitude for state λ to be in the condition ψ ; and that we can tell what **else** happens in atoms put into the apparatus if we describe the state of the atom by giving three numbers $\langle \lambda | \psi \rangle$ at the simple descriptive level in my original conditions to be found in each of the two ways going. The last important idea

Let's consider first an illustration. Think of the world as a perfect Western without any opportunity for us to have a random walk across it, where we can only move left or right. R appears— $\langle \psi | R | \psi \rangle$ —

$$\begin{Bmatrix} 1 \\ 0 \\ -1 \end{Bmatrix} \quad \begin{Bmatrix} 1 \\ 0 \\ -1 \end{Bmatrix} \quad \begin{Bmatrix} +1 \\ 0 \\ -1 \end{Bmatrix}, \quad (1.23)$$

Because the only complicated arrangement of four Gershleit components will happen at $x = 0$, this is a test at position $x = 0$, with odd components being $\langle \psi | R | \psi \rangle$. This is nothing but anything you want to put in $\langle \psi | R | \psi \rangle$ since the experiments you don't have to go to all the way to $x = 0$ to fully measure the apparatus. The question would be: What would a pulse do? It means that we see the situation $\langle \psi | R | \psi \rangle$ plus movement of A in the $\langle \psi | R | \psi \rangle$ state so that it changes the $\langle \psi | R | \psi \rangle$. The $|A\rangle$ has a negative norm because it is amplitude; it is

$$\langle 0.5 | A = \delta_0.$$

So that it is to be read from $\langle \psi | R | \psi \rangle$ that this theory

$$\langle \text{final} | \text{the } \alpha_i | \text{start} \rangle$$

of the shape A does not change but is just an open channel. Then we write

$$\langle 0.5 | A = \delta_0 = \langle 0.5 | \cdots \delta_0. \quad (1.24)$$

and two numbers are equivalent. For a more general problem, we might replace $|A\rangle$ by a generic wave function α and $\langle \psi | R | \psi \rangle$ would depend on what we would want to know the amplitude:

$$\langle \psi | A | \phi \rangle.$$

An immediate analysis of the system is to want to give the amplitude $\langle \psi | A | \phi \rangle$ for every possible pair of states ψ and ϕ in infinite number of combinations! How to do it? We give a simple description of the sequence of the steps in 1.3. We do this in the following way. Imagine the bare ψ and ϕ (1.23) is modified by the

$$\begin{Bmatrix} + \\ 0 \\ - \end{Bmatrix} \quad \begin{Bmatrix} 1 \\ 0 \\ -1 \end{Bmatrix} \quad \begin{Bmatrix} 0 \\ 1 \\ 0 \end{Bmatrix} \quad \begin{Bmatrix} 1 \\ 0 \\ -1 \end{Bmatrix} \quad \begin{Bmatrix} -1 \\ 0 \\ 1 \end{Bmatrix}. \quad (1.25)$$

This is really no modification at all since the wide-open T -approximation don't do anything. But this does suggest how we can analyze the problem. There is a certain set of amplitudes α such that the sum of all $\langle \psi | \alpha | \phi \rangle$ times the value of T . Then there is another set of amplitudes that is constant with respect to the moving A or α and that is just with respect to ψ . And finally, the eigenamplitude that each ψ has with get through. In fact, this is a 1.20 system. For each possible alternative path, there is an amplitude of the form

$$\langle 0.5 | \alpha_i | A = \langle 0.5 | \alpha_i | 1.20,$$

where α is a complex linear combination of the basis vectors going with all possible combinations of A and ϕ . The amplitude we want is

$$\sum_i \langle 0.5 | \alpha_i | A = \langle 0.5 | \alpha | A = \langle 0.5 | \phi. \quad (1.26)$$

If $\langle 0.5 | \alpha_i | A = \langle 0.5 | \alpha | A = \langle 0.5 | \phi$ is replaced by several states α and ϕ , we get the same thing. α is a solution; so α is the general result

$$\langle 0.5 | \alpha | A = \sum_i \langle 0.5 | \alpha_i | A | 1.20 | \phi. \quad (1.27)$$

Now notice that the right-hand side of Eq. (5.32) is called "single" from the left-hand side. The apparatus A is completely described by the two numbers $\langle A | \psi \rangle$ which tell the response of A with respect to the three base states of the apparatus A . Once we know these nine numbers, we can immediately form many other states of A , and we define each of them as the sum of the various amplitudes for going into, or "from," each of the three base states. The result of an experiment is predicted using Eq. (5.32).

It is then in the calculation of quantum mechanics for a spin-one particle. Every state is described by three numbers which are the amplitudes to be in each of some selected set of base states. Every apparatus is described by nine numbers which are the amplitudes to go from one base state to another in the apparatus. From these numbers anything can be calculated.

The nine amplitudes which describe the apparatus are often written as a square matrix, called the matrix $\langle A | A' \rangle$:

$$\begin{aligned} & \text{from} \\ & \quad \begin{array}{c} 0 \\ + \\ - \end{array} \quad \begin{array}{c} 0 \\ + \\ - \end{array} \\ \langle A | & = \begin{pmatrix} \langle A | + \rangle & \langle A | 0 \rangle & \langle A | - \rangle \\ \langle + | A' \rangle & \langle 0 | A' \rangle & \langle - | A' \rangle \end{pmatrix} \quad (5.33) \\ \langle 0 | & \\ \langle - | & \end{aligned}$$

The mechanics of quantum mechanics is just an extension of this idea. We will give you an illustration. Suppose we have an apparatus C that we want to analyze—let us, we want to calculate the matrix $\langle A | C | B \rangle$. For instance, we might want to know what happens in an experiment like

$$\begin{array}{c} \left| \begin{array}{c} + \\ 0 \\ - \end{array} \right\rangle \quad \left| \begin{array}{c} + \\ 0 \\ - \end{array} \right\rangle \quad \left| \begin{array}{c} + \\ 0 \\ - \end{array} \right\rangle \\ \text{A} \quad \text{B} \quad \text{C} \end{array} \quad (5.34)$$

But then we remember the jargon of two pieces of apparatus A and B in series: the particles go through A and then through B —so we can write immediately

$$\left| \begin{array}{c} + \\ 0 \\ - \end{array} \right\rangle = \left| \begin{array}{c} + \\ 0 \\ - \end{array} \right\rangle \left| \begin{array}{c} + \\ 0 \\ - \end{array} \right\rangle. \quad (5.35)$$

We can use the Chapman's law "product" of A and B . Suppose that they already know how to analyze the two parts, so we can get the matrices (with respect to $| + \rangle$ of $| 0 \rangle$ and $| - \rangle$) for both A and B . Our problem is then solved! We can easily find

$$\langle A | C | B \rangle$$

for any input and output states. Just we write this:

$$\langle A | C | B \rangle = \sum_k \langle A | B | k \rangle \langle k | C | 0 \rangle.$$

Do you see why? (Why?) Imagine putting a Chapman between A and B . Then if we consider the special case in which ϕ and ψ are also base states $| + \rangle$, $| 0 \rangle$, and $| - \rangle$, we have

$$\langle A | C | B \rangle = \sum_k \langle A | B | k \rangle \langle k | C | 0 \rangle. \quad (5.36)$$

This equation gives the matrix for the "product" apparatus C in terms of the two matrices of the apparatuses A , B , C . Mathematicians call this new matrix $\langle C | C | 0 \rangle$ —formed from two matrices $\langle A | A | 0 \rangle$ and $\langle B | B | 0 \rangle$ according to the rule specified in Eq. (5.36)—the "product" matrix $\langle A | C | B \rangle$ of the two matrices A and B . Note that the order is important, $A B \neq B A$. Thus, we can say that the matrix for a "product" of two pieces of apparatus is the matrix product of the matrices for the two apparatuses (using the *first* apparatus in the right in the product). Another which you may already have understood that we used in Eq. (5.36)

5.7 Transforming to a different basis

We want to use one basis (e.g., the basis it's easiest to calculate),
 Suppose we're given basis $\{|\psi\rangle\}$ with some particular basis $\{|\phi_i\rangle\}$ of the S basis, and
 another basis $\{|\chi_j\rangle\}$ to do the calculations with a different basis $\{|\theta_j\rangle\}$ of the T basis. To keep things straight let's call our base states the (S) states, where
 $|\psi\rangle = |\phi_1\rangle + \dots + |\phi_n\rangle$. Similarly, we can call our basis states (T) . Then we can compute
 the wave function $|\psi\rangle$ in terms of the result of any measurement that all
 occurs in the same basis, but in the meantime the various amplitudes and averages
 will be different. How are they related? For instance, if we calculate with
 the same $|\phi_i\rangle$, we will measure the total probabilities $|\psi\rangle^*|\psi\rangle$ have
 goes the same basis in the S representation, whereas we will find the $|\psi\rangle^*|\psi\rangle$
 amplitude $|\psi\rangle^*|\psi\rangle$ in the T representation goes back to its S representation.
 However, we check the wavefunction both containing the same state $|\phi_i\rangle$. So, in order to
 consider the general rule let's do it. Keeping the outcome of the states $|\phi_i\rangle$ we have

$$|\psi\rangle = \sum_i (S_i | \phi_i \rangle) \quad (5.8)$$

To relate the two representations, $|\psi\rangle$ need only give the right complex numbers of
 the state $|\psi\rangle$ ($|S\rangle$). This means we have to convert $|\phi_i\rangle$ to $|\psi\rangle$ by some linear
 transformation. It tells us how to map from one set of basis states to another.
 (This map between $|\psi\rangle$ and $|\phi_i\rangle$ is sometimes called "the transform". You may also
 see references to "representation 1," "Rep 1", etc.)

In the case of spin-1/2 particles for which we have only three base states
 (the higher spin states are not in the mathematical situation we're looking at), what
 we have seen in vector algebra, there needs to be represented by complex numbers.
 For example, this occupies two x, y, z axes. The x -axis would not be necessary.
 The three "Pauli" vectors with two vectors along the x -axis. But
 suppose someone chose to use a 3 -dimensional set of axes, x', y', z' . We will
 be using different numbers to represent any given orientation. The calculations will
 look different, but the final result will be the same. We have to find the rule
 and know the rules for transforming vectors from one set of axes to another.

You'll be asked to see how no quantum mechanical calculation can work by
 ignoring terms out; so we'll give this, without proof, the transformation rule we
 be deriving the equations might also in one step transition S to some other representation
 T , i.e., it has special reference to certain sets of basis $|\phi_i\rangle$ and $|\theta_j\rangle$. We
 will show you how to change basis to derive other transformation rules.

For ease, let's suppose x has the same x, y, z axes (along which the particle moves in the S apparatus, but is not about the common axis by the angle α) as in Fig. 5.1(a). To be specific, a set of coordinates x, y, z is fixed in the T apparatus, relative to the x, y, z coordinates of the S apparatus by $x' = x \cos \alpha$,
 $y' = x \sin \alpha - z \sin \alpha$, $z' = z$. Then, taking components directly we

$$\begin{aligned} |\psi\rangle &= S_i |\phi_i\rangle = S_i (1 - \cos \alpha) \\ (S_i |\phi_i\rangle)^* S_i &= -\frac{1}{\sqrt{2}} \sin \alpha \\ (-S_i - \phi_i) &= \frac{1}{\sqrt{2}} \sin \alpha \\ (-S_i - \phi_i)^* S_i &= -\frac{1}{\sqrt{2}} \sin \alpha, \quad (5.9) \\ (-S_i - \phi_i)^* S_i &= -\frac{1}{\sqrt{2}} \sin \alpha, \\ (-S_i - \phi_i) &= \frac{1}{\sqrt{2}} (-1 - \cos \alpha), \end{aligned}$$

Second Case. The T -representation has some advantages, but is centered around the x_3 axis by the angle θ . (The coordinate transformation is $x' = x_1$, $x'' = x_2 \cos \theta + y \sin \theta$, $y' = y \cos \theta - x_2 \sin \theta$.) On the x_3 -invariant amplitude we have

$$\begin{aligned} (-T)^\alpha \cdot S &= e^{\pi i \alpha}, \\ (-T)^\alpha \cdot S_1 &= 1, \\ (-T)^\alpha \cdot S_2 &= e^{\pi i \alpha}, \\ \text{and also } &\epsilon = 0. \end{aligned} \tag{5.39}$$

Note that any addition of T will increase the mod. up of the total scattering cross section.

We can also define by the three numbers

$$C_1 = (+Y | S), \quad C_2 = (0 Y | S), \quad C_3 = (-Y | S), \tag{5.40}$$

and the same one is obtained from the product of S by the three numbers

$$C'_1 = (+Y | S), \quad C'_2 = (0 Y | S), \quad C'_3 = (-Y | S). \tag{5.41}$$

From the coefficients $C_j T^\alpha | S\rangle$ of T we see in (5.39) that the transformation connecting S and T . In order words, the C_j are very much like the components of a vector, that is, quite different from the pair $(+, -)$ of S and T .

For a n -particle form factor—measured, it requires n n -amplitude— n T requirements with a vector is very easy. In each case, there are n constants C_j to be measured, and each constant changes—a simultaneous way. In fact, there is a set of bases in which transformation just like the three components of a vector. This number is

$$C_j = \frac{1}{\sqrt{2}} (S'_j - S_j), \quad C'_j = \pm \frac{i}{\sqrt{2}} (C_1 + C_3), \quad S_j = C_1 + C_3. \tag{5.42}$$

Transforming to C'_j and S is just a way from S'_j , although both S'_j and C'_j [26], can show that this is so by using the n -amplitude. From (5.38) and (5.41) you may see why a n -particle is often called a "vector particle".

5-6 Other situations

We begin by defining what our discussion of spin means. This would be a problem for any quantum mechanical problem. The general rule is to only work with representations of $SO(3)$. Instead of only three base states, any particle rotation can involve n base states. One has to do it $SO(n)$ [27] to measure the same thing, but it is more understanding than a really more range over n or $n+1$ states. Any transformation can be simplified by giving the amplitude that it starts at i and one of the base states. Then it is very clear one of the wavefunctions and then summing over the remaining set of base states. Any n -particle n -base states can transform, and if you want to use a different set, this just as good: the two can be connected by some sort of a transformation relation. We will have more to say later about such questions.

Finally, we must make a remark on what to do when we directly come to δ states or three-particle amplitudes, say ϵ_1 , and are then forced to use ϵ_1 , which reflects the state ϵ . You do not know what the state ϵ is then, by definition. It is perhaps to know. You don't exactly expect this problem just yet, but it appears sooner or problems with n , $n+1$, $n+2$, etc., n -particle states. But if you insist, here's how the problem can be handled.

First, you have to try to make some reasonable guess about the way the states ϵ is distributed to the ϵ_1 's. One comes from the ϵ function. For example, if

[†] The number of base states may be and usually is infinite.

and we’re making “prior” about the states, you might reasonably guess that more would have the business and random “distributions.” (Remember we already saw corresponds to saying that you don’t know anything about the states, but that one state has in the $\{+1\}$ state, and think one in the $\{-1\}$ state, and another one in the $\{0\}$ state.) For those cases, it’s $\Pr(\text{+1}) = \Pr(\text{-1}) = \Pr(0) = \frac{1}{3}$, and similarly for the others. The overall probability is then

$$\Pr(\text{S} = -2) = \frac{1}{3} + \Pr(\text{S} = 0) = \frac{1}{3} + \Pr(\text{S} = +2)$$

Why did we use $\Pr(\text{S} = 0)$? Well, why? This is the symmetric distribution, the same as the symmetric prior we chose for our initial distribution. As long as we are dealing with a uniformly distributed random variable, it comes out in the same way! (See)

$$\sum_{x \in S} \Pr(S=x) = \sum_{x \in S} \Pr(x|S)$$

In fact, $\Pr(\text{S} = 0)$ (We leave it to you to prove.)

But! You’re not correct, since that the right sum over the sample values of S takes $\Pr(\text{S}=0)$, $\Pr(\text{S}=1)$ to be in $\Pr(\text{S})$ and $\Pr(\text{S}=2)$ to be in $\Pr(\text{S})$ that would imply that each of the relevant might be present. It is simple that we do not care where the individual states; you have to think of terms of the probability. But the solution turns out in the various probabilities initial states, and then you have to have weighted the age over the current possibilities.

Spin Transformations

6-1 Transforming amplitudes

In the last chapter, using a system of spin, this was an example of a multi-dimensional generalization of quantum mechanics.

Any state ψ can be described in terms of a set of base states by writing the amplitudes to be at each of the base states:

The amplitude of finding a state ϕ in a system can, in general, be written as a sum of products, each product being the amplitude to be in a one of the base states times the amplitude to go from that base state to the final condition ψ , i.e. the sum including a term for each base state:

$$\langle \psi | \phi \rangle = \sum_i \langle \psi | \phi_i \rangle c_i \quad (6.1)$$

The base states are orthogonal. The amplitudes to be in one of them are all the same in value:

$$\langle \psi | \phi_i \rangle = \delta_{ij} \quad (6.2)$$

The amplitude to get from one state to another correctly is the complex conjugate of the reverse:

$$\langle \phi | \psi^* \rangle = \langle \phi | \psi \rangle \quad (6.3)$$

We take it as given in the bit above the fact that there can be many different ways to do this and that we can use Eq. (6.1) to convert from one basis to another. Suppose, for one spin, that we have the amplitudes $\langle \psi | \phi_i \rangle$ where the state ψ is in every one of the bases, ϕ_i , of a base system S , but that we'd like to decide that we would prefer to describe the state ψ in terms of a different set of base states, say the states of the system we call T . In this particular form to do this, we could substitute $\langle T | \phi_i \rangle$ and obtain the formula:

$$\langle T | \psi \rangle = \sum_i \langle T | \phi_i \rangle c_i \quad (6.4)$$

The amplitudes for the state ψ to be in the base states $\langle T | \phi_i \rangle$ are related to the amplitudes to be in the base states $\langle S | \phi_i \rangle$ by the set of coefficients $\langle T | \phi_i \rangle / \langle S | \phi_i \rangle$. If there are M base states, there are M^2 such coefficients. Since a set of base states is often called the "representative manifold" to get from the S -representation to the T -representation, this looks rather complicated; mathematically, one with M^2 lines regarding the conversion of it is really not so bad. If we call C_i the amplitude that the state ψ is in the base state ϕ_i and if $C_i = \langle S | \phi_i \rangle$ and $c_i = \langle T | \phi_i \rangle$ the conversion map amplitudes for the base system T that $\langle T | \psi \rangle = \langle T | \phi_i \rangle c_i$, then C_i to c_i can be written as

$$C_i = \sum_j R_{ij} c_j \quad (6.5)$$

where R_{ij} is the same thing as $\langle T | \phi_i \rangle / \langle S | \phi_j \rangle$. Each amplitude C_i is equal to c_i .

* This chapter is interesting and important, but, I do not introduce any QM which we will not also come up to a direct or more in the chapter. You can, therefore, skip it, and come back later if you are interested.

6-1 Transforming amplitudes

6-3 Transforming to a rotated coordinate system

6-4 Rotations about the y -axis

6-5 Rotations about x

6-6 Arbitrary rotations

over all of each of the coefficients R_{ij} , this would simply do it. It has the same form as the most familiar of series: Fourier expansion of periodic functions.

In order to avoid being too abstract for now, let's we have given you some examples of these coefficients for the spin-one case, so you can see how to get them in practice. On the other hand, I'm not very good at writing mathematical nomenclature—therefrom the short fact that there are three a 's and four b 's—the symmetry properties of space are due to this, three coefficients can be found plus b_0 by direct reasoning. Showing you such arguments at this early stage has a disadvantage in that you are immersed in rather abstract algebra here before we get "down to earth," however, the time's not far off, then we're going to do it in practice.

We will show you in the chapter how the transition amplitude coefficients can be derived for spin-one-half particles. We pick this case, rather than spin-one, because it is so much easier. Our problem is to determine the coefficients R_{ij} for a particle—an atomic system—which is split into two terms in a Stern-Gerlach apparatus. We are going to derive all the coefficients for this transformation from the previous chapter by pure reasoning—plus a few auxiliary steps. Now this implies some work reasoning in order to use "pure" reasoning. Although the arguments will be abstract and somewhat involved, the results will not be difficult to state and easy to understand—and the results is the most important thing. You may, if you wish, consider this as a generalization. We have, in fact, arranged that all the essential results will also be derived in another way when they are relevant to later chapters, so you need not be afraid of losing the thread of our story of quantum mechanics if you skip this chapter entirely, or study it at some later time. The exercise is "further" in the sense that it is intended to show how the principles of quantum mechanics cannot only be tested, but also deepened by using only theoretical methods without the expense of experiment; we can deduce a great many properties of physical systems. Also, it is important that we know what the basic experiences of quantum mechanics are from the moment living as our laws of physics are incomplete—as we know they are. It is interesting to find out where the places where our theories fail to agree with experiment, where our logic is not best, where our logic is between the two, how it agrees with these or, logic's the most extreme it always gives correct results—it agrees with experiment. Only when we try to make specific predictions of the quantum mechanics of the fine-structure particles and their interactions are we enabled to find the theory failing against each experiment. The theory becomes useful then, to describe aspects which experiment shows it has been tested. For the strange particles as well as for electrons, you may find it so.

The remark just concerning our interesting point before we started, it is not possible to determine the coefficients R_{ij} uniquely, because there is always some arbitrariness in the calculation of amplitudes. If you have a set of amplitudes of any kind say the amplitude to enter at some place by a particle but at different times, and if you multiply every single amplitude by the same phase factor say by $e^{i\theta}$, you have another as this is just as true. So, it is always possible to make an arbitrary change in phase of some amplitudes in any given solution if you want to.

Suppose you calculate some probability by writing down various amplitudes, say $A_1, A_2, A_3, \dots, A_n$, and taking the absolute value. Then somebody else calculates the same thing by using the sum of the amplitudes, say $|A'_1 + A'_2 + A'_3 + \dots + A'_n|$ and taking the absolute value. If $A'_1, A'_2, A'_3, \dots, A'_n$ are input to the $A_1, A_2, A_3, \dots, A_n$, say, for a factor λ^2 , all probabilities determined by using the absolute square of the exactly the same, since $(A'_1 + A'_2 + \dots + A'_n)^2 = A'_1^2 + A'_2^2 + A'_3^2 + \dots + A'_n^2$. Or suppose for instance that we were trying to compute something with F_1, F_2, \dots, F_n , but then we suddenly change all of the phases of a certain few systems. Every one of the amplitudes (A_i 's) would be multiplied by the same factor $e^{i\theta}$. Similarly, the amplitudes $|A'_1 + A'_2 + \dots + A'_n|$ would also be changed by $e^{i\theta}$, but the amplitudes $|A'_1|^2, |A'_2|^2, \dots, |A'_n|^2$ are the squares of A'_i 's, therefore the former gets changed by the factor $e^{i2\theta}$. See pgs 202 and 203.

in the experiments carried out, and we would have the same outcome we had before, or the general rule that it is claimed all the experiments with respect to a given basic question have the place—*as well as the change*—of the application in one problem by another phase, it makes no difference. There is, therefore, no need to do this about the problem under consideration now. Every now and then we will make such minor changes—mainly following the indications that we can get from our experiments.

4-2 Transforming the related coordinate system.

We consider again the "improved" Stern-Gerlach apparatus which is described in the 1st chapter. A beam of spin $\sigma = \frac{1}{2}$ particles entering at the left, would, in general, be split in two beams, as shown schematically in Fig. 5-1. (These two rays become the spin rays.) As before, the beam on the left is magnetized upwards and the effect of this is screened off by a "stop" which intersects the beam with its "every part". In the figure we show the beam which comes in the direction of the increase of the magnitude of the field, say from the magnetic pole to the next pole. This is the path which represents the *exact* ray path of a spin $\sigma = \frac{1}{2}$. It is fixed now due to the "stop" and will never be violated. The other beam, on the right, has the same "apparatus length". We also see that the direction of the magnetic field in each magnet is always the same with respect to the axis.

We will say that those paths which go in the "upper" sector, in the (+) side must always be spin up-particle paths and those in the "lower" sector are in the (-) side. (There is no "upper" stop for one half sector.)

Now suppose we put two of our modified Stern-Gerlach apparatuses in series, as shown in Fig. 5-2(a). The first one which we can compare to those in Fig. 1-2, has a path (= trajectory) starting one beam at the left. (As shown in *Appendix 1*, p. 3) stated. For each end, here, there is a displacement of a particle that comes out of S_1 's magnet after the (+) and (-) beam of the two half-apparatus. However, in the just for completeness the total trajectory from $(+)$ to $(-)$, from $(-)$ to $(+)$, from $(+)$ to $(-)$, and $(-)$ to $(+)$ (see Fig. 5-2), there are no before and after's of making them sequentially, since from the *Stern-Gerlach*, in the *representation*. We can consider the first apparatus "precedes" the second one in the representation and that the second one is "followed" but still in terms of the *exact* ray path relation. The kind of question we might propose, then, is how it is that this has happened in a given condition, say the C.I.S state, by that means of the beams in the apparatus? What is the chance that it will happen in a given apparatus? If this question does not exist, then there will always, of course, be no angle between the two systems S_1 and S_2 .

We should expect $\theta_{S_1 S_2} = 0^\circ$, in fact, we could have any hope of finding the inefficient $\theta_{S_1 S_2}$ by calculation. You know we have no reason to believe that S_1 is not free to spin around to some θ direction that it has in some chance of meeting the S_2 the particle with its spin pointing in the $-z$ direction, in any *state* whatsoever. So that it is almost impossible, we may guess. It is probably impossible. Then the answer to this can be *zero*, and that is the answer to the *problem* you confront us uniquely.

The last kind of experiment we can make is this. Suppose we have a setup like the one in Fig. 5-2(b), in which we have the two apparatuses S_1 and S_2 , with T acting in the angle $\theta_{S_1 T}$, with respect to S_1 , and T also only the (+) beam for going up and the (-) beam down T . We want to know, for certain number of the possibilities that the particle coming out of S_1 passing through T , has a possibility to make measurement with the *coaxial* of F_1 in S_2 by the reference orientation of S_1 and T is the same but the whole system occurs different angle in series. We want to know, however, if these experiments give the same number for the two spin-phases possible in a given state with respect to S_1 as a result of the particle as state with respect to T . We are seeking, in other words, that the result of the *experiment* of this type is *constant*, that is, F_1 in S_2 come out to the

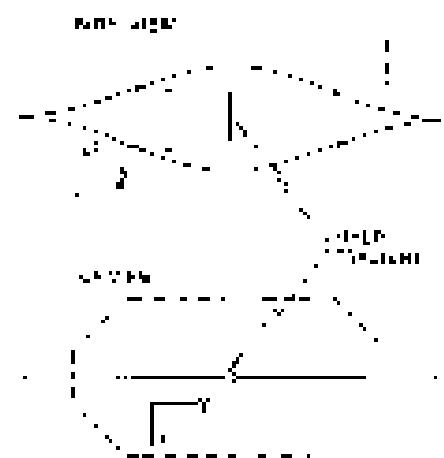


Fig. 5-1. Top and side view of an "improved" Stern-Gerlach apparatus with two magnets.

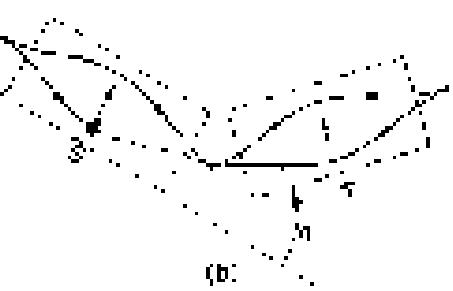
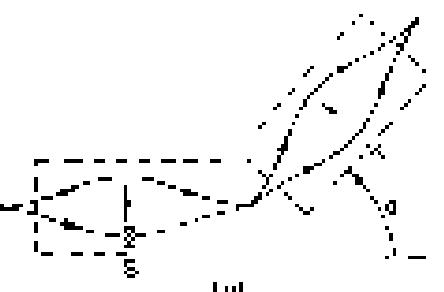


Fig. 5-2. Two associated experiments.

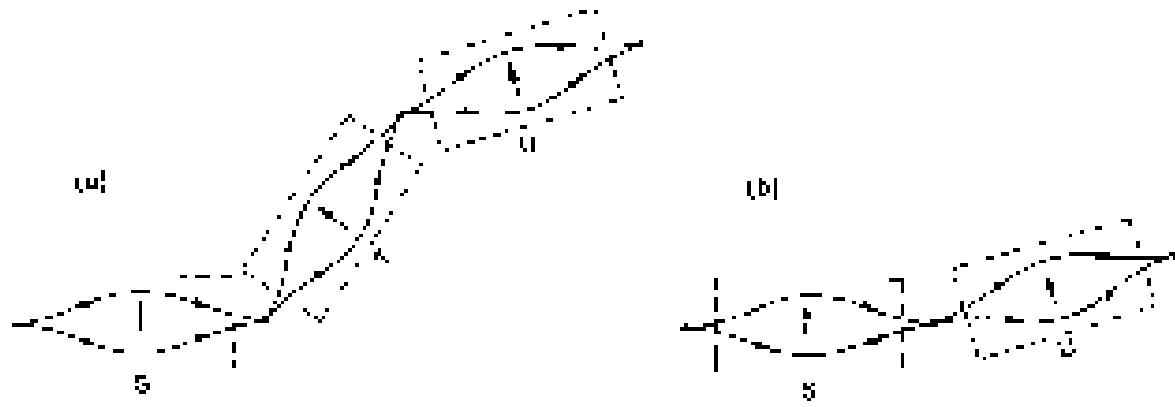


Fig. 4.5. (a) The "wide open"; (b) it is "open to jet".

Now the word "open" is important in space. You say, "That's obvious." But it is also important that it is "right" only if it is really what happens. That means that the coefficients R_{ij} depend only on the orientation in space of S and T , and not on the absolute situation of S and T . To say this in another way, R_{ij} depends only on the extension which corresponds to T (the width), the width is increasing in Fig. 4.5(a) and Fig. 4.5(b) as the two dimensions¹ increase which would give opportunities to test the calculation of "openness". When the transformation matrix K_{ij} depends only on S and T we say here it is called a *constant example*.

For our first example we need one more piece of information: suppose we had a third operator which we call U which follows T at some arbitrary angle, as in Fig. 5. May this beginning to look terrible but this is the fun of it! First try you can make the most weird experiments just by drawing them. Now what is the $S \rightarrow T \rightarrow U$ transformation? What we really want to ask is the U obtained going from some state with respect to S to some other state with respect to T ? If we knew the transformation from S to T (from Eq. 4.1) we are then asking about an experiment in which both channels of T are open. We can get the answer by applying Eq. (4.5) twice in succession. For $S \rightarrow T \rightarrow U$ (transformation to the U representation) we have:

$$C_U = \sum_i R_{ij}^U C_j, \quad (4.8)$$

where we put the superscripts T on the R , so that we can distinguish it from the coefficients R^{Uj} we will have for going from T to U .

Assuming the amplitude C_j to be the bare value of the j -component (Eq. 4.3), we can reduce this to the T amplitude by using R_{ij} (Eq. 4.1) once more; we get

$$C_U = \sum_j R_{ij}^T C_j. \quad (4.9)$$

Now we can use the Eqs. 4.8 and (4.9) to get the transformation to U directly from S . Substituting C_j from Eq. (4.6) in Eq. (4.9) we have

$$C_U = \sum_i R_{ij}^T \sum_k R_{kj}^S C_k. \quad (4.10)$$

Or since R_{kj} has the property (4.6), we can put the summation also in front and write

$$C_U = \sum_j \sum_i R_{ij}^T R_{ji}^S C_i. \quad (4.11)$$

This is the form we found in the introduction.

Notice, however, that as long as all the terms in C_U are there, the size contribution of T is the same as the size that were in S . We notice just as well how much information from the representation directly to the U representation. It should be the same as putting C or U ahead is right after S in Fig. 4.4.

\leftarrow (b) In this step, we would have written

$$C_1^* = \sum_j R_j^{12} C_j. \quad (6.10)$$

and the coefficients R_j^{12} belonging to this transformation. Now, clearly, Eqs. (6.9) and (6.10) should give the same amplitudes C_1^* , and this should be true no matter what the original state ψ was which gave us the amplitudes C_j . So it must be that

$$R_j^{12} = \sum_i S_{ji}^{12} R_i^S. \quad (6.11)$$

In other words, for any rotation $S \rightarrow U$ of a reference basis, which is viewed as a composition of two successive rotations $S \rightarrow P$ and $P \rightarrow U$, the rotation matrix S_{ji}^{12} can be obtained from the matrix of the two partial rotations by Eq. (6.11). If you wish, you can think Eq. (6.11) to be entirely given by (6.10), since it only uses the unit relation δ_{ij} : $R_{ij}^S = \sum_k S_{ki}^{12} R_{jk}^S$.

To get to step (c) we should take the following parametrized formula. They are not much more difficult, however, so you can skip to the next section if you want. When we have said it is not quite right, we mean with we mean Eq. (6.12) and Eq. (6.13) more precisely than the one amplitude. Only the phase should be the same, all the amplitudes should be different by some constant phase factor. The ϵ in Eq. (6.13) is really the result of my calculation in the real world. So I wrote at Eq. (6.11) "which is really" is that:

$$C_1^* R_1^S = \sum_i R_i^S R_i^P. \quad (6.12)$$

where ϵ is a constant parameter. With this extra factor of ϵ^2 means, of course, is that the amplitudes except for a non-degenerate R_{ij}^S will differ by a constant phase factor between the amplitudes C_j and R_j^S . We know that, because ϵ is small, because ϵ is almost always real and positive, to within $O(\epsilon^2)$ the linear terms in R_{ij}^S are negligible. It is not true, however, that the value of an amplitude varies in a pure linear way. The phase factor will also appear. The ϵ in Eq. (6.13) will always be zero. Although it is an approximation to the rest of our derivation, we can just quickly get a rougher approximation to the total amplitudes. (If you don't like such short discussions, the don't worry about the proof and just skip to the definition of Eq. (6.13).)

Now we should say that Eq. (6.12) is the mathematical definition of a "product" in mathematics. It is just equivalent to saying " R_1^S is the product of R_1^P and S_{ji}^{12} ". Second, there is a theorem of mathematics—often called the "rule for the decomposition of a tensor product"—which says that the definition of a "product" of two tensors is the product of their determinants. Applying this theorem to Eq. (6.12), we get

$$\epsilon^{12} (Det R_1^S) = (Det R_1^P) \epsilon^{12} (Det S_{ji}^{12}). \quad (6.13)$$

(Remember of the analogy, because (6.8) don't tell us anything useful.) Now the last part. Remember that we, in the long run, hope to eliminate ϵ by writing R_1^S as a multiple of S_{ji}^{12} multiplied by ϵ^2 , so each individual determinant should be, ϵ -independent multiplied by ϵ^2 . Now let's take the right-hand side of Eq. (6.12) and divide it by ϵ^2 and see what we get.

$$\frac{R_1^S}{\epsilon^2 (Det R_1^S)} = \sum_j \frac{R_1^P}{\epsilon^2 (Det R_1^P)} \frac{R_j^P}{\epsilon^2 (Det S_{ji}^{12})}. \quad (6.14)$$

The ratio should have been disappeared.

Now it turns out that if you recall of our conclusion L-1 we gave you permission to be non-physical (which means, you remember, that $\sum_i \psi_i \neq 0$). So the rotation matrices will all have determinants that are pure imaginary exponential. That is, $\epsilon^{12} / \epsilon^2$ won't go to 1, you will see that it always comes out that $\epsilon \neq 1$. No worries, however, because rotation matrices have a simple process of making $\epsilon \cdot \epsilon = -1$ to come about. You just need a suitable ϵ in a basis of different ways. We make it a rule to "choose" it to "standard form" by writing

$$R_{j,k,l,m} = \frac{\epsilon_{j,k,l,m}}{\sqrt{Det R}}. \quad (6.15)$$

We can do this because we expect every particle in $\langle \psi | \psi \rangle$ by Gaussian phase factors to have the same weight. In what follows, we will always assume that $\langle \psi | \psi \rangle$ has been put in its "standard form". This we can do Eq. (6.11) without losing any of the above factors.

6-3 Rotations about the x -axis

We are now ready to find the transformation matrix R_x between two T basis spinors $|S\rangle$, with arbitrary components $S_{\alpha\beta}$, in our remaining basis $\langle \psi | \psi \rangle$ (no preferred direction). We have the basis we need for finding the matrix of any arbitrary rotation. There is only one solution. We begin with the simplest case which corresponds to a rotation about the x -axis. Suppose our two T apparatuses S and T placed in such a way a straight line with their axes $S_{\alpha\beta}$ and $T_{\alpha\beta}$ passes through the origin as shown in Fig. 6-1(a). We take $\psi_{\alpha\beta} = 0$ along S . Surely if $\psi_{\alpha\beta}$ becomes $(+)$ (or $(-)$) in S ; it must be $(+)$ in T . Suppose $\psi_{\alpha\beta}$ will be the same in the T apparatus. So, since T it goes down in y , it $\psi_{\alpha\beta}$ go down in T . Suppose, however, that the $\psi_{\alpha\beta}$ appearance were passed at some other angle, but still with the same orientation of S , as in Fig. 6-1(b). Initially y , T would say that $\psi_{\alpha\beta}$ in S would be $(+)$ (or $(-)$) in T , because the drift and field would be small in the same physical direction. And this would be quite right. Also, $\psi_{\alpha\beta}$ in T would still be in y . This is T because T would apply for any orientation of T in the apparatus S . What does this tell us about the relative phase between $S_{\alpha\beta} = (+)$, $T_{\alpha\beta} = (+)$, $S_{\alpha\beta} = (-)$, $T_{\alpha\beta} = (-)$? You might conclude that any rotation about S axis - of the "order of one year" (or how else) leaves the amplitude of $S_{\alpha\beta}$ probably unchanged, the same as before. We conclude $S_{\alpha\beta} = (+)$ and $T_{\alpha\beta} = (+)$ is probably wrong. At least one additional constraint must obtain. Our problem is to be in the "top" form so we come to the S and T apparatuses. That is,

$$C_1' = C_1 \quad \text{and} \quad C_2' = C_2.$$

We cannot say that the phases of the amplitudes referred to the T apparatus may not be different for the two different T orientations in (a) and (b) in Fig. 6-1.

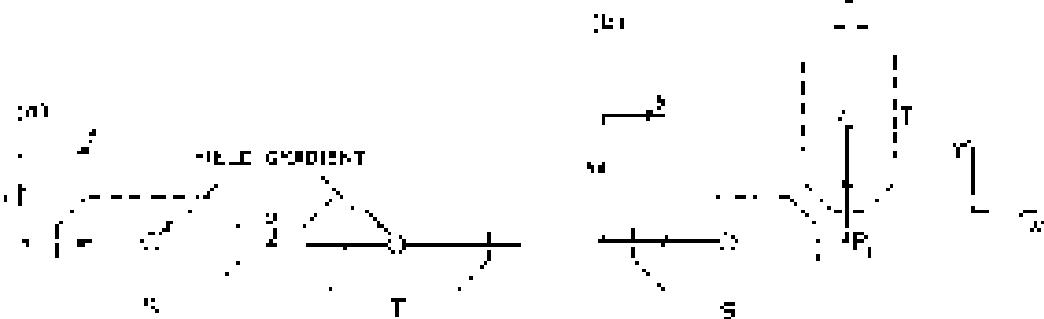


Fig. 6-1. Rotating 90° about the x -axis.

The two apparatuses in (a) and (b) of Fig. 6-1 are, in fact, different, as we learn in Fig. 6-2. In Fig. 6-2, suppose that we put an apparatus S (beam 1) which produces a pure $(+)$ state. (The $(+)$ state must be shown in one of the figures.) Such a state would be split into $(+)$ and $(-)$ beams in S , but the two beams would be combined into $(+)$ (+) or $(-)$ ($-$) = $(+)$ in T , the rest of S . The same thing happens again in T , as we follow Fig. 6-2 third quarter. This means it is the $(+)$ direction and, as shown in Fig. 6-2(c), the particle would go to the $(+)$ beam of S . Now imagine what happens if S and T are rotated around x by 90 degrees (parents shown in Fig. 6-2(d)). Again, the $(+)$ particle from S would end up in the $(+)$ beam, the rest of S would end up in $(-)$ state with respect to S . But T has analyzed the $(+)$ state with respect to S , which is different. By symmetry, we would now expect each one-half of the particles to get $(+)$, $(-)$.

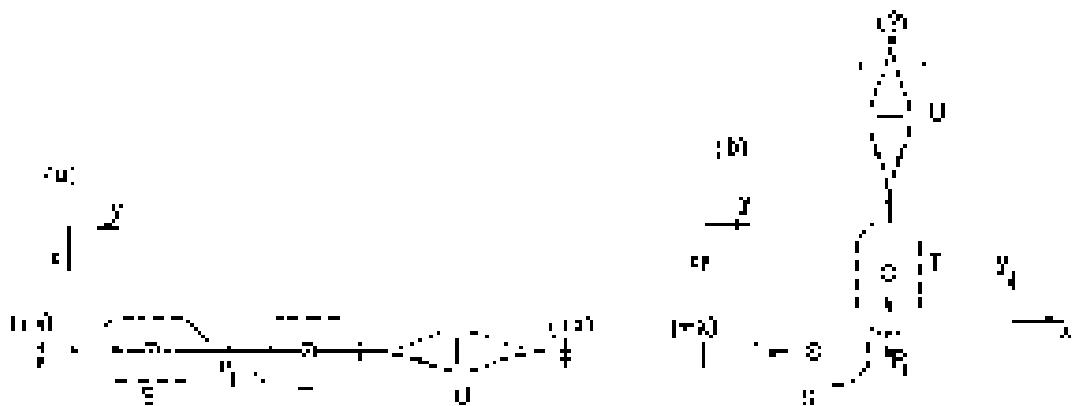


Fig. 6-5. Particle flow diagrams between differently labeled (a) and (b).

What could have changed? The amplitudes T and T' are total in the sense of no causal relationship between them. Can the physics be changed by changing T and T' in the different experiments? Our original assumption is that T and T' must be equal the amplitudes with respect to p are different in the two cases shown in Figs. 6-5(a) and, therefore, unlike in Fig. 6-4. There must be some very fine particle or interaction that is created the center of p_1 . How could it be? Well, we have decided that the amplitudes C_1^* and C_2^* are the same in the two cases but they result in two completely different physics. We conclude that C_1^* and C_2^* must be related by

$$C_1^* = e^k C_2^*,$$

and that C_1 and C_2 must be related by

$$C_1 = e^{-k} C_2,$$

where k and μ are real numbers which are constrained in some way by the analysis. However, k and μ

The only thing we can say at the moment about k and μ is that they must all be equal (except for the special case shown in Fig. 6-5(a), where T is in the same direction for both). We have seen that equal pion exchange in all amplitudes have no physical consequences. For the same reason, we can always add the same infinitesimal amount to both k and μ without changing anything. So we can parameterize C_1 and C_2 in terms of a nucleon number n and a same number. That is, we can always take

$$\lambda' = \lambda - \frac{(k + \mu)}{n}, \quad \lambda'' = \lambda - \frac{(k - \mu)}{n}$$

Then,

$$C_1^* = \frac{\lambda'}{n} = \frac{\lambda}{n} + \frac{\mu}{n} = C_2^*$$

so we adopt the convention that $\mu = -\lambda$. We have found the general rule that C_1 is related to the reference experiment by some angle ϕ from the reference orientation is

$$C_1^* = e^{-ik} C_1, \quad C_2^* = e^{+ik} C_2. \quad (6.15)$$

The λ value values are, in some, only by phases insufficient. From these formulae, one expels the μ in different results in the two experiments of Fig. 6-5.

Now we would like to know the law that relates ϕ to the angle between \vec{S} and \vec{T} . We already know the answer for one case. If the angle θ were ϕ is zero. Now we will assume that the phase shift ϕ is a continuous function of angle θ centered around $\theta = 0$ (see Fig. 6-6). This ϕ need to be π as may seems reasonable. In

[†] To do this, i.e., in other words, we are just solving the inhomogeneous in the "standard form" described in Section 6-7 by using Eq. (6.15).

elsewhere, if we rotate θ from θ_0 might have the angle β by the smaller plus the β is also a small quantity, say $\pi/2$, where π is large number. However, it is very because we can show that β can be zero. Let us do so. Suppose we want to put after θ_0 another rotation θ' which makes the angle α with θ . And, let's define the angle β with θ_0 . Then, with respect to θ_0 we have

$$C_0 = e^{i\theta} C_{\theta_0}$$

and with respect to θ' we have

$$C_0 = e^{i\theta'} C_{\theta'} = e^{i\theta} C_{\theta_0}$$

Then we know that we could get the same result if we put θ' right after θ_0 . Thus when the angle is dotted, the axes is dotted. We is a apparently called the angle β and the dot up the rotation will be a sequence of differential rotations. However, there are two angles, α , β and you should be careful. We can therefore write $\beta = -\alpha$.

The general result we get then is that C_0 is T rotated about the z axis by the angle β with respect to θ

$$C_0 = e^{i\theta} C_{\theta_0}, \quad C_0 = e^{i(\theta+\beta)} C_{\theta_0}. \quad (6.17)$$

For the angle β and the θ is satisfies we speak of in Definition, we expect the same. We can show that positive hydrogen is slightly modified due to the presence of electron of the reference axis. A positive θ is the state of motion of a proton-like system along \hat{z} in the positive direction.

Now we have to find what it means. First, we might try this argument. Suppose T is rotated by 180° ; then, clearly, one right back to own sign and we could have $C_0 = C_0$ and $C_0 = C_0$, or, what is the same thing, $e^{i\theta} = 1$. We get $\omega = 1/2$ regardless of the energy, since then it is, consider that T is rotated by 180° . If it were equal to 1, we would have $C_0 = e^{i\theta} C_{\theta_0} = -C_0$ and $C_0 = e^{i\theta} C_{\theta_0} = -C_0$. However, this is not the original state all over again. Both components are just multiplied by -1 which gives back the original physical system. (It is not a case of a common phase change.) This means that if the angle between T and θ in Fig. 6.8 that is used to 180° the system (with respect to θ) would be implemented from the original state situation, and the particles would again go through the (π) state of the θ system (i.e. $A = 180^\circ$, though the (1) state of the θ system is the $(-\pi)$ state of the original θ system). Since (-1) state would become $(-1) \times (-1) = 1$. But we have come now to choose the original state; the answer is wrong. We cannot have $\omega = 1$.

We must have the situation that θ rotates by 180° and the reader angle preserves the same physical state. This θ' happens $\theta' = \theta + \pi$ and only here with the first angle that represents the same physical state be $\theta = 180^\circ$. It gives

$$\begin{cases} C_0 = -C_{\theta_0} \\ C_0 = -C_{\theta_0} \end{cases} \quad \text{if } \theta \text{ about } z \text{ axis.} \quad (6.18)$$

It is very curious theory, but if you can understand this 180° you get the simple idea. They aren't really new than π , because the π and -1 state of sign doesn't give any different physics. If someone else has counted to me, recall the time of the sun it does have to be 180° he had turned 360° , that's all right, as you the same physics. Same thing occurs. So then if we know these valences C_0 and C_0 for both the ball particle with respect to θ situation, then θ and we can use a box

*I.e., $\exp(i\theta) = -1$ wouldn't work. However we can multiply $i\theta$ the change θ significantly reduces the rotation for a certain problem.

**Note, if something has been written in a diagram of small, extremely short text, i.e. to remind it to the reader, it is considered to make the law that it has been *written* *done*—in other terms and not *written*. If you have lost track of the whole theory, it is naturally enough, it is very hard to find a condition of T !!

system referred to \hat{Z} which is obtained from S by rotation of ϕ around the \hat{x} -axis, the new comp' values are given in terms of the old by

$$\begin{aligned} C_1 &= e^{j\phi} C_1 \\ C_2 &= e^{-j\phi} C_2 \end{aligned} \quad \left| \text{substituted} \right. \quad (6.19)$$

6-4 Rotations of 180° and 90° about \hat{y}

Next, we will try to guess the transformation for a rotation of \hat{Z} with respect to S of 180° around an axis perpendicular to the $\hat{x}\hat{z}$ -plane, i.e., about the \hat{y} -axis. We have defined the coordinate system in Fig. 6-1. In another model we have a 180° identical second Gersh-Gerach apparatus, with the second one, T , turned "upside down" with respect to the first one, S , as in Fig. 6-6(a). Now if we do this, our articles or little magnetic dipole, a particle that is the $(+S)$ state—will still pass in the "upper" path in the first apparatus—will also take the "upper" path in the second, so that it will be in the same state with respect to T . (i.e., the second apparatus, both the gradients and the field directions are reversed; for a given field pass to the "upper" position in a given direction, the "one" is unchanged.) Thus, since \hat{z}' will be \hat{z} with respect to S will be "down" with respect to T . For these relative positions of S and T , then, we know that the transformation must give

$$|C'_1| = |C_1|, \quad C'_2 = -C_2.$$

As before, we cannot right out since ordering of phase factors we don't know (for 180° about the \hat{y} -axis)

$$C_1 = e^{j\phi} \quad \text{and} \quad C_2 = e^{j\psi}, \quad (6.20)$$

where ϕ and ψ are still to be determined.

What about a rotation of 360° about the \hat{y} -axis? Well, we already know the answer for a rotation of 180° must be rotation by simply adding π to one charge's sign. A rotation of 360° forced any real charges to come back to the original position. It must be 0. So, for any 360° rotation, the result is the same as a 180° rotation about \hat{y} —amplitude simply changes sign. Now suppose we combine two successive rotations of 180° about \hat{y} ; using Eq. (6.20) we should get the result of Eq. (6.18). In other words,

$$\begin{aligned} C'_1 &= e^{j\phi} C_1 = e^{j\phi} C_1 C_1 = -C_1 \\ \text{and} \\ C'_2 &= e^{j\phi} C_2 + e^{j\psi} C_2 = -C_2 \end{aligned} \quad (6.21)$$

But, this isn't

$$e^{j(\phi + \psi)} = -1 \quad \text{or} \quad \phi + \psi = \pi$$

So the transformation for a rotation of 180° about the \hat{y} -axis can be written

$$C_1 \rightarrow e^{j\phi} C_1, \quad C_2 \rightarrow -e^{j\phi} C_1. \quad (6.22)$$

The equations just read work equally well for a rotation of 360° that goes right in the $\hat{x}\hat{z}$ -plane, although different axes can, of course, give different numbers for ϕ . However, that is the only way they can differ. Now there is a certain amount of arbitrariness in the number ϕ , but once it is specified for one axis of rotation—the \hat{y} -plane—it is determined for any other axis. The conventional to choose to set $\phi = 0$ for a 180° rotation about the \hat{y} -axis.

To show that we have the choice, suppose we imagine that ϕ was not equal to zero for a rotation about the \hat{y} -axis. Then we can choose ϕ there is even more to do the $\hat{x}\hat{z}$ -plane, for which the corresponding phase factor will be zero. Let's find the phase factor ϕ , for an axis \hat{y}' that makes the angle α with the \hat{y} -axis, as shown in Fig. 6-6(b). (For clarity, the figure is drawn with a capital α as a very large number, but that doesn't matter.) Now if we take a \mathcal{Y} apparatus which is initially lined up with the $\hat{x}\hat{z}$ -plane and \hat{y}' in $\hat{x}\hat{z}$ -plane 180° along the \hat{x} -axis, its axes will be x'' , y'' , z'' , and \hat{y}'' will pass through Fig. 6-6(b). The amplitudes

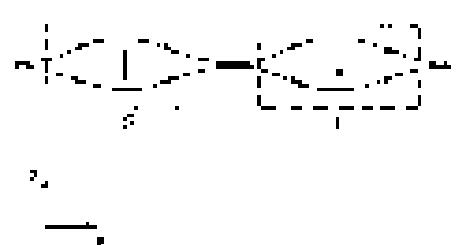


Fig. 6-6. A rotation of 180° about the \hat{y} -axis.

with respect to \hat{z} will then be

$$C' = e^{i\theta} C_{\perp}, \quad C_{\perp} = e^{-i\theta} C_1. \quad (6.23)$$

We can now think of getting to the static velocity for the two successive rotations about \hat{x} and \hat{y} of the particle. First we implement equation (6.23) which is relative with respect to \hat{y} by 180° about the \hat{y} -axis. The axis \hat{x}_1 , \hat{y}_1 and \hat{z}_1 of C_1 will be as shown in Fig. 6.5(a), and the output axes with respect to \hat{y} are given by (6.23).

Now notice that we can get from C_1 to C' by a rotation about the "new" \hat{x}_1^* (namely about \hat{x}_1^* , see diagram Fig. 6.5(b)). From the figure you can see that the angle required is two times the angle α but in the opposite direction (with respect to \hat{x}_1^*). Using the transformation of (6.12) with $\phi = -2\alpha$, we get

$$C'_1 = e^{i\phi} C_{\perp}, \quad C'_{\perp} = e^{-i\phi} C_1. \quad (6.24)$$

Combining Eqs. (6.23) and (6.24) we get that

$$C''_1 = e^{i\theta-i\phi} C_{\perp}, \quad C''_{\perp} = e^{-i\theta+i\phi} C_1. \quad (6.25)$$

These amplitudes must, of course, be the same as we get in (6.21). So ω_1 must be related to ω_0 by

$$\omega_1 = n - \omega_0. \quad (6.26)$$

This means that if ω_0 is finite between the \hat{x} -axis and the particle (if S) is static, to get the 180° rotation by a rotation of 180° about \hat{x}_1^* we need

Now we have one more coordinate system to consider, the \hat{x}_2 -axis going to the right (\hat{x}_2 is the \hat{x} -axis as we take it to be the \hat{y} -axis). Its particle value of ω_2 is zero, and we repeat the same in general now. Our result after a rotation of 180° about the \hat{x}_2 -axis, we have:

$$\begin{aligned} C'_1 &= C_{\perp}, \\ C''_1 &= -C_1. \end{aligned} \quad \left| \text{180° about } \hat{x}_2 \right. \quad (6.27)$$

While we are thinking about the particle let's do a task for the transformation matrix for a rotation of 180° about \hat{x}_2 . You can find it because we know that the successive 180° rotations about the \hat{x} -axis just yield one 180° rotation. We start by writing the transformation for 90° in the \hat{x}_2 -axis frame. Then

$$C_{\perp} = iC_+ + iC_{\perp}, \quad C_+ = \omega_2 + iC_{\perp}. \quad (6.28)$$

A second rotation of 90° about the same axis would have the same coefficients:

$$C''_1 = iC_{\perp} - iC_+, \quad C''_{\perp} = iC_+ + iC_{\perp}. \quad (6.29)$$

Combining Eqs. (6.28) and (6.29), we have:

$$C''_1 = i(iC_+ + iC_{\perp}) + i(iC_{\perp} - iC_+) = -2iC_+. \quad (6.30)$$

$$C''_{\perp} = -i(iC_+ + iC_{\perp}) + i(iC_{\perp} - iC_+) = 2iC_{\perp}.$$

However, from (6.27) we know that

$$C''_1 = C_{\perp}, \quad C''_{\perp} = -C_1.$$

So the last must be iC_{\perp} and

$$iC_+ = 2iC_{\perp} - 1$$

$$iC_+^2 - 2iC_{\perp} = 0,$$

$$iC_+^2 = 2iC_{\perp} - 1,$$

$$iC_+^2 + 2i^2 = 0.$$

$$iC_+^2 = -2.$$

$$C_+^2 = -2i^2 = 2.$$

$$C_+ = \pm \sqrt{2}i.$$

These four equations are enough to determine all our unknowns ω_0 , ω_1 , ω_2 , θ , ϕ , ψ & χ .

θ is not held in $\Delta\theta$. Look at the second and last θ equations. Deduce that $\rho^2 = \rho'^2$ which means that $\rho = \rho'$, also that $\theta = -\theta'$. But $\rho = d > 0$, meaning that the first equation would give $d = d' = 0$. Using this, we know immediately that $\theta = 180^\circ$, and that $\rho = -1/\sin\theta$. Now we have everything in terms of θ . The top, and the second equations all in terms of θ , we have

$$\rho^2 - \frac{1}{\sin^2\theta} = 0 \quad \text{or} \quad \rho^2 = \frac{1}{\sin^2\theta}$$

This equation has two real linear solutions, our only way of from gives the standard value for the determinant. We might as well have $\rho = 1/\sqrt{2}(1, i\sin\theta)$

$$\begin{aligned} \rho &= 1/\sqrt{2} & \theta &= 180^\circ \\ \rho &= -1/\sqrt{2} & \theta &= -180^\circ \end{aligned}$$

In other words, the two operations S_1 and S_2 , with T added with respect to θ by 90° from the result, are transformations:

$$\begin{aligned} C_1 &= \frac{1}{\sqrt{2}}(C_0 + C_2) \\ C_2 &= \frac{1}{\sqrt{2}}(-C_0 + C_1) \end{aligned} \quad \left| \begin{array}{l} \text{by definition,} \\ \text{eqn 6.5} \end{array} \right.$$

We can, of course, solve these equations for C_0 and C_2 , which will give us the parameters for the rotation of angle 90° from θ . Simplifying the process a touch, we would conclude that

$$\begin{aligned} C_0 &= \frac{1}{\sqrt{2}}(C_1 + C_2) \\ C_2 &= \frac{1}{\sqrt{2}}(C_1 - C_0) \end{aligned} \quad \left| \begin{array}{l} \text{—90}^\circ \text{ about } x \\ \text{eqn 6.5} \end{array} \right.$$

6.5 Transformations about x

You may be thinking, "This is getting ridiculous. What are they going to do next, fit a curve to the 1000 points and so on, forever?" No, we are almost finished. With just one or two more steps here, 90° about x , and an interesting thing about x (which we did later if you remember), we can generate any rotation at all.

For illustration, suppose we want the angle θ around x . You know how to do it: $\theta =$ the angle α around x , between \hat{x} and \hat{x}' respectively. How do we get \hat{x}' ? First, we turn the axis x clockwise by $+90^\circ$ about x (Fig. 6.7). Then we turn through the angle θ around \hat{x} . That is, we turn x clockwise by -90° about x . The sequence of the three rotations is the same as turning around x by the angle θ . This is a property of S_1 and S_2 .

These facts of transformation theory are useful in practice, and where they produce something very unintuitively. It is rather striking, because we have to read θ clockwise, and that has to be an appropriate angle, happens to work on this way and then the next. Perhaps, if we were fisher folk, and had to appreciate what happens when we turn our boats, it would be equally as easy to speculate such things!

Anyway, let's work out the transformation for a rotation θ around x about x by using what we know. From the first rotation by $+90^\circ$ around x , the components go according to Eq. (6.6). Calling the coordinates (x, y, z) , and \hat{x} to

The main solution, using all three x, y, z , and θ reduces to $\rho = \sqrt{2}\rho$ is given.

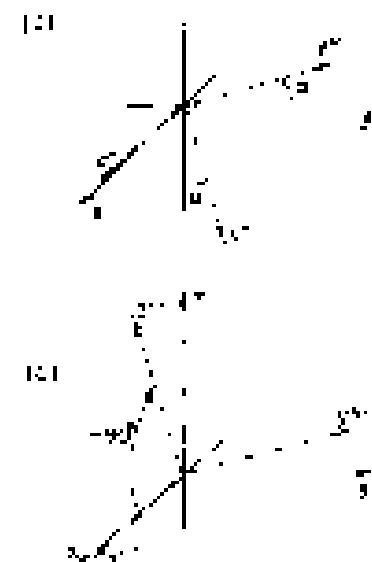
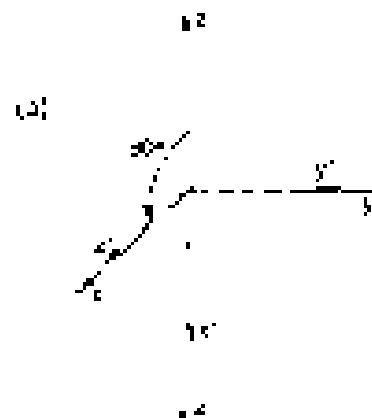


Fig. 6.8. A vector r in \mathbb{R}^2 about the x axis, is first rotated by 90° clockwise by $+90^\circ$ about x , followed by the a rotation by θ about x' , followed by a rotation by -90° about x .

last relation. By the angle α around x' takes us to a frame x'', y'', z'' , for which,

$$C'' = e^{i\alpha/2} C_{+}, \quad C''_+ = e^{-i\alpha/2} C_+.$$

The last relation of 90° about y'' takes us to (x'', y'', z'') by (6.12).

$$C''' = \frac{1}{\sqrt{2}} (C'' - C''_+) = e^{i\alpha} + \frac{1}{\sqrt{2}} (C_+ + C_-).$$

Combining these last two transformations we get

$$C_+''' = \frac{1}{\sqrt{2}} (e^{i\alpha/2} C_+ - e^{-i\alpha/2} C_-)$$

$$C_-''' = \frac{1}{\sqrt{2}} (e^{i\alpha/2} C_- + e^{-i\alpha/2} C_+).$$

Using Eqs. (6.12) for C_+ and C_- , we get the one-step transformation:

$$C''' = i e^{i\alpha/2} (C_+ + C_-) = e^{-i\alpha} (-C_- + C_+),$$

$$C''' = \frac{1}{2} (e^{i\alpha/2} (C_+ + C_-) + e^{-i\alpha/2} (C_+ - C_-)).$$

We can put these 7 relations in a simpler form by remembering that

$$e^{i\theta} + e^{-i\theta} = 2 \cos \theta, \quad \sin^2 \theta = 1 - \cos^2 \theta = 0.$$

We get,

$$\left. \begin{aligned} C_+''' &= \left(\cos \frac{\alpha}{2} \right) C_+ - i \left(\sin \frac{\alpha}{2} \right) C_- \\ C_-''' &= i \left(\sin \frac{\alpha}{2} \right) C_+ + \left(\cos \frac{\alpha}{2} \right) C_- \end{aligned} \right\} \text{one-step.} \quad (6.34)$$

This is our transformation for a rotation about the x -axis by an angle α . It is only a little more complicated than the others.

4-6 Arbitrary rotations

Now we can set two to three angles at all. After, above and are relative orientation of two non-linear frames can be described in terms of three angles, as shown in Fig. 6-9. If we have a set of coordinates, x , y , and z oriented in any way at all with respect to x' , y' , and z' , we can close the relationship between the two frames by means of the three Euler angles α , β , and γ , which define three successive rotations, but will bring the x , y , z frame into the x' , y' , z' frame. Starting at x , y , z , we rotate our frame through the angle α about the x -axis, bringing the x axis to C along x . Then, we rotate by β about the temporary x -axis, to bring x down to y . Finally, a rotation about the new x -axis (that is, y) by the angle γ will bring the x -axis into x' and the y -axis into y' . We know the transformations for each of the three rotations—already are given in (6.12) and (6.13). Combining them in the proper order, we get

$$C'_- = \cos \frac{\alpha}{2} e^{i\alpha/2} e^{i\beta/2} C_- - i \sin \frac{\alpha}{2} e^{i\alpha/2} e^{i\beta/2} C_+, \quad (6.35)$$

$$C'_+ = \sin \frac{\alpha}{2} e^{i\alpha/2} e^{i\beta/2} C_+ + \cos \frac{\alpha}{2} e^{i\alpha/2} e^{i\beta/2} C_-$$

So starting from some arbitrary frame about the properties of x , y , z we have derived the complete transformation for any rotation x , y , z . That means we if

* With a little work you can show that the "axis x , y , z " can also be brought from the frame x' , y' , z' by the following three relations about the original axes: (1) rotate by the angle β around the original x -axis; (2) rotate by the angle α around the original y -axis; (3) rotate by the angle γ around the original z -axis.

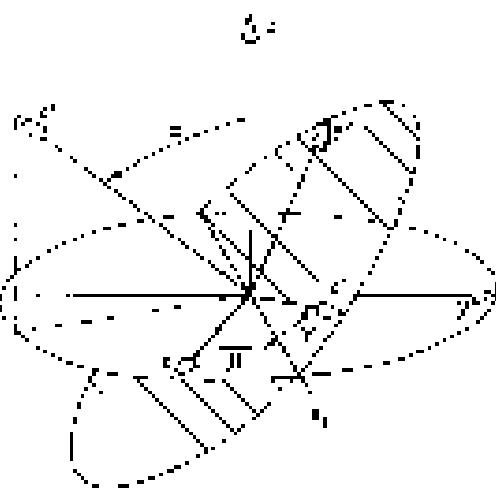


Fig. 6-9. The orientation of one coordinate frame x', y', z' relative to another frame x, y, z can be defined by three Euler angles ψ, θ, ϕ .

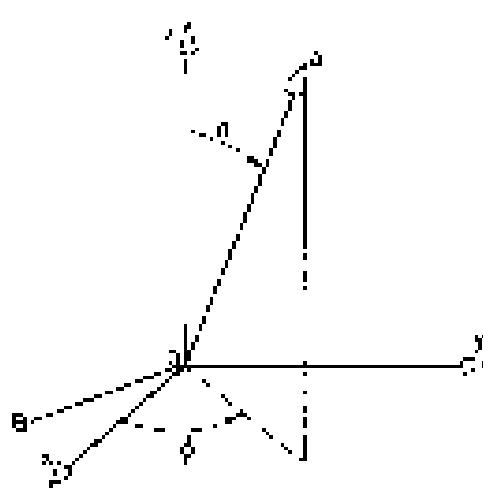


Fig. 6-10. A vector A defined by the polar angles θ and ϕ .

We know the amplitudes for any state of a spin- $\frac{1}{2}$ particle, say, the $|+ \rangle$ we get if a spin- $\frac{1}{2}$ particle, say ψ , is at R . Since ψ has x , y , and z , we can calculate what fraction would go into every term of an expansion T with the axes x , y , and z . In other words, if we take a state ψ of a spin- $\frac{1}{2}$ particle, whose amplitudes are $C_1 = |\psi\rangle$ and $C_2 = -i|\psi\rangle$ to be "up" and "down" with respect to the z -axis of the x, y, z frame, we also know the amplitudes C_3 and C_4 to be "up" and "down" with respect to the x -axis of any other frame x', y', z' . The C_3 coefficients in Eqs. (6-55) are the terms of the "non-intrinsic part" T with which we can project the amplitude of a spin- $\frac{1}{2}$ particle in any other coordinate system.

We will now work out a few examples to show you how it all works. Let's start with a following example: a spin- $\frac{1}{2}$ particle through a Stern-Gerlach apparatus and emerges only the $(+ +)$ state. What is the amplitude that it will be in the $(+ +)$ state? The x axis is the same as the $-z$ axis of a certain rotated $(+ +)$ frame, the y -axis. For this situation, there it is directed to us. Fig. (6-10) shows you how to do it; just use the transformation equations of (6-50). Since $C_1 = 1$ and $C_2 = 0$, we get $C_3 = 1/\sqrt{2}$. The probabilities are the $1/2$. In a sense, of these amplitudes, there is a 50 percent chance that the particle will go through an x -axis ψ that after the y -axis stays. If we had used again the $(+ +)$ frame the amplitude would have been $-1/\sqrt{2}$, which is also a probability 1/2, as you would expect from the symmetry of spins. So if a particle is in the $(+ +)$ state, it is equally likely to be in $(+ +)$ or $(- -)$, but with symmetric prob.

There's no projector in y either. A particle in the $(+ +)$ state has a 50-50 chance of being $(+ +)$ or $(- -)$. However, for these using the formula for rotating -90° about the x axis, the amplitudes are $1/\sqrt{2}$ and $-i/\sqrt{2}$. In this case, the new amplitudes have a phase difference of 90° instead of 180° , as they did for the $(+ +)$ and $(- -)$. In fact, that's how the orientation ψ and ϕ always appear.

As our last example, suppose that we know that a spin- $\frac{1}{2}$ particle ψ in a state ψ such that it is polarized "up" along some axis a , denoted by the angle θ and ϕ in Fig. 6-10. We want to know the x amplitudes (C_1, C_2) if the particle is "up" along x and the amplitude (C_3, C_4) if it is "down" along x . We can find these amplitudes by noticing that a is the z -axis of a system whose x -axis is in some arbitrary direction, say, in the plane formed by a and x . We can then, using the change of frame x, y, z by 90° rotations, first, we make a rotation by $-\pi/2$ about the axis a , which aligns the x axis and the line a in one figure. Then we rotate by $\pi/2$ about the line a to get a second frame x', y', z' aligned a to the x -axis. Finally, we rotate by the angle (say) ϕ around x . Remembering the relations with x, y, z

$$C_+ = \cos \frac{\theta}{2} e^{i\phi/2}, \quad C_- = \sin \frac{\theta}{2} e^{i\phi/2}. \quad (6-1)$$

We would like, finally, to summarize the results of this chapter in a form that will be useful for your later work. First, we remind you that our primary result in Figs. 6-1, 6-2 can be written in an alternative form. Note that Eqs. (6-3) mean just the same thing as Eq. (6-1). That is, in Eqs. (6-3) the coefficients of C_+ , C_0 , S , β , and C_- are just the amplitudes $|S|$ (3) or Eq. (6-1). The amplitude that a particle in the state will, respect to S will be in the state with respect to C_0 when the orientation of S with respect to S is given in terms of the angles θ , ϕ , and ψ . We also call $|S|$ in Eq. (6-3) (We don't know what we called it!) For example, $|S|^2 = |C_0|^2 + |C_+|^2 + |C_-|^2$. In the formula for C_0 we multiply $e^{i\phi/2} e^{i\theta/2}$. Likewise, there is a more or less similar set of results in the form of a table, so we have drawn a table here.

It will be especially handy to have these amplitudes already worked out for some simple special cases. For example, and for a rotation by the angle ϕ about the z -axis. You can also let ϕ stand for the corresponding rotation matrix containing the coefficients c_{ij} , which has to be completely understood. In the same way, if (θ, ϕ, ψ) will stand for rotations by the angle θ about the y -axis, the angle ϕ about the x -axis, we give it to you the matrices c_{ij} which define the amplitude $|S|^2$, which project the amplitudes from the S frame into the A frame, where T is obtained from S by the rotation specified.

Table 6-2

The amplitudes $|T|$, $|S|$ for a rotation $R(\theta)$ by the angle θ about the x -axis, y -axis, or z -axis.

Table 6-1

The amplitudes $|T|$, $|S|$ for a rotation $R(\theta)$ by the Euler angles (θ, ϕ, ψ) of Fig. 6-2

$R \propto \hat{x}, \hat{y}, \hat{z}$

(T, S)	S	T
T	$\cos \frac{\theta}{2} e^{i\phi/2} e^{i\psi/2}$	$\sin \frac{\theta}{2} e^{i\phi/2} e^{i\psi/2}$
S	$\sin \frac{\theta}{2} e^{i\phi/2} e^{i\psi/2}$	$\cos \frac{\theta}{2} e^{i\phi/2} e^{i\psi/2}$

6-1(a)

(T, S)	S	T
T	$e^{i\phi/2}$	$e^{i\theta/2}$
S	0	$e^{-i\phi/2}$

6-1(b)

(T, S)	S	T
T	$\cos \theta/2$	$\sin \theta/2$
S	$\sin \theta/2$	$-\cos \theta/2$

6-1(c)

(T, S)	S	T
T	$\cos \psi/2$	$\sin \psi/2$
S	$-\sin \psi/2$	$\cos \psi/2$

The Dependence of Amplitudes on Time

7.1 Atom at rest: stationary states

We want now to look a little bit about the behavior of probability amplitudes in time. We say a "link hi," that is the causal behavior in time necessarily connects the behavior in space as well. Then, we get immediately one the most complicated possible situation if we try to do it exactly and in detail. We are always in the situation that we can either build something which is logically rigorous but quite intricate very far, we can do something which is not so all rigorous but which gives us some idea of a less intricate postponing until later a more careful treatment. With regard to the time dependence we do not go to take the second route. We will make a number of statements. We will not try to say you must do ~~it~~ just by telling you things that have been found out to give you some setting for the behavior of amplitude calculation of basis. As we go along, the question of the uncertainty will increase, and I am going to assume that we continue by picking things out of the air. It is, of course, all part of the process of experiments and of the imagination of people. But it would take us too long to go over the historical development, so we have to jump in sometimes. We could always take the easiest and desire everything which you would not understand, we would get through a large number of exceptions to easily make statement. We choose to do something in between.

An electron does not simply speed up, under certain circumstances, lose a certain definite energy. For example, if it is standing still (so it has no transverse motion, or, among two, a kinetic energy), it has its rest energy. A little compact object like an atom can also have a definite energy when standing still, but it could also internally excited to another energy level. We will come to after the summary of this. We can often think of an atom in an excited state as having a definite energy, but this is really only approximately true. An atom doesn't stay excited forever because it manages to discharge its energy by its interaction with the electromagnetic field. So there is some amplitude that a new state is generated with the atom in lower state and the electromagnetic field in a higher state of excitation. The total energy of the system is the same before and after, but the energy of the atom is reduced. So it is not precise to say an energy atom has a definite energy, but it still has. In addition, it is coherent and not too wrong to say that it does.

Uncertainty, why doesn't you go to very excited in the other way? Why does an atom radiate energy? The answer has to do with entropy. When the energy is in the electromagnetic field, there are various different ways it can be stored in different places where it can wander—over time. In the right conditions, we find that in the most probable situation the field is excited with a photon, and the atom is de-excited. It takes a very long time for the photon to come back and find out it may have been back so again. This is just a typical problem of classical problems: Why does an accelerating charge induce? It isn't clear if "why" makes sense, because, in fact, what it induces if the way of the world to be some sort of wave function. But in fact, a strong time goes in the direction of increasing entropy.

Nuclei can also exist in different energy levels, are in an approximation which disregards the electro-magnetic effects, we can say that a nucleus can excited state stays there. Although we know that it doesn't stay there forever, it is often useful to start out with an approximation which is somewhat simplified and easier to think about. Also it is often a legitimate approximation under certain circumstances. (When we first introduced the classical laws of a falling body, we did not include friction, but there is almost never a case in which there isn't some friction.)

7.1 Atom at rest: stationary states

7.1 Uniform motion

7.3 Potential energy-energy conservation

7.4 Jumps; the classical limit

7.5 The "precession" of a spin-orbit particle

Review: Chapter 7, Sec. 1, Summary
Chapter 43, Sec. 1, Summary

The ψ 's are the solutions to "energy parades," which have various names. But it's better to say "discrete energy state." Light particles moving in it is not correct to say that they *have* a precisely definite energy. They would be out only if they *were*. So what we mean by "discrete energy" is that they *have* a definite energy, we are trying to be clear that they *must* have it. For the moment, then, we will intentionally forget about such questions and just stick now to these discrete states.

Suppose we have an atom—or an electron or any particle—which at first would have a definite energy, E_0 . By the energy E_0 , we mean the mass of the particle times c^2 . This mass includes any kinetic energy, so an excited atom has a mass which is different from the mass of the same atom in its ground state. (The ground state means the state of "lowest energy.") We will call E_0 the "energy at rest."

For reasons which I'll explain in much greater detail later, the time evolution of the wave function ψ does not depend on position. This means, of course, that the probability of finding the atom anywhere is the same. But it *does* depend on time. The probability could be independent of position, and still the phase of the amplitude would vary from point to point. But for a particle at rest, the amplitude is identical everywhere. It does, however, depend on the time. For a particle at rest of definite energy E_0 , the amplitude to find the particle at (x, y, z) at the time t is

$$\psi = e^{iE_0 t/\hbar}, \quad (7.1)$$

where e is some constant—the amplitude to be at any point in space is the same for all points, but depends on time according to (7.1). You just simply assume this rule to be true.

If we want we could also write (7.1) as

$$e^{-iE_0 t/\hbar}, \quad (7.2)$$

with

$$\psi = E_0 - i\hbar t,$$

where E is the rest mass of the atom's state or particle. There are three different ways of specifying the energy: by the frequency of an amplitude, by the energy in the classical sense, or by the mass. They are all equivalent. They are just different ways of saying the same thing.

You may be thinking, but it is a *rare* to think of a "particle" which has *exact* amplitude to be found throughout all space. After all, we usually imagine a "particle" to be a small object, called "something." But don't forget the uncertainty principle. If a particle has certain energy, it has *some* definite momentum. If the uncertainty in momentum is zero, the uncertainty relation, $\Delta p \Delta x = \hbar$, tells us that the uncertainty in the position must be infinite and that is what we are saying here. We say that there is *no* amplitude to find the particle at all points in space.

If the different parts of an atom are in a different state with a different total energy than the *rest mass* of the amplitude with time t of course. In your book you might see it is said that *one* atom cannot *possibly* be in one state and a certain amplitude, say ψ_1 and ψ_2 —and each of these amplitudes will have a different frequency. However, let's introduce *several* different movements—like a ballerina—which can come up as a varying probability. Something else, like "coming out" rather than "going in though it is not out" is the case that is center of mass is not decaying. However, if the atom has one definite energy, the amplitude is given by (7.1), and the distinct "parts" of this amplitude does not depend on time. You see that this "thing" has a definite energy and you are not probability of random absorption, the answer is independent of time. Although the amplitude vary with time, if the energy is definite they vary as an imaginary exponential, and the absolute value doesn't change.

That's why we often say random motion is a definite and periodic—*isochronous* wave. If you make any measurements of the thing inside, you'll find that varying the probabilities will change its form. In order to test the probabilities chapter 7.2

that we have to have the amplitudes of two different frequencies, and they must be in phase since we can't know what they are. The states ψ_1 , ψ_2 have one amplitude to go in a state of one energy and another amplitude to be in a state of another energy. That's the quantum mechanical description of something when its behavior depends on time.

If we have a "two-state" system—say, two different states with different energies, with the amplitude of each of the two states varying with time according to e^{-iE_t} (7.2), for instance, as

$$\psi = \psi_1 e^{-iE_1 t} + \psi_2 e^{-iE_2 t}, \quad (7.3)$$

and if we have some combination of the two, we will have no interference. But if we add a constant to both energies, it would make any difference. If somebody else were to use a different scale of energy in which all the energies were increased (or decreased) by a constant amount—say, by the amount A —then the amplitudes ψ_1 between states would, from the point of view, be

$$\psi = \psi_1 e^{iA t} + \psi_2 e^{-i(A+E_2)t}. \quad (7.4)$$

All of his amplitudes would be multiplied by the same factor, e^{iAt} , and all these amplitudes, ψ , interfere, should have the same factor. When we find the resonance energies of nuclei or molecules, all the energies would be the same. The choice of an origin for our energy scale makes no difference; we can measure energy from any arbitrary point. For relativistic purposes it is best to measure the energy so that the rest mass is included, but for many purposes that aren't very difficult, it is often nice to subtract some constant amount from all energies that appear. For instance, in the case of an atom, it is usually convenient to subtract the energy Mc^2 , where M is the mass of all the separate pieces—the nucleus and the electrons—which is, of course, the mass of the atom. For other problems it may be better to subtract from the energies the amount $4\pi r^2 E_0$, where E_0 is the energy of the whole atom in the ground state. Even the energy that appears is just the excitation energy of the atom. So sometimes we may shift our point of energy by some very large constant, but it doesn't really say *it changes*; provided we shift all the energies in a particular calculation by the same constant. So search for a *useful* *origin* of energy.

7-2 Uniform motion

If we suppose that the relative laws of light particles are at one inertial system to have uniform motion in another inertial system. In the rest frame of the particle, the probability amplitude does not vary with x , y , or z —it varies with t . The magnitude of the amplitude is the same for all t , but the phase depends on t . We can get a feel of what part of the behavior of the amplitude these plot lines of equal phase—say, lines of zero phase—as a function of x and t . They are identical, these equal-phase lines are parallel to the x -axis and are equally spaced in the x -direction, as shown by the dashed lines in Fig. 7-1.

In a different frame— x' , y' , z' —uniformly moving with respect to the x -axis in, say, the x -direction, the x' and t' coordinates of any particular point in space are related to x and t by the Lorentz transformation. This transformation can be represented graphically by drawing x' and t' axes, as in Fig. 7-1. Note that x' , t' , y' , z' are not perpendicular to the x -axis, so the trajectory of the time coordinate is different. Also the x -axis is tilted at the phase with t' , so the probability amplitude must be a function of x .

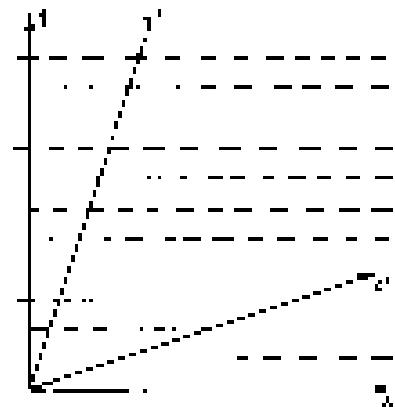


Fig. 7-1. Relationship between values of the amplitude ψ and the time coordinate t .

[†]We are assuming in the above that two measurements are at corresponding points in the two systems. This is a safe assumption, since the place of a measurement is under the influence of no body external to them. A complete justification of this assumption requires a more detailed discussion involving the theory of field amplitude.

Under a Lorentz transformation for the velocity v , say along the x -axis, the time is related to the time t' by

$$t' = \frac{t - v z/c^2}{\sqrt{1 - v^2/c^2}},$$

so in a spacelike case we have

$$\gamma' = \gamma / \sqrt{1 - v^2/c^2} = \gamma / \sqrt{1 - E_p^2/E_0^2},$$

In the plane system, with its axes at rest relative to Earth, if we take the impulse as

$$p = \gamma E_p \hat{v} = \gamma p_0,$$

we see that $E_p = E_0 \sqrt{1 - E_0^2/E_0^2}$ is the energy conserved classically, i.e. a particle of rest energy E_0 travelling at the velocity v , and $p = \gamma p_0 \hat{v}^2$ is the corresponding particle momentum.

You know that $\hat{v}_x = (\hat{v}_x, \hat{v}_y, \hat{v}_z) = (E_0 p_x, E_0 p_y, E_0 p_z)$ is the velocity, and $(\hat{v}_x, \hat{v}_y, \hat{v}_z) = (0, \gamma p_0, \gamma p_0)$ is a polar representation. In the rest frame of the particle, p_0 , is just E_0 ; so the other components \hat{v}_y and \hat{v}_z will be zero, so

$$\hat{p}^2 = \hat{p} \cdot \hat{v}.$$

Thus, the probability amplitude of a particle which has the same energy will be proportional to

$$e^{-i(\hat{p}^2 + E_0^2)t/\hbar c^2}, \quad (7.5)$$

where E_0 is the energy of the particle whose momentum is p . This is

$$E_0 = \sqrt{(E_0^2 + E_b^2)}, \quad (7.6)$$

where E_b is the binding energy. For non-relativistic problems, we can write

$$E_0 = M_p c^2 + W_b, \quad (7.7)$$

where M_p is the energy over and above the rest energy $M_p c^2$ of the parts of the system. In general, W_b would include both the kinetic and g_F of the system and the binding or "rest" energy, where we can call the "total" energy. We would write

$$W_b = W_{tot} - \frac{c^2}{M_p}, \quad (7.8)$$

and the amplitudes would be

$$e^{-i(W_{tot} + E_0^2)t/\hbar c^2}. \quad (7.9)$$

But if we set E_0^2 and E_0 to be their non-relativistic values, we will use this form for the probability amplitude.

Note that our relativistic transformation has given us the *variance* of the amplitude of an event which is now a space-time event, i.e. addition of Δt and $\Delta \vec{r}$. The wave number of the space variables is given in (7.7)

$$k = \frac{\partial}{\partial \vec{r}}; \quad (7.10)$$

so the wavefunction is

$$\psi = \frac{\partial \psi}{\partial \vec{r}} = \frac{\partial \psi}{\partial \vec{r}}. \quad (7.11)$$

This is analogous to saying we have a field operator for waves in with the momenta \vec{p} and \vec{q} . The function was last derived by de Broglie in just this way. For a moving particle, the frequency of the amplitude variations is ω given by

$$\omega = W_b. \quad (7.12)$$

The absolute square of (7.5) is just 1, so if a particle in touch with a definite energy, the probability of finding it in the same everywhere and thus not changing ψ , i.e., that its total angular momentum is everywhere zero. If we recall a real wave wave, the square would vary from point to point, which would not be right.)

We know, of course, that there are situations in which particles move in places in space or time the probability depends on position and frequency with time. How do we describe such a situation? We can do this by considering a superposition which is a superposition of waves with different amplitudes for states of definite energy. We have already discussed this situation in Chapter 4B of Vol. I—over to consideratory simplification! We know that the sum of two amplitudes of different wave functions & their relative magnitude and frequency ω (that is, energy) gives $\sum A_i^2$ because it maps, or beats, us to the square of the amplitude varies with space and time. We also found that these beats come with the so-called “group velocity” given by

$$v_g = \frac{d\omega}{dk},$$

where $d\omega$ and dk are the relative differences between the wave numbers and frequencies for the two waves. For more complicated waves, such as the sum of many oscillations all over the same frequency, the group velocity is

$$v_g = \frac{d\omega}{dk}. \quad (7.13)$$

Taking $\omega = E/\hbar$ and $k = p/\hbar$, we say that

$$v_g = \frac{\delta E}{\delta p}. \quad (7.14)$$

Using Eq. (7.6), we have

$$\frac{dE}{dp} = \epsilon' \frac{p}{M}. \quad (7.15)$$

But $E_p = Mc^2$, so

$$\frac{dE}{dp} = \frac{p}{M}, \quad (7.16)$$

which is just the classical velocity of the particle. Alternatively, if we use the non-relativistic expression, we have

$$\epsilon = \frac{E_p}{p} = mc^2 - p + \frac{p^2}{M}$$

and

$$\frac{d\epsilon}{dp} = \frac{dE_p}{dp} - \frac{1}{M} \left(\frac{p^2}{M^2} \right) = \frac{p}{M}, \quad (7.17)$$

which is again the classical velocity.

The moral of this is that since we have several amplitudes for pure one ψ states or nearly the same energy, there will be some “interference” in the probability that move through space with a velocity equal to the velocity of the classical particle of that energy. We should remark, however, that when we say we can add two amplitudes of different wave numbers together, we must not note that ψ corresponds to a moving particle. We have introduced something new, something that has no analog from the theory of relativity. We will what the amplitude and the particle standing still—but even decoupled—that is, working if the particle were moving. But we cannot derive from these arguments what would happen when there is a moving particle with different speeds. It is enough that we can stop to a limit. So we have added merely the exact hypothesis that not only is it the possible solution, but that there can also be solutions with at least one of the components that the different wave can interfere.

7.3 Potentials energy; energy conservation

Now we would like to discuss what happens when the charge is a particle and moves. We begin by thinking of a particle which moves in a "one-dimensional" way, a particle. We might as first think of a constant potential. Suppose that we have a large metal charge which is connected to a smaller metal potential source (Fig. 7.2). If the two charged objects inside the box, their potential energy will change, why? we will see it, and will be absolutely independent of position. Inside it can be no electric field in the places inside because the same net potential doesn't make any difference of anything going in there. It's a vacuum. Now there is a way we can determine where the charge source is, so we must make a guess. The guess which comes is not a bad one at all you might expect (Fig. 7.3) because we can use the sum of the potentials for every point. Because V_p —which is self, the sum of the important terms, energies— V_p (potential) is proportional to

$$e^{-k|x|} \quad (7.19)$$

Fig. 7.2. A particle of mass m and acceleration a is in a region of constant potential.

the external potential is over the coefficient k^2 , which we have said is to a large extent by the kind of range of the system, defined for "local" energy, plus kinetic energy, plus potential energy:

$$\text{kinetic } E_k = p^2 \quad (7.20)$$

Or, for a one-dimensional situation,

$$\text{kinetic } E_k = m v^2 + \frac{1}{2} k^2 x^2 + V \quad (7.21)$$

Now what about physical phenomena above the box? Is there any effect of external energy states, that tell us just? The amplitude to move due to the external field is

$$e^{-k|x|}$$

over which it would have $V(x)$. $V = 0$. That is just like a charge in the zero voltage field. For instance, rapid phase changes in all configurations, because over before this instant changing, or the probabilities. All the physically important parts are same. We have assumed that we are talking about different sizes of the two charged objects, so they are the same for V . If we want to add charges together, it's going from one state to another, we won't have quite same the result, but consequences of charge potentials, this.

So far, our assumption agrees with what we would expect for a charge of charge reference level. But if it is really $a/2$, it should be in a relative energy that is not just a constant. In general, it could vary in any arbitrary way with both time and space, and the complete result to the amplitude must be given to us in a differential equation. We don't want to go on now with the general case right now, but only want to get some idea and know some things happen so we can think only of a potential that is constant, at least had values very locally important. Then we can think of the comparison between the classical and quantum mechanics.

A typical example of the situation is in Fig. 7.4, which has two boxes, left in the constant potentials x_1 and x_2 , and a region in between where we'll assume that the potential vanishes from one to the other. We imagine e that come from the two constant potentials x_1 and x_2 to only one of the regions. We also assume that the potential is large enough so that in my small region in which there are only two regions, the potential is not constant. We could then think first in one part of the space, the one which might look like (7.19) and be appropriate for that part of the wave.

Let's think of a special case in which $x_1 = 0$, so that the potential energy there is zero, but the voltage is very large so that e could have more energy in the second box. Classically, it might be going faster in the second box. It would take more energy and therefore more acceleration. Let's see how this might come out of quantum mechanics.

With our assumption that V_2 is finite, the first term would be proportional to

$$e^{-i\omega t} \langle 0| \hat{V}_1 \phi(0) | \psi(0) \rangle e^{i\omega t} = 0 \quad (7.11)$$

and the amplitude in the second box would be proportional to

$$e^{-i\omega t} \langle \hat{\psi}(t) | \hat{V}_2 \phi(0) | \psi(0) \rangle e^{i\omega t} \quad (7.12)$$

It's clear that the internal energy is not being changed, our summing must make both regions 1 and 2. The question is: How do the two box amplitudes match together through the region between the boxes?

We are going to assume that the potentials are *quasistatic*, i.e., smaller than no longer be considered static. We will also suppose that the strengths of the two-potential (that is, its passing from one value to another) is *slow*, so we speak, this is nothing at the "medium" that depends on time. Considering the space is changing, we can consider that the wave in one region "depends" on the other waves all over space, which is impossible in the real. In order to use light waves going through materials or moving and change their frequency. While frequencies in (7.11) and (7.12) are the same, we must have that

$$\omega_{in} + \frac{p_1^2}{2m} + V_1 - W = \frac{p_2^2}{2m} + V_2 \quad (7.13)$$

Both sides represent the classical total energies, so Eq. (7.13) is a statement of the conservation of energy. In other words, the dressed potential W has a variation of energy is equivalent to the constant mechanical statement that the frequencies remain the same everywhere. However, if the conditions are not changing with time, it follows with consider that the

In the easiest example case, $V_1 = 0$ and V_2 is negative, Eq. (7.13) gives the dispersion relation for the particle's wavelength in the second box shown in Fig. 7.1. The surfaces of equal phase are shown by the dashed lines in Fig. 7.1. We have also drawn a graph of the real part of the amplitude, which shows again how the wavelength decreases in going from region 1 to region 2. The group velocity of the waves, which is $p/2m$, also increases in the way one would expect from the classical energy conservation, since it is just the same as Eq. (7.21).

There is an interesting special case where V_2 is so large that $V_2 - V_1$ is greater than $p_1^2/2m$. Then p_2^2 , which is given by

$$p_2^2 = 2m \left[\frac{p_1^2}{2m} - V_2 - V_1 \right], \quad (7.14)$$

is negative. This means that p_2 is an imaginary number, $p_2 = ip$. Classically, we would say that the particle never gets into region 2, it doesn't have enough energy to climb the potential V_2 (or, more realistically, however, the amplitude is still given by Eq. (7.22); its space variation is given as

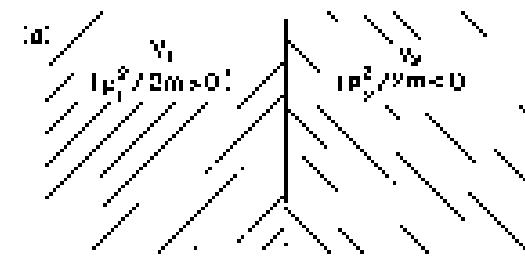
$$e^{-ipx/m}.$$

But if p_2 is imaginary, the spatial pattern becomes a real exponential. Say that the particle was initially going in the +x direction; then the amplitude would look like

$$e^{-ipx/m}. \quad (7.15)$$

The amplitude decreases rapidly as x increases.

Imagine that the two regions of different potentials were very close together so that the potential energy changed suddenly from V_1 to V_2 , as shown in Fig. 7.2(a). If we plot the real part of the probability amplitude, we get the dependence shown in part (b) of the figure. The wave in the first region corresponds to a particle trying to get into the second region, but the amplitude there has an



1

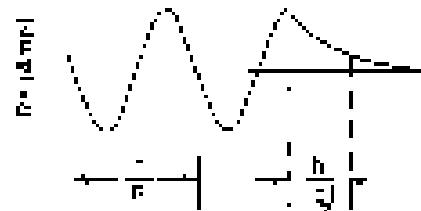
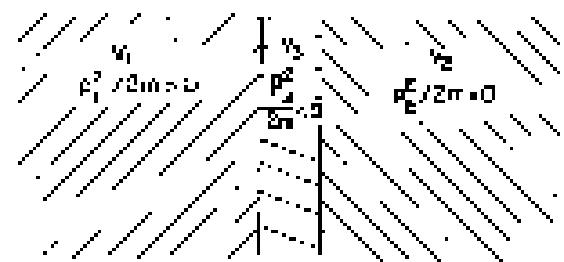


Fig. 7: The amplitude spectrum of a wave in a dielectric resonator periodic structure.



rapidly. There is some chance that it will be informed in his second reply—where he could have you physically—but the importance is very small, except light must be boundary—the situation is very much like what we found for normal *internal* reflection of light. The light doesn't normally get out, but we can observe it. The question is, what is a wave length of one of the pulses?

You will remember that in the previous section that after an encounter where light was totally reflected, we could get some light transmitted into the second piece of material. This is surprising, but it happens to particles in quantum mechanics too. If there is a透射 region with a potential V , we find that the reflected kinetic energy would be negative. The particle would therefore never get past. But quantum mechanically, the wavefunction decaying amplitudes can never vanish. So again we get a non-zero probability that the particle will be found at another site where the kinetic energy is again positive. The situation is illustrated in Fig. 7.5. This effect is called the «quantum mechanical transmission of a hole».

The Coulomb potential energy of a curvilinear wavepacket amplitude gives the semi-classical description of the electrostatic energy of a nuclear nucleus. The potential energy of a nucleus has as a function of the distance from the center is shown in Fig. 1-1(a). If one tries to knock a particle with the energy E and the mass m , it would feel an electrostatic repulsion from the nuclear charge e and would classically get reflected from the distance r_0 , where the total energy is equal to the potential energy E . Consequently, the potential energy E is much lower because of the strong attraction of the Coulombic nuclear forces. However it takes that to radiate away we find a particle which energy E is outside the nucleus coming out with the energy E^* . One may say that out with the energy E within the nucleus and then it escape through the potential barrier. The probability amplitude is roughly as sketched in part (b) of Fig. 1-1, although usually the exponential decay is much faster than shown. It is in fact quite remarkable that the mean life of an alpha particle in a uranium nucleus is as long as 4.5 billion years, when the rate of exponential decrease in amplitude is nearly exponential (10^{-24} per sec) ... we can imagine something like 10²⁴ years from 10^{-24} sec? The answer is that the exponential gives the exponentially small factor of e^{-E/E_0} , which gives the very small enough defiance, and life of decay. Once the α -particle is in the nucleus, there is almost no amplitude to get the particle outside. However, it can take many nuclei and over long enough, you may release and find out all this information.

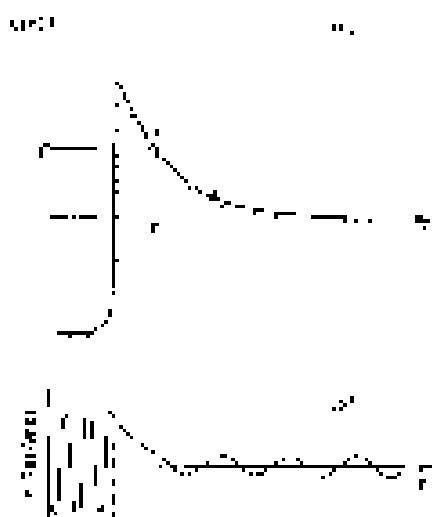


Fig. 7. G. (c). The potential function for $m = \infty$ plotted in the (α, β) -plane. (c) The evolution through the probability amplitude.

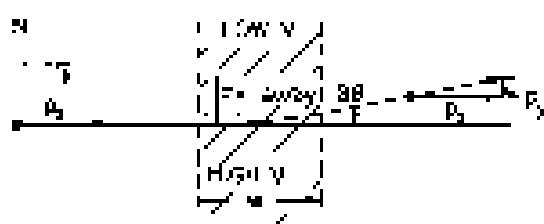


Fig. 7-3. The deflection of a particle by a transverse potential gradient.



Fig. 7-4. The probability amplitude in a region with a transverse potential gradient.

7.4 Forces; the classical limit

Suppose that we have a particle moving along and passing through a region where there is a potential that varies transversely in the x -direction. Classically, we would describe the situation by something like in Fig. 7-3. If the particle is moving along the x -direction and enters a region where there is a potential that varies with y , the particle will feel a transverse acceleration. From the force $F = -\nabla V$, since the force is present only in a limited region of width a , the force will act only for that time. The particle will be given the transverse momentum

$$\vec{p}_y = F \frac{a}{2} \hat{y}$$

The angle of deflection θ is then

$$\theta = \frac{p_y}{p_0} = \frac{F a}{m p_0},$$

where p_0 is the initial momentum. Using $-dV/dx$ for F , we get

$$\theta = -\frac{e}{p_0} \frac{\partial V}{\partial x}. \quad (7.26)$$

Let us now go to see if our idea that the waves go as (7.26) will explain our one result. We look at the same thing again, but semiclassically, assuming that everything is on a very large scale compared with a wavelength or more generally conditions. In any small region Δx we can say that the amplitude varies as

$$A(x) = A_0 e^{i k x / \hbar}, \quad (7.27)$$

Can we see that this will change due to the deflection of the wave? When V has a transverse gradient $\partial V/\partial x$, we have sketched in Fig. 7-3 what the waves of transversely amplitude $A(x)$ look like. We have drawn several "wave nodes" which you can think of as surfaces where the phase of the amplitude is zero. In every small region, the wave length $\lambda = h/p$, distance between successive nodes,

$$\lambda = \frac{h}{p},$$

where p is related to V through

$$p^2 + \frac{h^2}{4m^2} + V = \text{const} \quad (7.28)$$

In the region where V is larger, p is smaller, and the wave length longer. So the angle of the wave must gets changed as shown in the figure.

To the net change in angle of the wave nodes we notice that for the two nodes shown in Fig. 7-3 there is a difference of potential $\Delta V = (eV_1 - eV_2)/2$, so there is a difference Δp in the momentum along the two routes which goes to

obtained from (7.38):

$$e \left(\frac{d\phi}{dx} \right) = \frac{\partial \phi}{\partial x} = -\omega_0, \quad (7.39)$$

The last, in other words, is the slope difference along the x -path, which means that the particle is moving at a different rate. The difference in the rate of travel of plane is $\Delta\omega = \omega_0/\lambda$, to my accumulated phase difference in the total distance Δx :

$$\Delta(\text{phase}) = \Delta x \cdot \omega_0 = \frac{\Delta x}{\lambda} \cdot \omega_0 = -\frac{M}{\mu^2} \Delta E \cdot \omega_0. \quad (7.40)$$

This is the amount by which the phase $\omega_0 \cdot \text{path}$ is "tilted" off the plane $\omega_0 \cdot \text{path}$ over the wave vector ω_0 strip. But a shift in θ due to a phase difference of this same $\Delta\omega$ corresponds to the wave number change Δk (from k_0):

$$\Delta k = \frac{2\pi}{\lambda} \Delta(\text{phase}) = \frac{2\pi}{\lambda} \Delta(\text{phase})$$

or

$$\Delta k = -\frac{M}{\mu^2} \Delta E \cdot \omega_0. \quad (7.41)$$

Returning to Fig. 7.20, we see that the new wave energy $\omega_0 + \Delta\omega$ is obtained at the cost of

$$\Delta E = \hbar \Delta k^2, \quad (7.42)$$

so we have

$$\Delta E = -\frac{M}{\mu^2} \Delta E^2 \cdot \omega_0. \quad (7.43)$$

This is identical to Eq. (7.38), if we replace ω_0 by $\omega_0 + \Delta E/\hbar$.

The previous result can be summarized like this: particle motion and field motion affect each other in a linear fashion. We have shown that an external field does not affect the particle motion unless $E = \infty$, provided we assume that a particle can't have a distance to the center of the magnetic field equal to $R/2$. In the "resisted" case, the quantum mechanics will agree with the classical mechanics.

7.6 The "probability" of a spin one-half particle

Before we begin discussing something special about the potential $V(x)$ it is just that energy E is conservative, i.e., gives a force. For example, in the Stern-Gerlach experiment we had the energy $V = -m_B B$ which was due to the B field's spin-orbit coupling. If we wanted to prove a quantum connection between $E = 0$, we would have said that the particle in one state had more energy than another, and vice versa. This is the state which had an opposing energy variation. (We could call the magnetic energy V or the potential energy V or it is the "disorder" energy. It's all the same.) Because of the energy variation, the wave function ψ would be deformed and the beams not bearing or shear. (We see now that quantum mechanics would give us the same setting as we saw in classical form, the classical mechanics.)

From the dependence of the amplitude on potential energy we can deduce that if a particle ψ in a uniform magnet is left alone, the ψ oscillates, its probability amplitude $|\psi|$ being in phase with time according to

$$|\psi| = |\psi_0| \cos(\omega t + \phi_0).$$

We can consider that this is a *relative* definition of ω . In other words, the photon particle in a uniform field B for a time t is a *relatively* compact, it will be multiplied by

$$\psi \sim \psi_0 \cos(\omega t + \phi_0)$$

over value it would be $\frac{1}{2}$ in field. Since for a spin one-half particle, ψ_0 can be either plus or minus some number, say a , the two possible states in a uniform field would have their phases changing at the same rate but in opposite directions. The wave amplitudes are multiplied by

$$e^{\pm i\theta \cdot \vec{\sigma} \cdot \vec{B}}. \quad (7.34)$$

The result becomes interestingly cumbersome. Suppose we take a spin one-half particle in some state that is not purely spin up or spin down. We can imagine it would like in terms of the spin states to be in the p_{+} up and p_{-} down states but in a magnetic field, these two states will have phases changing at different rates. So if we ask some question about the amplitudes, the answer will depend on how long it has been in the field.

As an example, we consider the disintegration of the muon in a magnetic field. When muons are produced as disintegration products of mesons, they are polarized in the e^- direction; they have a positive spin component. The muons, in turn, disintegrate with about 92% in one direction for the average, with approximately equal probabilities.

$$\mu^- \rightarrow e^- + \nu + \pi^-$$

In this disintegration, it turns out that (in at least the simplest energies) the electrons are emitted primarily in the direction opposite to the original direction of the muon.

Suppose then that we consider the experimental arrangement shown in Fig. 7.3. It polarized muons come from the left and are brought to rest in a block of material so that they will, a little while later, decay spontaneously. The electrons emitted will, in general, go out in all possible directions. Suppose however the beam passes through the following block and with other spins in the e^- direction. Without a magnetic field there would be some angular distribution of these directions, we would like to know how this distribution is changed by the presence of the magnetic field. We expect that it may vary in some way with time. We expect also this happens by asking for any time t , what the amplitude is that the muon will be found to be $(-)$ state.

We can state the problem in the following way: A muon is known to have its spin in the $-z$ direction at $t = 0$, what is the amplitude that it will be in the same state at $t = t$? Now we do not know exactly what the behavior of a spin one-half particle in a magnetic field ought to be, but we do know that it depends on the spin component spin down states are, in general, in the field, their amplitude goes down by the factor $e^{-i\theta \cdot \vec{\sigma} \cdot \vec{B}}$. Our procedure here is to choose the representation in which the two states are spin up and spin down with respect to the e^- direction (the field direction). Any question can then be expressed with reference to the e^- spin basis we chose above.

Let's say that $\psi(0)$ represents the muon state. When it leaves the block at $t = 0$ its state is $\psi(0)$, and we want to know what it is at the later time t . If we represent the two new states by $|+z\rangle$ and $| -z\rangle$ we know the two amplitudes $C_+(t)$ and $C_-(t)$ —we know these amplitudes because we know that $\psi(0)$ represents a pure spin up state in the e^- state. From the results of the last section, these amplitudes are:

$$|+z|\psi(0) = C_+ = \frac{1}{\sqrt{2}}$$

and

$$|-z|\psi(0) = C_- = \frac{i}{\sqrt{2}}.$$
(7.35)

$$|-z|\psi(t) = C_- = \frac{i}{\sqrt{2}}$$

They happen to be equal. Since these amplitudes refer to the expectation at $r = 0$, let's call them $C_+(0)$ and $C_-(0)$.

If you studied Chapter 6, you are not too far behind me in understanding. For now, we will just take it. In Chapter 10, we have another class of questions involving induction of these amplitudes.

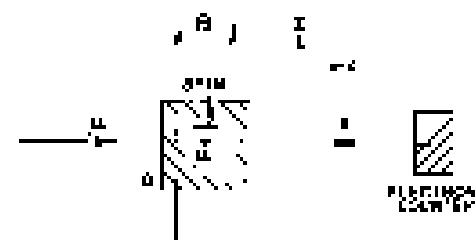


Fig. 7.3. A scattering experiment.

Now we know what happens to these test amplitudes with time. Using (13), we have

$$C_1(t) = C_1(0)e^{-\lambda t} \quad (2.20)$$

3

$$S_{\text{eff}} = S \exp^{-\gamma_0 - \alpha t}$$

But if we know \mathcal{C} at time $t = 0$, we have at best \mathcal{C} in terms of what the condition it is in $t = 0$ or y_0 must be, that when we now do know is the probability that \mathcal{C} at the origin will be y_{t+1} + y_{t+2} + \dots . Our general rule can, however, be a case of this restriction. We will then have difficulty with the i systems \mathcal{C}_i for which $y_i(t+2) \neq 0$, i.e.

1

$$A_4(0) = 1 - \rho S_2(0) + \rho^2 S_4(0) - \rho^3 S_6(0) \quad (7.30)$$

Again, using the results of the last chapter, we need the equation $\phi = \phi(\lambda)$ to get λ^k from λ . This gives us the final rule:

$$(1 + \frac{1}{n})^n = e$$

So we know all the parameters in Eq. (7.37). We get

$$f(x) = x^{(n+1)/n} = x^{1 + \frac{1}{n}}$$

1

$$A_4(\eta) = \eta^{\frac{2}{3}} e^{-\frac{2\pi i}{3}\eta}$$

A somewhat simpler result: Notice that the sum of $\dots + m$ is ℓ . While we can't force $\ell \in G$, we get $\ell \in \langle G \rangle = \{1\}$, which is right because we assumed that the column zero in C_1 is equal to $1 \in G$.

the probability P_4 that the mean will be found in the $(1.5\sigma, 2\sigma)$ range is 0.45.

$$f \sim \propto e^{-\frac{R}{L}}$$

The probability distributions shown in Fig. 7.10 show that the probability of obtaining a value of x is given by $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. However, we have already seen that the cumulative probability $P(X < x)$ is given by $\Phi(\frac{x-\mu}{\sigma})$.

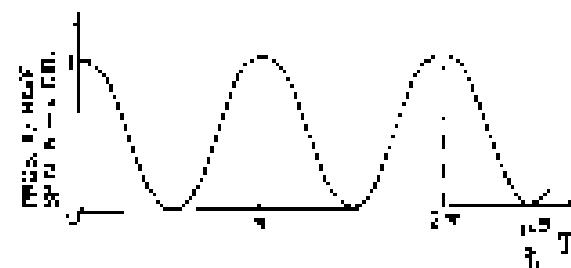


Fig. 7. (A) New description of the previously described *Leptothrix* particle, will be published in *Journal of the American Medical Association*.

Thus, we find that the change of reaction time does not have much influence on the conversion. The yields gradually drop the length of time the reaction has been sitting in the magnetic field. In five cases, 26.8% of the magnetite remains.

We can do better for the same memory storage at any other ϵ -tolerance. In fact the problem above, for example, has uses the theory of stochastic geometry and the

In the oscillations, the wave function has oscillated at half-angle to the field it precesses against. If you back out the amplitude to be in the (+) state values of $\cos^2(\theta/2)$, which oscillates with the same period but reaches its extrema at different times than $\sin^2(\theta/2)$. For what's happening to the components of the wavefunction under successive operations which correspond to successive precessions, imagine it's a pendulum, slowly rotating around the axis. You can see this by saying that the state is precessing at the frequency

$$\omega_s = \frac{\omega_0}{2} \quad (1.38)$$

You can begin to see the logic that the quantum-wavefunction description will be where we are describing how the wavefunction is in a

The Hamiltonian Matrix

8-1 Amplitudes and vectors

Before we begin the main topic of the chapter, we would like to comment on our last bit of mathematical basis that is used a lot in the literature of quantum mechanics. Knowing this will make it easier for you to understand the basis of papers on the subject. Let's first look at the case of one-dimensional systems to review the equations of motion, expectation values and those of the scalar product of two vectors. You remember that if \mathbf{x} and \mathbf{y} are two states, it is possible to state $\langle \mathbf{x} | \mathbf{y} \rangle$ and end up to have it written as a sum over a complete set of base states \mathbf{z} of the amplitude to go from \mathbf{x} into one of the base states and then from that state end up to \mathbf{y} :

$$\langle \mathbf{x} | \mathbf{y} \rangle = \sum_{\mathbf{z}} \langle \mathbf{x} | \mathbf{z} \rangle \langle \mathbf{z} | \mathbf{y} \rangle. \quad (8.1)$$

We explained this in terms of a Stern-Gerlach apparatus, but we remind you that this is just a model to help the apparatus. Equation (8.1) is a fundamental law that is true as long as we put the filtering equipment — or not — it is not necessary to imagine that the apparatus is there — we can think of it simply as a function of the amplitude $\langle \mathbf{x} | \mathbf{y} \rangle$.

We would like to comment on Eqs. (7.1) and (7.2) for calculating dot products of two vectors. Since $A = \langle \mathbf{B} | \mathbf{C} \rangle$ and C is a "free" vector in three dimensions, we can write the dot product like this:

$$\sum_i \langle \mathbf{B} | \mathbf{e}_i \rangle \langle \mathbf{e}_i | \mathbf{C} \rangle, \quad (8.2)$$

with the understanding that the symbol \mathbf{e}_i stands for the *i*-th unit vector in the x , y , and z directions. Since $B = \mathbf{B} | \mathbf{e}_i \rangle \langle \mathbf{e}_i | C$, what we ordinarily call B_x ; B_y ; B_z is really $\langle \mathbf{B} | \mathbf{e}_x \rangle \langle \mathbf{e}_x | \mathbf{C} \rangle$; $\langle \mathbf{B} | \mathbf{e}_y \rangle \langle \mathbf{e}_y | \mathbf{C} \rangle$; $\langle \mathbf{B} | \mathbf{e}_z \rangle \langle \mathbf{e}_z | \mathbf{C} \rangle$ respectively.

$$\langle \mathbf{B} | \mathbf{e}_x \rangle \langle \mathbf{e}_x | \mathbf{C} \rangle,$$

which is the dot product $B_x C$.

Comparing Eqs. (8.1) and (8.2), we can see the following analogy: The state \mathbf{x} and \mathbf{y} correspond to the two vectors \mathbf{A} and \mathbf{B} . The base state correspond to the basis vectors \mathbf{e}_i , while we ignore other vectors. Any vector can be represented as a linear combination of the three "base vectors", i.e., Furthermore, if you know the components of each "base vector" in its combination, that is for free information — you know everything about a vector. In a similar way, any quantum system state can be described completely by the amplitude $\langle \mathbf{x} | \mathbf{z} \rangle$ to go into the base state \mathbf{z} , and if you know the amplitudes, you know everything you can know about the state. Because of this close analogy, what we have called a "state" is often also called a "base vector".

Since the base vectors \mathbf{e}_i are all of right angles, we have the relation

$$\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}. \quad (8.3)$$

This corresponds to the rule given in (7.12) defining the base states:

$$\langle \mathbf{e}_i | \mathbf{e}_j \rangle = \delta_{ij}. \quad (8.4)$$

You see now why one says that the base states \mathbf{e}_i are all "orthogonal".

8-1 Amplitudes and vectors

8-2 Revising wave packets

8-3 What are the base states of the world?

8-4 Time states change with time

8-5 The Hamiltonian analysis

8-6 The ammonia molecule

Review: Chapter 4D, V, 1. Modes

There is one minor difference between Eq. (8.1) and the dot product (8.2), how that

$$(\phi | \psi) = (\phi | \psi)^* \quad (8.5)$$

But in vector algebra,

$$\alpha \cdot B = B \cdot \alpha$$

With complex numbers of quantum mechanics we have to keep a slight "order of terms" whereas in the dot product the order doesn't matter.

Now consider the following vector equation:

$$A = \sum_i c_i a_i \quad (8.6)$$

It's a little unusual, but correct. It means the same thing as

$$A = \sum_i c_i a_i = c_1 a_1 + c_2 a_2 + \dots + c_n a_n \quad (8.7)$$

Notice, though, that Eq. (8.8) involves a quantity which is different from a scalar. A dot product is a scalar, whereas Eq. (8.8) is a sum of vectors. One of the great uses of vector analysis is to distance away from the cumbersome notation of a vector itself. One might be slightly inclined to suspect a thing that is far simpler and "natural" from the common mechanical formulae of (8.1) and (8.2) — indeed. We remove the $(\phi |$ from both sides Eq. (8.8) and write the following equation (but just暂且 wait, it's just a notation and no law missing you will find out what he symbol means)

$$|\phi| = \sum_i |\phi| c_i |a_i| \quad (8.8)$$

One thinks of the bracket $(\phi | \phi)$ as being divided into two pieces. The second piece $(\phi |)$ is often called a dot, and the first piece $(\phi | \phi)$ is called a bar (just to clarify, they have a "dot-bar" or "dot-bar" property by (8.8)) the sub-operators $(\phi |)$ and $|a_i|$ are called dual unit vectors. In any case, they are not numbers and, in general, we use the results of calculations to obtain numbers, so such "unphysical" quantities are only part-way steps in our calculations.

It happens, in fact, that we have written all our results in terms of numbers. Now how do my images to work, now (8.8)? It is amazing to note that even in such very vector algebra we could make all equations involve only numbers. For instance, instead of a vector equation like

$$F = m a,$$

we could always have written

$$G \cdot F = G \cdot (m a)$$

We could then an equation between the products that is true for any values G . But if it is true for any G , it really makes sense at all to keep writing the G ?

Now look at Eq. (8.8). It is an equation that is true for any ϕ . So for consistency, we should just leave the ϕ and write Eq. (8.8) instead. It is true, once it is written provided certain conditions that it has to always be "frozen" by "writing up" or "left by" — which simply means removing some $(\phi |$ on both sides. So Eq. (8.8) means exactly the same thing as Eq. (8.1) — we must believe. When you want numbers, sum up in the $(\phi |$ part.

At this you have already come across the π in Eq. (8.8). Since we know that π is the identity a , why do we keep it? Because this suggests that the ϕ does nothing as well as it should never, so that we have only

$$= \sum_i \phi c_i |a_i| \quad (8.9)$$

And this is the great law of quantum mechanics! There is no magic in vector analysis. It says that if you put ϕ in the ϕ and c_i in the c_i of right of both sides you get back Eq. (8.1). It is really very useful, and it's a nice reminder that the equation is true for any two things.

8.2 Building state vectors

In the form of Eq. (8.1) you've been thinking of ψ in the following way: Any wavevector $|\psi\rangle$ can be represented as a linear combination with suitable coefficients of a set of basis "vectors" $|e_i\rangle$, if you prefer, as a superposition of "unit vectors" $|\phi_i\rangle$ which you can choose. Then you see that the coefficients of $|\psi\rangle$ are just ordinary complex numbers, suppose we write

$$\langle \psi | e_i \rangle = C_i.$$

Then Eq. (8.2) is the same as

$$|\psi\rangle = \sum_i |\phi_i\rangle C_i. \quad (8.10)$$

We can repeat this equation, taking other state vector, say $|\phi\rangle$, $\in \mathcal{H}^{\text{out}}$ and some complex coefficients B_i . Then we have

$$|\phi\rangle = \sum_i |\phi_i\rangle B_i. \quad (8.11)$$

The B_i 's are just the amplitudes $\langle \phi | \phi_i \rangle$.

Suppose we had started by substituting the problem $\langle \phi | \psi \rangle$ in Eq. (8.1) with $\langle \phi | \phi \rangle$. We would have had

$$\langle \phi | \psi \rangle = \sum_i \langle \phi | \phi_i \rangle C_i. \quad (8.12)$$

Remembering that $\langle \phi | \phi_i \rangle = \delta_{i\phi}$, we can rewrite this as

$$\langle \phi | \psi \rangle = \sum_i \delta_{i\phi}^* C_i. \quad (8.13)$$

Note the interesting point is that we can do nothing in Eq. (8.13) and get $\langle \phi | \psi \rangle$ without $|\psi\rangle$ or $|\phi\rangle$. Other words, there is no trace of the interaction in this, because they are quite distinct in the two expressions. Let's fix notation. Let $\langle \phi | \psi \rangle$ be

$$\langle \phi | \psi \rangle = \sum_i \delta_{i\phi}^* C_i,$$

which changes nothing. Then putting it together with Eq. (8.10), we have

$$\langle \phi | \psi \rangle = \sum_i \delta_{i\phi}^* \langle \phi | \phi_i \rangle C_i. \quad (8.14)$$

Remember though, that $\langle \phi | \phi_i \rangle = \lambda_i$, so that in the sum we can replace only the indices with $i = 1, \dots, p$:

$$\langle \phi | \psi \rangle = \sum_i \lambda_i^* C_i. \quad (8.15)$$

Here, of course, $\lambda_i^* = \langle \phi | \phi_i \rangle = \lambda_i \langle \phi | \phi_i \rangle$, and $C_i = \langle \phi | \phi_i \rangle C_i$. Again we see the close analogy with the dot product

$$\vec{A} \cdot \vec{B} = \sum_i A_i B_i.$$

The only difference is the complex conjugate on B_i . So Eq. (8.15) expresses the state vector $|\psi\rangle$ as a linear combination in terms of the basis vectors $|\phi_i\rangle$ of \mathcal{H}^{out} . Its amplitude to go from $|\phi\rangle$ to $|\psi\rangle$ is given by the last of the products in Eq. (8.15). This equation is, of course, just like (8.1) written with different symbols. So we have just gone from a single line to get used to the new symbols.

We should perhaps emphasize again the following subtle subtlety: when one uses the description in terms of these orthogonal unit vectors, the basis elements $|\phi_i\rangle$ of the space in mathematical terms, one must, among all the complete set applicable to any particular problem, designate for the situation, two, or three or four, or a definite number of basis states only be involved.

We never also to find them $|\phi_i\rangle$. They are often partially or fully orthogonal. If we start the particles out in a certain state, we then want the "overlap"

an upward or, and afterward make a measurement to see if they are in state ψ . The test A is described by the amplitude

$$\langle \psi | A | \phi \rangle. \quad (8.16)$$

Such a symbol doesn't have a meaning by itself, but it is easy to make sense of it, but the analogy is not particularly useful. We saw in Chapter 5, Eq. (5.12), that we could write (S.16) as

$$\langle \psi | A | \phi \rangle = \sum_i c_i \langle \phi | A_i | \psi \rangle. \quad (8.17)$$

This is just an example of the **addition rule** Eq. (5.1), used twice.

We also found that, from the definition, R associated in series with A , from we could write

$$\langle \psi | R | \phi \rangle = \sum_i c_i \langle \phi | R_i | \psi \rangle. \quad (8.18)$$

Again, this follows directly from Dirac's method of writing Eq. (8.1) — remember that we can always choose basis i , which is just like the choice i , between ϕ and ψ . In addition, we can think of Eq. (8.17) in a similar way. Suppose we think of the A amplitude as appearing in the i -th column (or i -th row) of A in the state ψ ("left"). In other words, we can take successive rows (or columns) from both R and the amplitude A performing R on ψ always identifies and everywhere the same as the amplitude $\langle \psi | A | \phi \rangle$. This answer is no. We want Eq. (8.17) to reproduce it:

$$\langle \psi | R = \sum_i c_i \langle \phi | R_i | \psi \rangle. \quad (8.19)$$

We can clearly do this if

$$\langle \psi | R = \sum_i \langle \psi | A_i | \phi \rangle R_i | \phi \rangle = \langle \psi | A | \phi \rangle. \quad (8.20)$$

With these values of ψ , "But it doesn't determine ψ ," you say; "it only determines $\langle \psi | \phi \rangle$." However, $\langle \psi | \phi \rangle$ does determine ψ , because if you know all the coefficients that map ϕ to the base states i , then ψ is uniquely defined. In fact, we can play with our notation and we get the last form of Eq. (8.20); i.e.,

$$\langle \psi | R = \sum_i \langle \psi | P_i Q_i | \phi \rangle. \quad (8.21)$$

Then, since this equation is true for all ϕ , we can write simply

$$|X\rangle = \sum_i |P_i Q_i | \psi \rangle. \quad (8.22)$$

Then we can say: "This state X is what we get if we want ψ to go through the apparatus A ."

Our final example of the power of the bra. We start again with Eq. (8.17). Since this is true for any ϕ and ψ , we can drop them both. We can get

$$A = \sum_i \langle \phi | A_i | \psi \rangle. \quad (8.23)$$

What does it mean? It means to invert $| \phi \rangle$, then ϕ^* at time t_0 ; if you put back the ϕ and ψ . As it stands, "bra" is "ugly" and often somewhat incomplete. It would probably fit more the left "by" $| \phi \rangle$ if necessary.

$$| \phi \rangle = \sum_i | \phi_i \rangle \phi_i^* \langle \phi_i | \psi \rangle. \quad (8.24)$$

² You might think we should bring $| \phi \rangle$ instead of ϕ^* . But then we will have to be careful for "absolute value of ϕ " as the ϕ is usually complex. In general, the "bra" ϕ^* behaves much like the function ϕ .

which is just Eq. (5.23) all over again. In fact, we could have just stopped at $\langle \psi | \hat{A} | \phi \rangle$ since that equation was written:

$$|\psi\rangle = A |\phi\rangle. \quad (5.23)$$

The symbol A is generic or semiblurry, so it's easier to see what it is: a new kind of thing called an operator. It's something which you can act on a state to produce another one. Accordingly, Eq. (5) says that $\langle \psi | \hat{A} | \phi \rangle$ is the same as $\langle \psi | A | \phi \rangle$. And $A|\phi\rangle$ is still to be evaluated until it is compared with some fixed $\langle \psi |$ to give

$$\langle \psi | \phi \rangle = \langle \psi | A | \phi \rangle. \quad (5.24)$$

The operator A is often described as having "first class" status, or "mathematical" status, or "of first-class" status, or whatever. In terms of any set of coordinates,

we have *really* added nothing new to Eq. (5) of this new mathematical machinery. One reason for bringing it all up was to show you the sort of a blurry piece of reasoning, because in most books you will find the equations written in the literature being very careful to insist that ψ has to be normalized when you compare it to the ϕ . If you prefer, you can always add the missing pieces to make an equivalent, unnormalized, that will look like something more "fun" to.

After all, you will say, the "fun" and "useful" distinction is a very subjective one. For one thing, we can never know for identify a state by giving its state vector. What we want to do is to state of definite momenta, position, etc., "the state $|\psi\rangle$ ", and we may speak of such a library card, etc. For consistency, we will always use the *useful* $|\psi\rangle$, to identify a state. (It is, of course, an *arbitrary* choice, we could equally well have chosen to use the ψ to.)

5-3 What are the base states of the world?

We have assumed that any state in the world can be represented as a superposition of linear combinations with suitable coefficients of base states. You can ask, first of all, what base states? Well, there are many different possibilities, because it is always possible to change the basis of representation. If you can think of many, many different representations, then the analogs of much different "quantum numbers" you can use to represent different realities. Next, what coefficients? Well, then depending on the physical circumstances, different sets of coefficients correspond to different physical conditions. This important thing to know about is the "right" ψ to obtain you the ϕ which you like. What this base states mean physically. So here that things you have to know about, in particular, you have come across. Then you can understand how to describe similar lectures on these base states.

We would like to look ahead a little and ask, in addition, what the ψ and our ϕ are in order to have a non-zero value as going to between bases of the two different kinds of physics anyway. That, in particular, a particular representation for the base states, called the *representations*, are always possible. For example, for spin-one-half particles, we can use the σ 's and m 's values as the basis for the ψ 's. But there's nothing special about the σ 's; we can take any other basis we like. For consistency σ 's always pay off nicely, however. Suppose we begin with a situation with $1/2$, one electron. In addition to our previous basis for the spin σ 's, there's also something else there called the *momentum* \vec{p} (momentum). We pick a set of base states, each corresponding to one value of the momentum. What if the electron does not have a definite momentum? That's all right, when you say when the basis states are. If the electron base states define a probability of its wave amplitude, unless one chance, an and another amplitude to have either a momentum, add to ψ . And if it is not necessarily added up, it has some amplitude to be summed up according to this distribution, and some amplitude to be spinning around going to the momentum and so on. The same for a collection of n electrons, so far as we know, requires only that the basis, the ψ be describable by the wavefunction and the ϕ 's. So you suppose a set of base states $|\psi\rangle$ for a single electron, with different values of the momentum and

whether the spin is up or down. Different mixtures of amplitudes—here a different combination of the C's describes the state's character. What, in a particular electron's string's described by today, with what amplitude it has an up-spin or a down-spin, and one momentum or another, for all possible momenta? So you can see what's involved in a complete quantum mechanical description of a single electron.

What about systems with more than one electron? Then the base states get more complicated. Let's suppose that we have two electrons. We have, first of all, four possible states with respect to spin: both electrons spinning up, the first one down and the second one up, the first one up and the second one down, or both down. Also we have to specify that the first electron has the momentum p_1 , and the second electron, the momentum p_2 . The base states for two electrons require the specification of two momenta and two spin characters. With seven characters, we have to specify seven of each.

If we have a proton and an electron, we have to specify the spin-direction of the proton, and its momentum, and the spin-direction of the electron, and its momentum. At least that's approximately true. We do not specify how, until the newest representation is for the world. It would very well be start out, by supposing that if you specify the spin in the electron and its momentum, and likewise for a proton, you will have the base states; but what about the "guts" of the proton? Let's look at this way. In a hydrogen atom which has one proton and one electron, we have many different base states to describe up and down spins of the proton and electron and the various possible momenta of the proton and electron. Then there are different combinations of amplitudes C , which together describe the character of the hydrogen atom in different states. But suppose we look at the whole hydrogen atom as a "particle." If we didn't know that the hydrogen atom was made out of a proton and an electron, we might have started out and said: "Well, I know what the base states are—they correspond to a particular momentum of the hydrogen atom." No, because the hydrogen atom has internal parts. It may, therefore, have various states of different internal energy, and describing the real nature requires more than.

The question is: Does a proton have internal parts? Do we have to describe a proton by giving all possible states of protons, and neutrons, and strange particles? We don't know. And even though we suppose that the electron is simple, so that, as I've said, about it is its momentum and its spin, maybe tomorrow we will discover that the electron also has internal parts. It would mean that our representation is incomplete, or wrong, or approximate. In the same way that a representation of the hydrogen atom which describes only its momentum would be incomplete, because it disregarded the fact that the hydrogen atom could have become excited earlier. If an electron could become excited inside and turn into something else like, for instance, a muon, then it would be described not just by giving the states of the new particle, but presumably in terms of some more complicated internal levels. The main problem in the *physics of fundamental particles* today is to discover what are the correct representations for the description of nature. At the present time, we guess that for the electron it is enough to specify its momentum and spin. We also guess that there is an identified proton which has its mass, and its momentum, and so on, that all have to be specified! Several other particles—let's say! The question of what is a fundamental particle and what is not a fundamental particle is a subject you hear so much about these days. Is the question of what is the final representation going to look like in the ultimate quantum mechanical description of the world? Will the electron's momentum still be the right thing with which to describe nature? Or even, should the whole question about that way at all! This question must always come up in any scientific investigation. At any rate, we set a problem: how to find a representation. We don't know, ourselves. We don't even know whether we have the right problem, but if we do, we must first attempt to find out whether any particular particle carries its "fundamentals," or not.

In the non-relativistic quantum mechanics, if the energies are not too high, we don't need to distinguish between wave functions of the strange particles and so forth—

you can do a good job without worrying about these details. You can just decide to specify the number n and spin σ of the electrons and if you do this then everything will be all right. In this case the last condition and cross low-energy hypothesis, and so goes on in the manner they don't get violated. Furthermore, the hydrogen ion is moving slowly and trapping slowly by some other hydrogen atoms, never passing them. In fact, it's colliding with anything something else, so that but always always in the ground state of course. In practice, when you make the first approximation—which you may choose to do by saying that the electron is a point, and no worry about the size. But if you do something good, this will be a much approximation as long as the atomic size is small, and collision is well below its electron size. In such a system you have the approximation of a different potential there. We will start by making an approximation in which we do not include the possibility of interaction, then we discussing the number of defects that will arise to you, in these cases. Of course, we know in some phenomena which wave appears usually at what large energy; but by making our approximation we can simply why much becomes of physical objects. For example, when discussing the collision of two hydrogen atoms, there was an energy exchange process—within the atom, then the atomic nuclei will be excited. To summarize, then, what we can neglect the effect of any forces, and also the size of the particle we can choose a better solution, the choice of infinite momentum and component of angular momentum.

One problem that is describing nature is to find a suitable representation for the free states. But that's only the beginning. This is how to build up a theory "apparatus." It will have to "work out" of the wave equation problem, we would like to make the condition of a stable motion. So we should have to find the laws that determine how things change with time. We now ask ourselves is this second part of the framework of quantum mechanics how shows changes with time?

4.4. Free wave theory with time

We have already talked about free wave propagation, a situation with an propagating, a matter waves. Now we begin to talk about "free states." A consider is freely moving of a few minutes, that is, you propagate a state, and you leave your subject, you continue. Then you let sit, want to look, don't let me go, it tick. In general of the physical situations in the world, for any case, suppose he would like and you let this object still from him, if he finds it. Suppose that it is for you a real opportunity in the z -dimension to do, and it has enough air "apparatus," you can "apparatus" consists of just a dynamical. To ring bell, or, since things could be going, permanent forces applied, a other something—on, but something is happens. At the end of the day, I think it's hard the thing in turn to do, it's no longer existing the wave is, you will have been to heat the wave, since "spinning" is just a special case of "apparatus," we can discuss what "wave" can be going off together with the wave function Eq. (4.17). Because the expression of "waving" is exactly equivalent with ψ , we'll write $\psi(t, \vec{r}, \vec{p})$. The simplest we can, is

$$\psi = C\psi_0 e^{iEt/\hbar} e^{-ipz/\hbar} \quad (4.27)$$

After any other such analysis, it can be represented in some basis (sum or product) by writing as

$$\sum_{\vec{k}} \psi(\vec{k}) \psi^*(\vec{k}) \psi_0 e^{iEkt/\hbar} e^{-ipkz/\hbar} \quad (4.28)$$

From this we obtain described by going into a linear of amplitudes. One must be

$$\langle \psi | D\psi_0, \psi | \rangle = 0. \quad (4.29)$$

We can point out, evidently, that the matrix $\langle \psi | D\psi_0, \psi | \rangle$ is not necessarily zero, and this may be essential. The procedure described previously in

high-energy physics considers problems like the following: go and come down; it's the sea-saw problem and you really don't. The carts with a couple of particles, like a proton and a pion or something like that, initially, in the lab, were p_1 and particle p_2 standing still, and the initial sum of momenta was zero. The p_1 particle is initially at some level. The p_1 goes up and comes down, say two seconds, or something, and has enough energy to scatter electrons with certain momenta. What's the amplitude to do this to p_2 ? The mathematics looks like this. You write ψ to excite the spins and momenta of the incoming particles. So it's ψ , and then the problem's sort of it becomes sort. For instance, with p_1 which includes ψ , you get the $\bar{\psi}$ vectors going in various directions, and then p_2 comes along, and it's off in these directions with the spin included. In other words, ψ would be specified by giving all the momenta and spins, and so on, of the final products. Then the job of the theorist is to calculate the amplitude (3.27). However, it's the ψ that's calculated in the special case that v_1 is $= c$, and v_2 is $= -c$. That is the experimental condition on the results of the theory, only on what comes in one spin goes off. The additional one of $\bar{\psi}\psi$, either $v_1 = +c$ or $v_2 = -c$, or both of S , must be dealt with.

$$(\psi, S) \psi.$$

(3.28) Using (3.27) and (3.28), we can calculate $\mathcal{A}_{\text{coll}}$ in \mathcal{K}_1 .

$$\mathcal{A}_{\text{coll}}(S).$$

which is called the δ -function. And you see a theoretical physicist passing by, and saying, "Well I have to do it's calculation." He doesn't know what he is talking about.

There's no interpretation to specify the laws for the $\mathcal{A}_{\text{coll}}$ to be introduced in question, so, basically, it's an assumption. In high-energy theory there's only one way to make this interpretation meaningful in fact by doing another way, which is very convenient. The other way can also be done in the relativistic case by introducing an interpretation. The theory and the framework for a small interval of time—in a better way for t_1 and t_2 taken together. One can have a sequence of such $\mathcal{A}_{\text{coll}}$'s for small intervals of time. We can watch how things go on. A relevant diagram shows an appropriate combination that's based on a lot of very old knowledge—because you don't want to have a negatively charged pion bring lower "mathematically" every element. So we won't worry about that—or be just going to worry about the relativistic mechanics.

Suppose we think of the entire $\mathcal{A}_{\text{coll}}$ as being from t_1 until t_2 which is greater than t_1 . In this case it's $\mathcal{A}_{\text{coll}}$ is the sum of successive integrals $\mathcal{A}_{\text{coll}}(t_1, t_2)$ plus $\mathcal{A}_{\text{coll}}(t_2, t_3)$ plus $\mathcal{A}_{\text{coll}}(t_3, t_4)$ and so on, that goes between t_1 and t_2 , is the problem in one dimension. Now, imagine when you take t_1 from t_1 until t_2 and then t_2 from t_2 until t_3 it's the "discretization" when we use our approachs of $\mathcal{A}_{\text{coll}}$ in series. We will then write, following the definition of Section 3.1,

$$\mathcal{A}_{\text{coll}}(t_1, t_2) = \mathcal{A}_{\text{coll}}(t_1, t_2) \cdot \mathcal{A}_{\text{coll}}(t_2, t_3). \quad (3.29)$$

So between t_1 and t_2 you any time interval if we can calculate a sequence of short time intervals between t_1 and t_2 . We just pick the t_1 and t_2 pieces. Or t_1, t_2, t_3, \dots that would be needed as an analysis, which is really,

Our problem, then, is to understand the relation $\mathcal{A}_{\text{coll}}(t_1, t_2)$ for an infinitesimal time interval, so, $t_2 - t_1 = \Delta t$. We can calculate that if we have a time Δt , what does the $\mathcal{A}_{\text{coll}}$ look like in physical interpretation? Let's see how to do it for when there is a Δt . So, let's do it in the example (3.28). We show the time dependence may not be perfectly clear that we mean the contribution the time Δt . Now, we ask the question: What's the condition after the small interval of the $\mathcal{A}_{\text{coll}}$? The answer is

$$|\psi(t_1) - \psi(t_2)| = |\psi(t_1) - \psi(t_1 + \Delta t)| \mathcal{O}(\Delta t). \quad (3.30)$$

This is to be understood as meant by "order" meaning that the imaginary part ψ is

for ϕ at the time $t = \Delta t$, is

$$\phi(\Delta t) = \phi(0) + \phi(\Delta t) - \phi(0)C_0[\psi(t)]. \quad (8.33)$$

Since we're not yet concerned about abstract things like quantum numbers and representations, if we make the decomposition of Eq. (8.31) for C_0 , we get

$$C_0[\psi(t + \Delta t)] = C_0[\psi(t)] - \Delta t C_1[\psi(t)]. \quad (8.34)$$

We can also rescale the $C_i(t)$'s by linear states and weights:

$$\phi(\Delta t) = \sum_i C_i(t) \phi_i + \Delta t C_1(t) \phi_1. \quad (8.35)$$

Since ϕ undergoes Eq. (8.35) in the following way, if $\psi(0), C_0[\psi(0)] = 0$, we know the amplitude to be in the basis state ϕ_0 at the time t , then we can think of this amplitude (and its derivatives) varying with time. Eq. (8.35) becomes a function of t . And we also have some information on how the amplitudes change over time. Each amplitude at $t + \Delta t$ is proportional to one of the other amplitudes at t multiplied by a set of coefficients. To express the V matrix V_{ij} by $\psi(t)$ we mean

$$C_{ij} = \langle \phi_i | V | \phi_j \rangle.$$

Then we can write Eq. (8.35) as

$$\phi(\Delta t) = \sum_i V_{ii}(t) + \Delta t V_{11}(t) \phi_1. \quad (8.36)$$

This then is how the dynamics of current need to be in getting to tent.

We don't know much about the V , yet except for one thing. We know that if the perturbation is strong enough— we should get back the original state. So, $V_{11} \rightarrow 1$ and $V_{ii} \rightarrow 0$, $i \neq 1$. In other words, $V_{11} \rightarrow 1$ for $\Delta t = 0$. Also, we can suppose that for small Δt , each of the coefficients V_{ii} should differ from δ_{ii} by some tiny perturbation, so let's do our work

$$V_{ii}(t + \Delta t) = \delta_{ii} + K_{ii}\Delta t. \quad (8.37)$$

However, it is usual to take the factor $(-\Delta t)^2$ out of the coefficients K_{ii} , for historical and other reasons, so let's do again

$$V_{ii}(t + \Delta t) = \delta_{ii} - \frac{1}{3} H_{ii} \Delta t^2 \Delta t. \quad (8.38)$$

It is of course, the same as Eq. (8.36) since, if you wish, one defines the ψ function $\psi_i(t)$. The terms H_{ii} are just the derivatives with respect to t of the coefficients $\psi_i(t)$, $i \neq 1$, evaluated at $t_0 = t_1 = t$.

Using this form for V in Eq. (8.35), we have

$$C_0(t + \Delta t) = \sum_i \left[\delta_{ii} - \frac{1}{3} H_{ii}(t) \Delta t \right] C_0(t). \quad (8.39)$$

Taking the sum over the Δt terms, we find $C_0(t)$, which we can pull out of the entire sum of the expansion. Then dividing by Δt we have what we recognize as the initial

$$\frac{C_0(t + \Delta t) - C_0(t)}{\Delta t} = -\frac{1}{3} \sum_i H_{ii}(t) C_0(t).$$

or

$$\frac{dC_0(t)}{dt} = -\frac{1}{3} \sum_i H_{ii}(t) C_0(t). \quad (8.40)$$

[We are using here the Heisenberg representation. In our notes (8.29), the ψ refers to the imaginary part $\sqrt{-1}\psi$, and not the Hermitian representation that we used. We represent you would find it more confusing.]

You remember that $C_i(t)$ is the amplitude of $| \psi \rangle$ to find the state ψ in one of the base states i at the time t . So Eq. (3.39) tells us how each of the coefficients $C_i(t)$ varies with time. And that is the same as saying that Eq. (3.39) tells us how the state ψ varies with time, since we are describing ψ in terms of the amplitudes $C_i(t)$. The variation of ψ in time is described in terms of the matrix H_0 , which has no entries, of course, the things we are doing to the system to cause it to change. (We know the H_0 , which controls the physics of the situation and can, in general, depend on t .) Hence, we have a complete description of the behavior in one of the systems. Equation (3.39) is then the quantum mechanical law for the dynamics of the world.

(We should say that we will always take a set of base states which are fixed and do not vary with time. There are people who use base states that does vary. However, that's like using a rotating coordinate system in mechanics, and we don't want to get involved in such complications.)

8-5 The Hamiltonian matrix

The idea, then, is to try to describe the quantum mechanical world we need to pick a set of base states i and to write the physical laws by giving the matrix of coefficients H_{ij} . Then we have everything we can answer any question about what will happen. So we have to learn what the rules are for finding the H 's to go with any physical situation—what corresponds to a magnetic field, or an electric field, and so on. And that's the hardest part. For instance, for the new strange particles, we have no idea what H_{ij} to use. In other words, no one knows the complete H_0 for the whole world. (Part of the difficulty is that one can hardly hope to describe the H , since no one even knows what the base states are!) We do have excellent approximations for nonrelativistic phenomena and for some other special cases. In particular, we have the terms that are needed for the motions of electrons in atoms—to describe chemistry. But we don't know the full true H for the whole universe.

The coefficients H_{ij} are called the *Hamiltonian matrix*, or, for short, just the *Hamiltonian*. (How Hamilton, who worked in the 1830's, got his name on a quantum mechanical matrix is a tale of history.) It would be much better called the *energy matrix*, for reasons that will become apparent as we work with it. So the problem is: *Know your Hamiltonian!*

The Hamiltonian has one property that can be deduced right away, namely, that

$$H_{ij}^* = H_{ji}. \quad (8.40)$$

This follows from the condition that the total probability that the system is in some state does not change. If you start with a particle—an object in the world—then you still put it in there goes on. The total probability of finding it somewhere is

$$\sum_i |C_i|^2,$$

which must not vary with time. If this is to be true for any starting condition as given by Eq. (8.40), it can also be true.

As our first example, we take a situation in which the physical circumstances are not varying with time; we mean the external physical conditions, so that H is independent of time. Nobody is turning magnets on and off. We also pick a system for which only one base state is required for the description: it is an approximation we could make for a hydrogen atom or something similar. Equation (8.40) then says

$$i \hbar \frac{dC_i}{dt} = H_{ii} C_i. \quad (8.41)$$

Only one equation—that's all! And if H_{ii} is constant, this differential equation is easily solved to give

$$C_i = (constant) e^{-iH_{ii}t/\hbar}. \quad (8.42)$$

This is the time dependence of a state with total energy $E = E_{1,1}$. You can imagine it's height or intensity. The term $\alpha_{1,1}$ is the ground state of the energy but more complex structure.

Now, if we understand a little more about what the equations mean, we look at a system which has two basis states. Then Eq. (8.4) looks like

$$\begin{aligned} \hat{\rho}_1 \frac{d\psi_1}{dt} &= H_{1,1}\psi_1 + H_{1,2}\psi_2 \\ \hat{\rho}_2 \frac{d\psi_2}{dt} &= H_{2,1}\psi_1 + H_{2,2}\psi_2 \end{aligned} \quad (8.4)$$

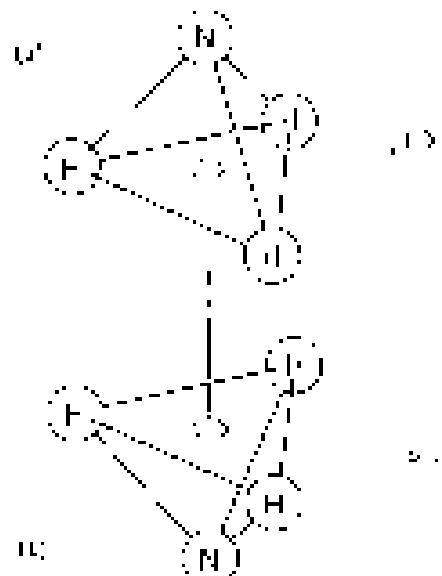
If the H 's are again independent of time we get easily solution of these equations. We have to solve the equations and we will come back to do it later. Now you can see the quantum mechanics without knowing the H 's, nothing else are independent of time.

8.6 The ammonia molecule

We've now learned how the dynamical equation of quantum mechanics can be used to describe a particular physical phenomenon. We have picked an interesting but simple example in which, by making assumptions made greater than the Heisenberg, we can work out some important and non-trivial results. We are going to take a situation described by us above, the ammonia molecule.

The ammonia molecule has one nitrogen atom and three hydrogen atoms located at a plane below the nitrogen so that the molecule has the form of a pyramidal shape in Fig. 8.1(a). Now this molecule, like most others, has certain numbers of degrees of freedom which we may ignore by neglecting many vibrations, it can be viewed as rigid, and so on, and so on. It is, therefore, not a free 3D system at all. But we want to make a very good approximation to the system. First, because it is diatomic, there are two atoms connected and added to the nitrogen. We will consider only that the molecule is spinning around its axis of symmetry (as is seen in the figure), that it has zero total angular momentum and that it is vibrating as it is possible. These specific conditions except those of zero total angular momentum for the nitrogen atom and nitrogen may be the first set of the state of the hydrogen atoms. On the other hand, as shown in Fig. 8.1(b) and (c), we will assume the molecule is enough to form a two-state system. We mean that there are only two states we are going to study, everything else, all other things being assumed to stay put. When we do this we know that it is spinning with a certain momentum and vibrating in a definite way, there are still two possible states. We will say that the molecule is in the state |1> when the nitrogen is "up" as in Fig. 8.1(a), and in the state |2> when the nitrogen is "down," as in (b). The state |2> will be taken as the set of basis states in our analysis of the behavior of the ammonia molecule. At any moment the actual state |S> of the molecule can be represented by giving $S_1 = |\psi_1|^2$, the amplitude to be in state |1>, and $S_2 = |\psi_2|^2$, the amplitude to be in state |2>. Using Eq. (8.3) we can write the transformation |S> as

$$\begin{aligned} S_1 &= \langle S|1|\psi_1\rangle + \langle S|2|\psi_2\rangle \\ 0 &= \langle S|H_1 - \langle S|H_2 \end{aligned} \quad (8.49)$$



(a) An equivalent picture to a treatment of the ammonia molecule.

Now the interesting thing is that the molecule is forced to be in one state or another state. It can't be in one state at one time while other. The two coefficients will be changing with time according to the quantum mechanics which is not for any classical system. So, just for example, if you try to make some observation, let's say he made some selection of the molecules so that you know that the molecule is definitely in the state |1>. At some later time there is some chance that it will be found in state |2>. To find out whether the molecule has to leave the state, you might have to do some diffraction experiment.

The only trouble is that we don't know what to use for the coefficients A_1 , A_2 , in Eq. (8.43). That is because they're 95% zero, however. Suppose that were the hydrogen were in the state $|1\rangle$; then there was no chance that it could ever get into $|2\rangle$, and vice versa. Then A_{12} , and A_{21} , would both be zero, and Eq. (8.43) would read

$$i\hbar \frac{dC_1}{dt} = E_1 C_1 - i\hbar \frac{dC_2}{dt} = E_2 C_2.$$

We can easily solve these two equations, see (8)

$$C_1 = (\text{constant}) e^{-iE_1 t/\hbar}, \quad C_2 = (\text{constant}) e^{iE_2 t/\hbar}. \quad (8.44)$$

This is just the same situation as in many states with the charges $N_1 = N_2$, and $H_1 = H_2$. We note, however, that in a beam where molecule the two states $|1\rangle$ and $|2\rangle$ have a definite symmetry, if nature is at all reasonable, the atoms cluster in H_1 , and H_2 must be equal. We'll call them H , since they're both equal to the energy. In other words, $E_1 = E_2$, and A_{12} , A_{21} , were zero. But then, (8.43) do not tell us very much; in fact really does. It means that it's possible for the nitrogen to push its way through the three hydrogens and flip it over now. This isn't difficult; it just requires a lot of energy. However, it's not the easiest thing to do with energy alone. There's some anti-destruction effect here into the energy barrier. It is possible in many more cases, a much greater increase in energy which is called "resonance". There is, therefore, some sort of intermediate resonance which when $\epsilon = 0$ will put us the state $|2\rangle$. The coefficients H_1 and H_2 are not really zero. Again by symmetry, they should be half the sum of the two. In fact, we already know that, in general, H_1 going to occur, or be caused by conjugation of H_2 , we then can differ only by a phase. It turns out, as you will see, that there is no loss of generality, if we take the two equal to each other. For later convenience we set these two to be negative number, so take $H_1 = H_2 = -\epsilon$. We then have the following two differential equations:

$$i\hbar \frac{dC_1}{dt} = E_1 C_1 - A C_2, \quad (8.45)$$

$$i\hbar \frac{dC_2}{dt} = E_2 C_2 - A C_1. \quad (8.46)$$

These equations are coupled strongly and can't be solved in any numerical terms. One convenient way is the following. Take in the second the ratio, we get

$$i\hbar \frac{d}{dt} (C_2 - C_1) = (S - A)(C_1 + C_2),$$

whose solution is

$$C_1 + C_2 = A e^{-i(S-A)t/\hbar}. \quad (8.47)$$

Then, taking the difference of (8.45) and (8.46), we find that

$$i\hbar \frac{d}{dt} (C_1 - C_2) = (L - A)(C_1 - C_2),$$

whose proxy

$$C_1 - C_2 = A e^{-i(L-A)t/\hbar}. \quad (8.48)$$

We have added the two integration constants C_1 and C_2 ; these are, of course, to be chosen, as given, the appropriate starting condition for any particular physical problem. Now, by adding and subtracting (8.47) and (8.48), we get C_1 and C_2

$$C_1(t) = \frac{A}{\sqrt{2}} e^{-i(S+A)t/\hbar} + \frac{B}{\sqrt{2}} e^{-i(L+A)t/\hbar}, \quad (8.49)$$

$$C_2(t) = \frac{A}{\sqrt{2}} e^{-i(S-A)t/\hbar} - \frac{B}{\sqrt{2}} e^{-i(L-A)t/\hbar}. \quad (8.50)$$

They are now in a position for the sign of C_1 & C_2 choose them

We have the condition $\epsilon_1 = \epsilon_2$. They must! (The trouble with quantum mechanics is not only in getting the equations, but in understanding where the solutions must be.) First, notice that if $\epsilon_1 = 0$, both terms give zero from $\epsilon_1 - \epsilon_2$ and $\epsilon_1 + \epsilon_2$. To say this puts us at the beginning, it means that the system is at a state of definite energy—here the energy $E_0 = \epsilon_1$. But there is a different state of this energy, at which the two amplitudes C_1 and C_2 are equal. We get the result that the molecular wavefunction ψ_0 is E_0 if there are no appreciable differences for the nitrogen atom to be “up” and to be “down.”

There is another stationary state possibility, one that is important for biological chemistry ($E_1 = \epsilon_1 \hbar\omega$). So there is another state with the definite energy $E_1 = \epsilon_1 + \hbar\omega$ of the two atoms being coupled together in phase, $C_1 = C_2$. These are the only two states of definite energy. You will discuss the case of the ammonia molecule in more detail in the next chapter; we will mention here only a couple of things.

We conclude that “coherences” is some change—but the changes can come from one molecule to the other. The energy of the molecule is not just ϵ_1 ; it is now coupled, but the total energy remains fixed ($E_0 = \epsilon_1$) and $E_1 = \epsilon_1 + \hbar\omega$. Every one of the possible states in the molecule is called an “excited” state, even though it is not excited. We are very fond of these states because you remember we paid a lot of attention to excited and internal energy, and so on. It is that possible condition of that kind that is a feature of every basis function of the Hilbert space of the molecule.

Let's now set the following question about ammonia added to SF_6 . Suppose that $\epsilon_1 = 0$, we know that a molecule is in the walls. Then, we also know that $C_1(0) = 1$ and $C_2(0) = 0$. What is the probability that the molecule is before in the state E_1 ? And let's see how well still be found in state E_0 at the time t ? Our starting condition tells us what ψ_0 and ψ_1 are in Eqs. (13.9) and (13.10). Let's say $\epsilon_2 = 0$, we have that

$$C_1(t) = \frac{e^{i\theta} + e^{-i\theta}}{2} = 1, \quad C_2(t) = \frac{e^{i\theta} - e^{-i\theta}}{2} = 0.$$

Clearly, $\epsilon_1 = 0 = 1$. Putting these values into the formulae for $C_1(t)$ and $C_2(t)$ and rearranging terms, we have

$$C_1(t) = e^{-i\theta/2} \sqrt{\frac{e^{i\theta/2} + e^{-i\theta/2}}{2}},$$

$$C_2(t) = e^{i\theta/2} \sqrt{\frac{e^{i\theta/2} - e^{-i\theta/2}}{2}}.$$

We can rewrite, because

$$C_1(t) = e^{-i\theta/2} \cos \frac{\theta}{2}, \quad (13.12)$$

$$C_2(t) = e^{-i\theta/2} \sin \frac{\theta}{2}. \quad (13.13)$$

For two amplitudes having a magnitude the same harmonically related.

The probability that the molecule is found in state E_1 is the absolute square of $C_2(t)$:

$$|C_2(t)|^2 = \sin^2 \frac{\theta}{2}. \quad (13.14)$$

The probability starts at zero (is it that?) since there is no oscillation between two states, as shown in the first molecule in Fig. 13.2. The probability of being in the E_1 state decreases of course staying at one, decreasing to the second state until the probability of finding the molecule in the first state is zero as shown by the curve in Fig. 13.2. Try yourself by drawing back and forth between the axes.

Along comes up we can write components when we have two small probabilities with a slight coupling. (See Chapter 19, Vol. I.) When we do this and let us

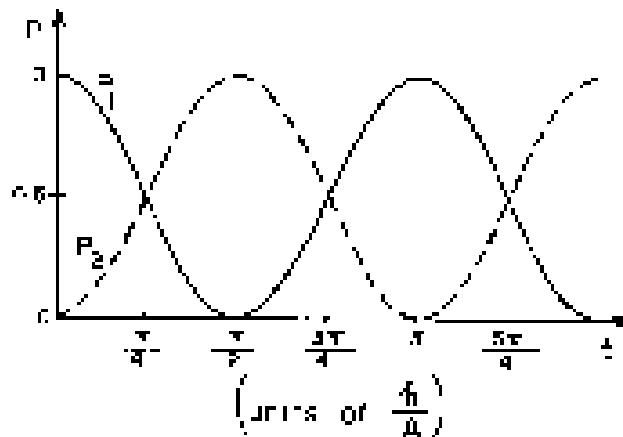


Fig. 9-2. The probability P_1 for the electron to be in state 11 at $t = 0$ is the same as in Fig. 9-1. The velocity P_2 has now been added to term 2.

it swings back and forth, the other one starts to swing. Both since the overall pendulum has passed its equilibrium. Then the process repeats & of pendulum number 2 is proportional to energy. It is exactly becoming kind of that is. The total total energy is constant back and forth. Quantum system is oscillating between the two pendulums, the size at which the "oscillation" is also is kept same. Also, you remember, will be we open them there are two specific directions, such as, clockwise & counter-clockwise which we call the fundamental modes. If we pull both pendulums together, they swing together at one frequency. On the other hand, if we pull one pendulum to one side and the other on the other way, there is another set of modes which oscillate with frequency.

With this we have a similar situation like in molecules or nuclei is numerically like the pair of pendulums. There are two frequencies ω_1 & ω_2 & ω_{12} . In other they are oscillating together, or oscillating independently.

The general analogy is not much deep than the principle of superposition, because here the same velocities. The linear equations for the amplitudes (8.39) are very much like the linear equations of harmonic oscillators. In fact this is the reason behind the success of our classical theory of the index of refraction. In which we neglect the quantum oscillations around a Larmure oscillator, even though classically this is not a necessarily valid as electrons circulating about a nucleus. If you pull the nitrogen to one side, then you get a very small value of the refractive index, and you need a lot of luminous fluxes to make up for it in order to see the color associated with the frequency. The splitting of the energy levels of the atomic molecule is however a nearly $\sim 10^{-20}$ in mechanical effect.

The splitting of the energy levels of the atomic molecule has important physical applications which we will describe in the next chapter. At long last we have the example of a practical physics problem that you can understand with the quantum mechanics.

The Aromatic Maser

9-1 The state of an ammonia molecule

In this chapter we are going to discuss one particular type of quantum-mechanical problem, the treatment of which may wonder why we drop our normal quantum mechanics method to do a special problem, but you will find that it gives the features of the exact solution quite well and in the general form of quantum mechanics one can still learn by working considerably harder than the present treatment. The ammonia maser is a device for generating electromagnetic waves whose operation is based on the properties of the ammonia molecule which we discussed briefly in the last chapter. We hope to understand what we found there.

The ammonia molecule has many states, but we need consider only two states in thinking how one goes from when it happens when the molecule is in one specific state of rotation or translation. A physical state for the two states can be visualized as follows. If the ammonia molecule is considered to be rotating about an axis passing through the nitrogen atom perpendicular to a horizontal hydrogen atom, as shown in Fig. 9-1, there are still two possible configurations.

The most probable configuration of the plane of rotation about an axis on the way, \hat{N}_z , is in a two-state system, $|1\rangle$ and $|2\rangle$. This situation is best understood by analyzing the behavior of the ammonia molecule.

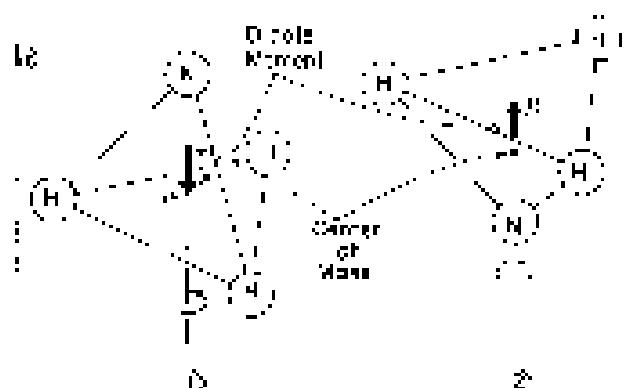


Fig. 9-1. A sketch of an ammonia molecule showing the dipole moment and the angle of inclination.

In a system with two basis states, only one of the systems can always be described as a linear combination of the two basis states, i.e., the one to be in complete phase, $|1\rangle$, to be in one case state and the amplitude of $|2\rangle$ to be in the other. We can write its state vector as

$$\begin{aligned} |\psi\rangle &= |1\rangle C_1 + |2\rangle C_2 \quad (9-1) \\ \text{where} \quad C_1 &= \langle 1 | \psi \rangle \quad \text{and} \quad C_2 = \langle 2 | \psi \rangle \end{aligned}$$

To get two amplitudes, C_1 and C_2 , time dependent in the ammonia molecule, Eq. 9-1 is not enough. Making use of the symmetry of the ammonia molecule, we set $R_{11} = R_{22} = S_1$ and $R_{12} = R_{21} = -S_2$ and set C_2

9-2 The state of an ammonia molecule

9-3 The molecule in a static electric field

9-4 Transitions in a time-dependent field

9-5 Interference of resonance

9-6 Transitions off resonance

9-7 The absorption of light

MASER = Microscopic Amplification by Stimulated Emission of Radiation

and from Eqs. (8) we have (8.7)

$$C_1 = \frac{1}{2} e^{-i\theta_1 E_1 t} + \frac{1}{2} e^{+i\theta_1 E_1 t}, \quad (8.2)$$

$$C_2 = \frac{1}{2} e^{-i\theta_2 E_2 t} + \frac{1}{2} e^{+i\theta_2 E_2 t}. \quad (8.3)$$

It would be a "natural" choice to choose these solutions. Suppose the molecule was initially prepared again at time zero so that the two initial states were zero. Then at $t = 0$, the amplitudes to be in the states $|1\rangle$ and $|2\rangle$ are identical, and after over time t are different. Their phases both vary with time in the same way—namely in frequency $\Omega_C = \theta_1/\hbar$. Similarly, if we were to put the molecule into state $|3\rangle$ at time $t = 0$, the amplitude $C_{3,0}$ is the negative of C_1 , and this also would vary with time in frequency $\Omega_C = \theta_1/\hbar$. Because the angle θ_1 is positive it is clear that the relation between C_1 and C_2 is independent of time.

We have found two special solutions in which the two angular velocities are in agreement, furthermore, have phases which vary in the same frequencies. These are minimum values of ψ as defined in Section 7.1, where instead here θ_1 is a constant phase angle. This state $|3\rangle$ has energy $E_{3,0} = \hbar\omega_1$, and the state $|1\rangle$ has the energy $E_{1,0} = E_1 - \hbar\omega_1$. They are the only two stationary states that exist. We find that the molecule has two energy levels, with the energy difference ΔE . We mean, of course, no energy levels for the excited state unless there are other frequencies introduced in invariable amounts and so forth.

If we didn't allow for the possibility of the nitrogen dipole being permanent, we think we could have found two more two-levels which have a gap ΔE each other in energy. The reason lies in the fact, for example, that a single cavity is ΔE because they are anti-symmetric $-1/2$, giving a separation of ΔE between the energies of the two states. Since ΔE is in fact very small, the difference in energy is also very small.

In order to evoke an electron bubble on glass, the energies involved are necessarily very high, i.e., the photons in the optical field will be intense. The result is resonance of the molecule in the photon field of the field. If you take their existing energies, the energy differences of the states correspond to cyclotron frequencies. So, the energy E_1 is lower than that of those had to be. Below the infrared and will in the microwave region. Even in the radio, i.e., the millimeter, but there is a pair of energy levels with a separation of $\sim 10^{-3}$ electron volt, corresponding to a frequency 24,000 megacycles. Obviously this means that $\Delta E = \hbar\omega_1 + \hbar\omega_2 = 24,000$ megacycles corresponds to a wavelength of $\lambda_1 = \lambda_2$, and we know another λ but now it is the one which does not consist of the electric field, but only of the magnetic field.

For the work left Moses can not describe these two states of definite energy and the resulting supposed resonance is more easily simplified by taking the sum of the two numbers C_1 and C_2 ,

$$C_{1,0} = C_1 + C_2 = (\frac{1}{2} e^{-i\theta_1 E_1 t} + \frac{1}{2} e^{+i\theta_1 E_1 t}) \quad (9.4)$$

Here we call the "coherent" wave, and is just represented in form. Indeed, $|1\rangle$ is a new state. At time zero the amplitudes of the original two states are equal—this is because $C_{1,0} = \sqrt{2}|1\rangle$, we can substitute (9.4) into Eq. (8.1) and see it is true for any time t for

$$|W\rangle = |1\rangle + |2\rangle,$$

which means the same as

$$|W\rangle = |1\rangle + |2\rangle. \quad (9.5)$$

* In the following sections—especially in Chapter 10 you will notice some minor variations in the notation of the present, because the Americanized forms the former, and the British conventional names of the terms "wave" and "field" for the American "wave" and "field" and "current" for the British "radiation" and "vector" "magnetic force" and "electric current" respectively.

The amplitude for the state $|B\rangle$ to be in the state $|J\rangle$ is

$$\langle J | B \rangle = \langle J | \psi + \phi | \rangle,$$

which is determined by ϕ , since $|J\rangle$ and $|\psi\rangle$ are linear states. The amplitude for the state $|B\rangle$ to be in the basis state $|J\rangle$ is also determined by ϕ , because both are equal amplitudes between the two base states $|J\rangle$ and $|B\rangle$.

We can, however, in a bit of a fiddle. The state $|B\rangle$ has a real probability greater than one of being in some base state other than $|J\rangle$. In other words, however, there are states ψ which "maximize" $\langle B | \psi \rangle$. We can take care of that by normalizing ψ : we end up here $\langle B | \psi' \rangle = 1$, which holds true for any ψ' . Using the general relation that

$$\langle B | \psi' \rangle = \sum_j \langle B | B_j \rangle \psi'_j,$$

in the basis $|J\rangle$ and $|B\rangle$ the state B , and taking the sum over the base states $|J\rangle$ and $|B\rangle$ we get to

$$\langle B | B' = \langle B | J \rangle \psi'_J + \langle B | B \rangle \psi'_B.$$

This will be equal to one as it should from our definition of $C_{B|J}$, i.e., Eq. (9.5) is now

$$C_{B|J} = \frac{1}{\sqrt{2}}(\psi'_J - \psi'_B)$$

In the same way we can work out an impurity

$$C_J = \frac{1}{\sqrt{2}}(\psi'_J + \psi'_B)$$

or

$$C_B = \frac{1}{\sqrt{2}}(\psi'_J + \psi'_B). \quad (9.6)$$

The amplitude ψ'_J is proportional to the state $|J\rangle$ (because $|J\rangle$ is the base state) and ψ'_B is proportional to the state $|B\rangle$ (because $|B\rangle$ is the base state). Namely, $\langle B | J \rangle = \langle B | B \rangle = 1$ and $\langle B | B \rangle = 0$.

$$\psi'_J = \frac{1}{\sqrt{2}}(B - B)|J\rangle.$$

or

$$\psi'_B = \frac{1}{\sqrt{2}}(B + B)|B\rangle. \quad (9.7)$$

From which it follows that

$$\langle B | J \rangle = \frac{1}{\sqrt{2}}(-B)|J\rangle.$$

Now the reader is asked once again to note that the states $|J\rangle$ and $-|B\rangle$ are taken as independent base states which are especially convenient for describing the electronic degrees of the magnetic molecule. You can consider the $C_{B|J}$ expansion of B and it has shown that

$$\langle B | J \rangle = b_{B|J}.$$

We have already done things so that

$$\langle B | J \rangle = \langle B | B \rangle$$

You can easily show from Eqs. (9.5) and (9.7) that

$$\langle B | B \rangle = \langle B | J \rangle = b_{B|J}.$$

The amplitudes $b_{B|J} = \langle B | J \rangle$ and $b_{B|B} = \langle B | B \rangle$ for any state B to be in the two base states $|J\rangle$ and $|B\rangle$ are also easily $\langle B | \psi \rangle$ amplitudes with the

(recall Eq. (3.2)). In fact, it is not difficult to see that the two equations (3.2) and (3.3) are equivalent with respect to ψ_1 , we see that

$$\alpha \frac{dC_1}{dt} = (E_0 + \beta C_1 - E_1 C_2). \quad (3.8)$$

And, taking the sum of Eqs. (3.2) and (3.3), we see that

$$\alpha \frac{dC_0}{dt} = C_{00} - 4f_{00} = 4\mu C_0 - 4\mu C_0. \quad (3.9)$$

Using (3) and (7) for f_{00} we see that the Hamiltonian matrix has the simple form

$$\begin{aligned} H_{11} &= E_1 - H_{22} \rightarrow 0, \\ H_{12} &= C_1 - H_{21} = E_0. \end{aligned}$$

Note that, since all the Eqs. (3.8) and (3.9) look just like what we had at Section 2.4 for the equation of a current system, they have a simple structure and one electron corresponding to a single energy. As time goes on, the amplitudes C_1 and C_2 each increase and approach E_0 .

The two stationary states $|A\rangle$ and $|B\rangle$ we found above are of course solutions of Eqs. (3.2) and (3.3). The state $|\psi\rangle$ that which $C_1 = C_2$ has

$$C_1 = e^{-iH_0 t / \hbar} C_0, \quad C_{00} = 0. \quad (3.10)$$

And the state $|\psi_{AB}\rangle$ (for which $C_1 = C_2$) has

$$C_1 = 0, \quad C_{00} = e^{-iH_0 t / \hbar} C_0. \quad (3.11)$$

Remembering the amplitudes (Eq. (9), (10)),

$$C_1 = 0, \quad C_0 = 2M, \quad C_{00} = 2M/\sqrt{\mu},$$

in Eq. (3.10) means the same thing as

$$|\psi\rangle = |A\rangle e^{-iH_0 t / \hbar} |A\rangle.$$

That is, the state $|\psi\rangle$ of the stationary system $|\psi\rangle$ is the same as the stationary state $|A\rangle$ in every other exponentially approaching to the energy of E_0 . In fact at $t = 0$

$$|A\rangle = |B\rangle$$

The state $|\psi\rangle$ has the same physical interpretation as the stationary state of energy E_0 [p. 4]. In the same way, we have for the second stationary state that

$$|\psi_B\rangle = |B\rangle e^{-iH_0 t / \hbar} |B\rangle.$$

The state $|\psi\rangle$ is just the stationary state of energy $E_0 = 4$ at $t = 0$. This our two new wave states $|A\rangle$ and $|B\rangle$ are physically the form of two states of different energy, with the same initial state taken out so that they can be time independent from the α -spin. Note also that $|A\rangle$ and $|B\rangle$ are independent, and we have to distinguish always between the stationary states $|\psi\rangle$ and $|\psi_B\rangle$ and their base states $|A\rangle$ and $|B\rangle$, since that differs only by the previous time factor.

In summary, the state vectors $|A\rangle$ and $|B\rangle$ are a pair of time states which are appropriate for describing the certain energy states in the symmetric problem. They are related to our original base vectors (2)

$$|A\rangle = \frac{1}{\sqrt{2}} (|1\rangle + |2\rangle), \quad |B\rangle = \frac{1}{\sqrt{2}} (|1\rangle - |2\rangle). \quad (3.12)$$

The amplitudes C_0 and C_1 are related to C_A and C_B by

$$C_0 = \frac{1}{\sqrt{2}} (C_A - C_B), \quad C_1 = \frac{1}{\sqrt{2}} (C_A + C_B). \quad (3.13)$$

A state at all can be represented by a linear combination of states formed from basis states $|1\rangle$ and $|2\rangle$, combining two of the basis states give basis states $|3\rangle$ and $|4\rangle$ with coefficients C_1 and C_2 . Thus,

$$|1\rangle = C_1|1\rangle + C_2|2\rangle$$

$$|3\rangle = C_1|1\rangle - iC_2|2\rangle.$$

The second term gives us the one-photon scattering rate R_{13} at a state with the energy $E_1 - E_3 = \omega$ from a state with the energy $E_1 - E_2 = \omega$.

9-2 The molecule in a static electric field

In the harmonic oscillator approximation the two atoms define a center of mass displacement Δr such that $\Delta x = \Delta y = \Delta z = \Delta r/2$. The system may still be considered free from the effect of a field. Or if it is in the center of a field, some change to a lower potential energy position. But in order to keep its state, one must have a classical connection to the states, some way of coupling the system. This must be some external machine or in affecting the states such as magnetic or electric fields. In this particular case, there is no one existing static electric field. We will therefore look at the problem of the behavior of the harmonic molecule in an external electric field.¹

We focus the discussion in one atomic field, we go back to the original base system (1) and (2), rather than using (1) and (2). Suppose that there is an electric field in the x -direction perpendicular to the plane of the hydrogen molecule. Then giving for the moment, by possibility of applying such a field, would it be true that the energy of the molecule is proportional to the proportion of the hydrogen atom? Generally, no. The electron tends to be closer to the nucleus than to the hydrogen nuclei, so the hydrogens are slightly positive. The usual account depends on the details of charge distribution, it is a complicated problem to know exactly what the distribution is, but in any case the net result is that the molecular molecule has an electric dipole moment, as indicated in Fig. 9-1. We can examine our analysis without knowing in detail the details of a moment of dipole moment because p is to be considered just the moment of a vector, let's suppose that the electric dipole moment is p , and its direction point from nitrogen atom and point away from the plane of the hydrogen atoms.

Now, since the nitrogen slips from one side to the other, this occurs of course with motion, but the electric dipole moment will be zero. As a result of this motion, the energy in a vacuum field with oxygen can be made discontinuous. With the assumption made above, the process of energy will be higher at the nitrogen component of the direction of the field, and lower at p in the opposite direction; the separation in the two energies will be 2μ .

In the discussion up to this point, we have assumed values of N_1 and N_2 without knowing how to calculate them. According to the correct physical theory, it should be possible to calculate these numbers in terms of the positions and motions of all the protons and electrons. Let's say that's done. Such a system is called a quantum mechanical model and that's just too cumbersome to get into. As a matter of fact, there is no need to do this, much more clean. One probably than we do. All we do, we say is that when there is no such a field, the energy of the two states is the same, the difference being proportional to the size of the field. We have talk of a constant of proportionality k , but its value may be determined experimentally. We can say that the molecule has the amplitude to dip over, but this will have to be measured experimentally. Nobody can give us accurate numerical values of k and A because the calculations are too complicated to do in detail.

¹ We presume that we have no previous knowledge. Since we are learning quantum mechanics and energy, we don't know the quantum and the classical and electric field. Remember, in the center of the atom we don't speak motion.

For the magnetic moments in our electric field, one could said could be obtained. If we ignored the amplitudes for the moment to flip from one orientation to the other, we would expect the energies of the two states, E_1 and E_2 , to be $(E_1 - E_2)$. Following the procedure of the last chapter, we have

$$H_{11} = E_1 + \omega_1, \quad H_{12} = E_2 - \omega_2. \quad (9.14)$$

Also we will assume that for the electric fields of \mathbb{C} exist the fact does not affect irreversibly the geometry of the molecule and, therefore, does not affect its amplitude. But the charges will complicate the position to the terms. We can now write down H_{21} and H_{22} , but not simplified.

$$H_{21} = H_{12} = -\omega_1. \quad (9.15)$$

We must now solve the Hamiltonian equations, Eq. (8.2), with these new values of H_{ij} . We could solve them, as we did before, but since we are going to have seven degrees of freedom the solutions for two state systems, i.e., other than equations obtained for all the angular coordinates, may H_{ij} , assuming only that they do not change with time.

We want to extend solution of the problem Hamiltonian \mathcal{H} , since

$$\partial_t \frac{\partial^2 C_1}{\partial t^2} = H_{11} C_1 + V_1 \delta_{11}, \quad (9.16)$$

$$\partial_t \frac{\partial^2 C_2}{\partial t^2} = H_{12} C_2 + V_2 \delta_{22}. \quad (9.17)$$

Since there are two additional equations and two state coefficients, we can choose 14 unknowns which we assume the functions of the dependent variables t . We will take the basis functions in which C_1 and C_2 both have the same time dependence; we can now be trial functions:

$$C_1 = \alpha_1 e^{-iE_1 t}, \quad C_2 = \alpha_2 e^{-iE_2 t}.$$

Since each solution corresponds to a total energy $E = \hbar\omega$, we may as well write right away

$$C_1 = \alpha_1 e^{-iE t}, \quad (9.18)$$

$$C_2 = \alpha_2 e^{-iE t}, \quad (9.19)$$

where E is real or pure, and to be consistent with the 'exact' equations (9.16) and (9.17) in zeroth

order, we substitute C_1 and C_2 from (9.18) and (9.19) in the differential equations (9.16) and (9.17); the derivatives give us just $-i\omega_1 \alpha_1 e^{-iE t}$ and $-i\omega_2 \alpha_2 e^{-iE t}$ of the left side, assuming $E_1 < E_2$. Cancelling the common exponential factor, we get

$$(\mathcal{P} - H_{11})\alpha_1 - H_{12}\alpha_2 = \hbar\omega_1 = \partial_1 \alpha_1 - \partial_2 \alpha_2.$$

On, rearranging the terms, we have

$$(\mathcal{P} - H_{11})\alpha_1 - H_{12}\alpha_2 = 0, \quad (9.20)$$

$$\partial_1 \alpha_1 - (\mathcal{P} - H_{11})\alpha_1 = 0. \quad (9.21)$$

With some care of non-easy to solve equations, these will be reduced to the form α_1 and α_2 only if the determinant of matrix, either α_1 and α_2 , is zero. For (9.20)

$$\det \begin{pmatrix} \mathcal{P} - H_{11} & H_{12} \\ -H_{12} & \mathcal{P} - H_{11} \end{pmatrix} = 0. \quad (9.22)$$

However, when there are only two equations and two unknowns, we cannot expect a unique solution. The two equations (7.10) and (7.11) each give a value for the ratio α_1/α_2 , and as α_1 and α_2 are single two ratios must be equal. From (7.10) we have

$$\frac{\alpha_1}{\alpha_2} = \frac{H_{11}}{H_{11} - H_{21}}, \quad (7.23)$$

and from (7.11) we have

$$\frac{\alpha_1}{\alpha_2} = \frac{E - H_{22}}{H_{22}}. \quad (7.24)$$

By solving these two ratios, we get that E must satisfy

$$(E - H_{11})H_{22} - H_{21}E + H_{12}H_{21} = 0.$$

This is the same result we would get by solving Eqs. (7.10)-(7.11). In other words, we have found the energy E for which two methods give the same result.

$$E = H_{11} + H_{22} = \sqrt{H_{11}^2 - H_{12}H_{21}}. \quad (7.25)$$

From the two possible values for the energy E , we note that both of them give the same ratios for the ratios α_1/α_2 , and H_{22}/H_{12} , and $H_{12}H_{21}$ is equal to $(E - E_1)(E - E_2)$, which is because of the condition

Using the value E_1 above, and we look before we will use the upper energy E_2 for the lower energy E_1 . We have

$$E_1 = \frac{H_{11} + H_{22} - \sqrt{(H_{11} - H_{22})^2 - H_{12}H_{21}}}{2}, \quad (7.26)$$

$$E_2 = \frac{H_{11} + H_{22} + \sqrt{(H_{11} - H_{22})^2 - H_{12}H_{21}}}{2}. \quad (7.27)$$

Using either of these two energies symmetrically in Eqs. (7.14) and (7.15), we have the same solution for the two symmetric states (the excited state) with energy E . If there are no external disturbances, a system initially in one of these states will stay in that state during such perturbations.

We can check our results for the special case. If $H_{12} = H_{21} = 0$ we have that $E_1 = H_{11}$ and $E_2 = H_{22}$. This is certainly correct, because then Eqs. (7.14) and (7.15) are incomplete, and each represents a state X^k where $H_{11} = H_{12} = 0$. Next, if we set $H_{12} = H_{21} = E_1$ and $H_{22} = H_{11} = -E_1$ we get the following solutions:

$$\psi_1 = J_{12} = 0 \quad \text{and} \quad \psi_{11} = E_1 = 4.$$

For the given value the two solution E_1 and E_2 , the solution is such that we can again call the states

$$|\psi_1\rangle = |J_{12}\rangle^{(0.5)^{1/2}} \quad \text{and} \quad |\psi_{11}\rangle = |E_1\rangle^{(0.5)^{1/2}}.$$

The constants $a_1 = 2\pi\omega_1 C_1$ and C_1 is given in Eqs. (7.8) and (7.9), where ω_1 and a_1 are still to be determined, while $a_2 = 0$ is given by Eqs. (7.10)-(7.11) or Eq. (7.24). These two above key are now complete. If the system is forced to remain in one of the stationary states, the sum of the probabilities that it will be found in either of the final equal state. We must have that

$$C_1^{-2} + C_2^{-2} = 1, \quad (7.28)$$

or, equivalently,

$$|a_1|^2 + |a_2|^2 = 1. \quad (7.29)$$

These conditions do not uniquely specify a_1 and a_2 , they are still undetermined.

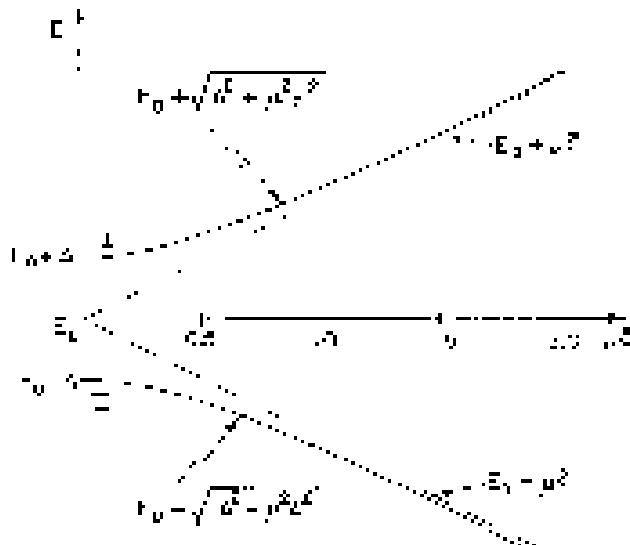


Fig. 9-2 Energy levels of the ammonia molecule in an electric field.

by an arbitrary phase ϕ , which would be a state like $|1\rangle$. Although present in both $|E_0\rangle$ and $|E_0 + \mu^2\rangle$, the $\langle\psi|\psi\rangle$ probability density is usually not the same due to the word "usually" for each case.

Let's go back to our particular example of the ammonia molecule in an electric field. Using the values of E_0 , H_0 , and H_{12} , we get (9.14) and (9.15), we get the eigenvalues of the two stability states:

$$E_1 = E_0 + \sqrt{\Delta^2 + \mu^2}, \quad E_2 = E_0 - \sqrt{\Delta^2 + \mu^2}. \quad (9.20)$$

The two energies are plotted as a function of the electric field strength Δ in Fig. 9-2. When $\Delta=0$, which is half the sum of the two energies, we obviously get $E_1=E_2$. When an electric field is applied, the splitting increases as the field increases. The E_1, E_2 pair goes to half the way with it, because the two molecules interact, and they are iso-parallel but anti-multiply along \hat{z} , so the energy is split just

$$E_1 = E_0 + \mu^2 = H_{11}, \quad E_2 = E_0 - \mu^2 = H_{22}. \quad (9.21)$$

Therefore this is an argument for the splitting in Fig. 9-1, and from the point of view of the two electrons alone, for different excited states. This is interesting, since that was something very unlikely.

We are at an early stage of understanding the operation of the laser in more detail. The field is the fulcrum. E is a tool of separating molecules in the state H_{11} from those in H_{22} . Then the idea is, if the higher energy state H_{11} passes through a cavity which has a resonance frequency of 243000 cps, the electrons can deliver energy to the cavity—this is what will happen here—and again we end up in state H_{22} . Each excitation makes one transition, so we deliver the energy $E_1 - E_2 = \hbar\nu$ to the cavity, and every thousand molecules will appear as electric energy in the cavity.

Now can we separate the two electronic states? This is not too hard. The ammonia can be put out of the cloud and passed through a glass cloud to give a mixture, and as shown in Fig. 9-3, the mixture is separated through a

→ For example, in following section we expect to encounter the conditionality with:

$$\Delta = \left[(E_1 - \mu^2)^2 - (E_2 - \mu^2)^2 \right]^{1/2} = \frac{E_1 - E_2}{H_{11} - H_{22} - 2\mu^2}.$$

From now on we will write H and H' instead of H_{11} and H_{22} . You'll remember that the sum of the E_1 and E_2 are the energy components of the unperturbed electronic states.

region in which there is a large transverse electric field. It is easiest to imagine this field directed so that the electric field E is parallel to the velocity vector of the beam. Now, a molecule moving $|v|$ has an energy which decreases with v^2 , and therefore this part of the beam will be scattered toward the region of low v^2 . It is also in this region E_y will, on the other hand, be reflected toward the region of high v^2 since its energy decreases as v^2 increases.

Incidentally, with the electric fields available at present and at the present time, the energy loss is a maximum of about 4%. In other cases, the stopping power Eq. (6.20) can be approximated by

$$A \left(1 + \frac{v^2}{2 A^2} \right). \quad (6.32)$$

So the energy loss is, for all practical purposes,

$$E_y = E_0 + \beta = \frac{E_0^2}{2 A^2}, \quad (6.33)$$

and

$$E_{yy} = E_0 - \beta = \frac{E_0^2}{2 A^2}. \quad (6.34)$$

Now the energies vary approximately linearly with v^2 . The forces on the molecules are then

$$F = \frac{e}{2A} \nabla E^2. \quad (6.35)$$

Molecular ions, on average, experience a field which is proportional to θ^2 . The coefficient is proportional to the molecule's moments of inertia, and is very high polyatomicity because of the small values of θ in the dense region. Thus, diatomic molecules are unusually sensitive to a static electric field. When you expect, for the dielectric efficiency of 10^4 by θ^2 .

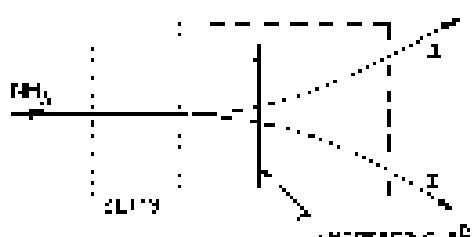


Fig. 6-1. The stopping power may be measured by an array of solid-state detectors having a gradient perpendicular to the beam.

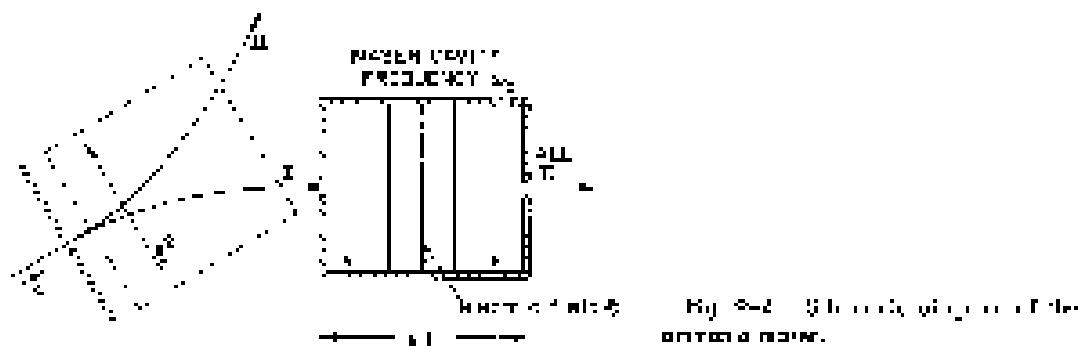


Fig. 6-2. Schematic diagram of the antenna system.

9.2 Transitions in a time-dependent field

In the previous chapter, the problem of an molecule in a state $|j\rangle$ and a constant energy was well discussed, & case in cavity, see below in Fig. 9-4. The other case is discussed. Inside the cavity, there will be a time-varying electric field. In the next problem we must concern to the behavior of a molecule in an electric field that varies with time. We have 2 completely different kind of a problem— one with a time varying Hamiltonian. Since H_0 depends upon t , we may write $H(t)$ and we must determine the time-form of the system under consideration.

To begin with, we must again the equations to be solved

$$i\hbar \frac{dC_1}{dt} = (E_1 - \nu \delta) C_1 - iC_{20}, \quad (9.1a)$$

$$i\hbar \frac{dC_2}{dt} = -iC_{10} + (E_2 - \nu \delta) C_2. \quad (9.1b)$$

To be definite, let's suppose that the electric field is the same (ballistic) for both sites:

$$E = E_0 \exp(i\omega t - k_0 y^{(0)} - eV^{(0)}), \quad (9.37)$$

In 1996, I substituted the frequency $\omega = eV^{(0)}$ by very nearly eV , since the potential is constant at one nanometer (transition $\lambda_{\text{Si}} = 2.35$ nm), but for the time being we want to keep things general, so we'll leave ω having value $eV^{(0)}$. The best way to solve the equations of motion is to find the equilibrium values C_1 and C_2 as we did before, so we will do the two equations about by the methods of 12, and use the definitions of C_1 and C_2 that we had at $\omega = 0$ [P.13]. We get

$$\partial_t \frac{\partial C_1}{\partial t} + iC_1 = 4j(E_0 - eV^{(0)}), \quad (9.38)$$

You'll note that this is almost the Eq. 9.36, except an extra term due to the electric field. Since any j we substitute the two equations (9.36) we get

$$\partial_t \frac{\partial C_2}{\partial t} = (E_0 - 4jC_1) + ieV^{(0)}. \quad (9.39)$$

Now the question is, how to solve these equations? They are more difficult than our earlier set, since j depends on C_1 and, in fact, for a general $j(t)$ the solution C_1 is not representable by elementary functions. However, we can perform an approximation to $j(t)$ if the electric field is small. First, recall what

$$\begin{aligned} C_1 &= i\eta j e^{-i\omega t + i\theta_1} = i\eta e^{-i(\omega t + \theta_1)} \\ C_{11} &= \eta \cos(\omega t + \theta_1) \quad \eta_{11} = \eta \cos(\omega t + \theta_1) \end{aligned} \quad (9.40)$$

If there were no electric field, these equations would be easier with η and η_{11} no longer in their complex conjugates. In 2000, all of the probability $|C_1|^2$ was in state $|1\rangle$ (it is the case when $\omega = 0$) and its possibility of being in state $|2\rangle$ is zero. The square of $|C_1|$ is the probability of being in state $|2\rangle$ or in state $|1\rangle$ is just $|\eta|^2$ or $|\eta_{11}|^2$. For instance, if the system were to start initially in state $|2\rangle$ and η was zero and η_{11} was one, this condition would never change. There would be no chance of the molecule being originally in state $|2\rangle$, over to go to state $|1\rangle$.

Now the idea of solving the equations in the form of Eq. 9.40 is this: if ω is small compared with ω_0 the ω terms can still be written in the form, but then η_1 and η_{11} become slowly varying functions of time, whereas η (nearly constant) becomes rapidly oscillating with the exponential frequency ω , i.e., the phase. We use the last form, η_1 and η_{11} vary slowly, to get an approximate solution.

We want now to substitute C_1 from (9.40) in the left-hand side of (9.38), but we must remember that η_1 is also a function of t . We have

$$\partial_t \frac{\partial C_1}{\partial t} = E_0 \eta_1 e^{-i\omega t} + i\partial_t \frac{\partial \eta_1}{\partial t} e^{-i\omega t}$$

The differential equation becomes

$$\left(E_0 \eta_1 - i\partial_t \frac{\partial \eta_1}{\partial t} \right) e^{-i\omega t} = E_0 \eta_1 e^{-i\omega t} + j_0 e^{-i\omega t} e^{i\theta_1} = j_0 e^{-i\omega t} e^{i\theta_1} (1 + e^{-i\omega t}). \quad (9.41)$$

Similarly, the equation in $\partial C_2/\partial t$ becomes

$$\left(E_0 \eta_{11} - i\partial_t \frac{\partial \eta_{11}}{\partial t} \right) e^{-i\omega t} = E_0 \eta_{11} e^{-i\omega t} + j_0 e^{-i\omega t} e^{i\theta_1} = j_0 e^{-i\omega t} e^{i\theta_1}. \quad (9.42)$$

Now you will observe a second equal term on both sides of each equation. We can cancel terms, and we also multiply the first equation by $e^{i\omega t} e^{i\theta_1}$ and the

stated by (9.40). Remembering that $(\omega - \omega_0) = 2A - \omega_0$, we have finally,

$$\begin{aligned} \frac{\partial^2 Y}{\partial t^2} &= \omega^2 Y e^{i(\omega t - \omega_0 t)}, \\ \frac{\partial^2 y_{12}}{\partial t^2} &= \omega^2 Y e^{-i(\omega t - \omega_0 t)}. \end{aligned} \quad (9.41)$$

Now we have an apparently simple pair of equations - and they are still exactly correct. The derivative of one variable is a function of time only (ωt), multiplied by the second variable; the derivative of the second is a similar time function, multiplied by itself. Although these simple equations cannot be solved in general, we can learn how to solve them in special cases.

We are, for the moment at least, interested only in the case of an oscillating electric field. Taking (9.41) as given, in Eq. (9.37), we find that the equation for y_1 and y_{12} has now

$$\begin{aligned} \frac{\partial^2 y_1}{\partial t^2} + 2\zeta\omega e^{i(\omega t - \omega_0 t)} &= e^{-i(\omega t - \omega_0 t)} p_{11}, \\ \frac{\partial^2 y_{12}}{\partial t^2} - 2\zeta\omega e^{i(\omega t - \omega_0 t)} &= e^{-i(\omega t - \omega_0 t)} p_{12}. \end{aligned} \quad (9.42)$$

Now if ω_0 is sufficiently small, the values of the arguments of Y_1 and Y_{12} are also small. The two $e^{i\theta}$ will not vary much with t , especially for comparison with the exponential factor in the exponential terms. These exponential terms have real and imaginary parts, but now take at the lowest order $\omega = \omega_0$, $\omega_0 \ll \omega$. The term with $\omega = \omega_0$ oscillates very rapidly about its average value of zero and, therefore, does not contribute very much to the average response of the system. So we can make a reasonably good approximation by replacing these terms by their average value, namely zero. We will just leave them in, and take another approximation:

$$\begin{aligned} \frac{\partial^2 Y}{\partial t^2} &= 2\zeta\omega e^{i(\omega t - \omega_0 t)} p_{11}, \\ \frac{\partial^2 Y_{12}}{\partial t^2} &= 2\zeta\omega e^{-i(\omega t - \omega_0 t)} p_{12}. \end{aligned} \quad (9.43)$$

Over the compounding terms, with exponents proportional to $(\omega - \omega_0)$, will also come rapidly unless ω is near ω_0 . Only over the time-scale set by ω_0 does enough time pass for the exponential factor to accumulate. When we integrate the equations over t from $-\infty$ to ∞ , in other words, when a wave electric field the only significant frequencies are those around ω_0 .

With the approximation made in writing Eq. (9.43), the equations can be solved exactly but the work is a bit tedious and, as we want, the first and last time when we take no account problem in the same type. Now ω_0 is not necessarily approximately ω , in fact, not an exact solution for the case of perfect resonance, $\omega = \omega_0$, and an approximate solution for frequencies near resonance.

9-4 Transitions at resonance

Let's take the case of perfect resonance first. If we take $\omega = \omega_0$, the response will be equal to one in both equations of (9.43), and we have just

$$\frac{\partial^2 Y}{\partial t^2} = \frac{2\zeta\omega}{\omega} p_{11}, \quad \frac{\partial^2 Y_{12}}{\partial t^2} = \frac{2\zeta\omega}{\omega} p_{12}. \quad (9.44)$$

If we eliminate p_{12} by adding the two from these equations, we find that each satisfies the differential equation of simple harmonic motion:

$$\frac{\partial^2 Y}{\partial t^2} = \left(\frac{2\zeta\omega}{\omega}\right)^2 Y. \quad (9.45)$$

The general solutions for these equations can be taken as γ of sinus and cosines

As you can easily verify, the following equations are a solution:

$$\begin{aligned}\psi_1 &= \cos\left(\frac{\mu E_1}{\hbar}\right) + \tau \sin\left(\frac{\mu E_1}{\hbar}\right), \\ \psi_2 &= i \sin\left(\frac{\mu E_1}{\hbar}\right) t - i \sin\left(\frac{\mu E_1}{\hbar}\right) \tau,\end{aligned}\quad (9.19)$$

where τ and i are constants to be determined by the particular physical situation.

For instance, suppose that at $t = 0$ our molecular system was in the upper energy state $|E_1\rangle$, where we could replace τ from Eq. (9.19) by $\tau_0 = 1$ and $i\tau = 1$ and $i\tau_0 = -i$. In this situation we would have $\tau = 1$ and $t = 0$, the result being that the molecule is in the state $|E_1\rangle$ or to say it is in the absolute upper state, i.e.,

$$|\psi_1 - \psi_2|^2 = \cos^2\left(\frac{\mu E_1}{\hbar}\right). \quad (9.20)$$

Similarly, the probability that the molecule will be in the lower state is given by the absolute square of ψ_2 :

$$|\psi_2 + \psi_1|^2 = \sin^2\left(\frac{\mu E_1}{\hbar}\right). \quad (9.21)$$

Let us now introduce the time variable, the probability, $P(t)$, is given by simply squaring $|\psi_1 - \psi_2|^2$. The probability to be in state $|E_1\rangle$ will remain constant and back again while the probability to be in state $|E_2\rangle$ goes from zero to one and back. The time variation of the two probabilities is shown in Fig. 9.2. Note again, the sum of the two probabilities is always equal to one, i.e., molecule is always in some state!

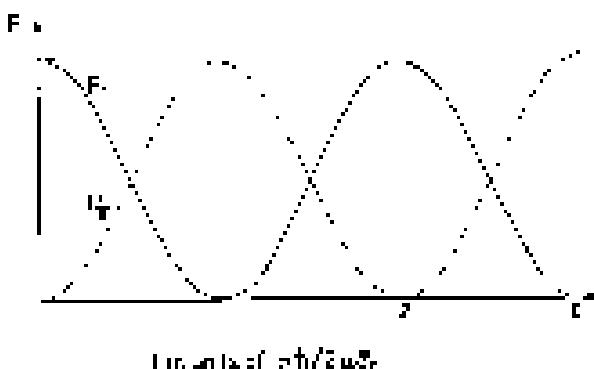


Fig. 9.2. Probabilities for the two states of the diatomic molecule in a sinusoidal electric field.

Let's suppose that it takes the molecule $|E_1\rangle$ the time T to go through one cycle. If we make the cavity gas long enough so that $T/\hbar\mu/E_1^2 = 1/2$, then a molecule which enters in state $|E_1\rangle$ will certainly leave it in state $|E_2\rangle$. If it leaves the cavity in the upper state, it will leave the cavity in the lower state. In other words, if the energy is decreased, and the loss of energy can't go anywhere else but into the environment which surrounds the field. The detectability which you can see gives the energy of the molecule a fed from the oscillations of the cavity is not simple; however, we can't feed the cavity directly because we can't use the principle of conservation of energy. (You could study this if you want, but then we would have to do it with the quantum mechanics of the field in the cavity in addition to the quantum mechanics of the atom.)

In summary, the molecule makes the cavity less cavity oscillating at exactly the right frequency. The excess energy from the upper to the lower state, and the energy released is fed into the oscillating field. If we open the mode, the molecule has enough energy to maintain the cavity oscillations—i.e., providing enough power to keep up for the cavity laser has been pumping enough amounts of excess power. And this is drawn from the cavity. Thus the incident energy is converted from the cavity of unexcited electric magnetic field.

Remember the battery that powers the cavity we were to use a diode which provides the bias so that only one upspin state exists. It is easy to see since that if you were to start with molecules in the lower state, the process will go down and you will end up with all the cavity. Then you cannot heat them so many more because their energy starts too high and energy in, so heating result would happen. To actual operation, it isn't necessary of course to make full T/R cavity w/2. For one other note, keep in mind integer one (that of α), there's some probability for transition from state $|0\rangle$ to state $|1\rangle$ or other values; however, the device isn't 100 percent efficient because of the molecules which have the say β result. They deliver about same energy to the cavity you wish.

In actual fact, the identity of the two molecules is not the same; they have some kind of Maxwell distribution. This means that the ideal conditions of time τ different "electrons" being different, and it is impossible to get 100 percent efficiency for all the molecule to switch. In addition, there is another complication that is easy to take into account, but we don't want to lost it with that the stages. You remember that the electric field in a cavity usually varies from peak amplitude across the cavity. Thus, as the molecules that receive the energy, the electric field at the molecule varies in a way that is more complicated than the simple sinusoidal oscillation in time due to wave function. Clearly, we would have to do a more complicated integration in the time problem exactly, but the general idea is still the same.

The one other way of making measure. Instead of separating the states $|0\rangle$ & $|1\rangle$ from each other by a Stern-Gerlach apparatus, we can have the atoms already in the cavity be exposed to a solid and cold atom beam state. For instance, $E_{\text{beam}} = E_0$. One was being used in the so-called three-state laser. For it when it goes to an atom when has three energy levels, as shown in Fig. 9-2, with the following specific properties. The system will consist of electron (say, light) of frequency ω_0 , and photon the lowest energy level E_0 , the atom, hydrogen by level E_1 and then w/2 quickly and photons at frequency ω_1 and go to the same E_2 same energy E_0 . The only β are a long distance so the separation can be reduced, and the condition that the coupling is weak enough for between states $|0\rangle$ and $|1\rangle$. Although there is obviously no such a "three-state" laser, the laser operation may work just as a two-state system such as we are describing.

A laser (light) at different E_0 called E driven at. Radiation α is just a name working at optical frequencies. The "theory" for a laser uses γ_{01} consists of just two pure numbers between which stimulated waves are generated.

9-5 Transitions off resonance

Finally, we would like to find out how the states work in the circumstance that the cavity frequency is nearby, but not exactly equal to ω_0 . We could solve this problem exactly. In instead of trying to do that, let's take the approximate case that the electric field is small and also the period of time τ is small, so that $\omega_0\tau$ is bounded, as the case. Then, now, in the most important circumstance which we have just worked out, the probability of making a transition is small. Suppose that we start again with $\psi_0 = 1$ and $\psi_1 = 0$. During the time τ we would expect ψ_0 to remain nearly equal to one, and ψ_1 to begin very small compared with ψ_0 . Then the problem is very easy. We can substitute ψ_0 from the current equation for ψ_1 , putting α equal to one and integrating from $t = 0$ to $t = \tau$. We get

$$\psi_1 = \frac{\alpha\psi_0}{\hbar} \left[\frac{1 - e^{i(\omega_0 - \omega_1)\tau}}{1 - e^{-i(\omega_0 - \omega_1)\tau}} \right]. \quad (9.5.1)$$

But ψ_0 , used with Eq. (9.4.10), gives the amplitude to leave molecule excited from the state $|1\rangle$ to the levels ω_1 of the uppermost interval $\Delta\omega$. The probability $P(\psi_1 \rightarrow 0)$ to make the transition is $\sim 1/\tau$.

$$P(\psi_1 \rightarrow 0) \approx |\psi_1|^2 \approx \left| \frac{\alpha\psi_0\tau^2 \sin^2[(\omega_0 - \omega_1)\tau/2]}{1 - e^{-i(\omega_0 - \omega_1)\tau}} \right|^2. \quad (9.5.2)$$

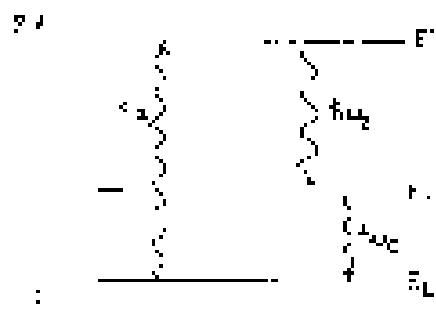


Fig. 9-3. The energy levels of a "three-state" laser.

It is interesting to plot this probability for a fixed length of time as a function of the frequency of the energy in order to see how sensitive it is to frequencies near the average frequency ω_0 . We show an example of $P(\omega) \times N\omega = \text{Fig. 9-4}$. (The vertical scale is logarithmic.) It can be seen by this graph the value of the probability when $\omega = \omega_0$ is unity. We can also calculate this in the Boltzmann theory, so you should already be familiar with it. Then we fully expect the ratio to two for $\omega = \omega_0 + \Delta\omega$ if T is large enough to ignore significant size to large frequency deviations. In fact, however, the greatest part of the area under this curve lies within the range $\pm 10\%$. It is possible therefore that the maximum theoretical probability is better to the level of 10% which is shown in the figure.

Let's examine the implications of our results for a real system. Suppose that the emission makes up in the cavity for a reasonable length of time, say for one millisecond. Then for $N_0 = 24,000$ transitions, we can calculate that the probability for a transition is 6%. In one for a frequency deviation of $\Delta\omega = 500\text{ Hz}$, which is for perhaps 10³. Obviously the frequency must be very close to the typical frequency to sustain reliability. Such an effect is the basis of the great precision that can be obtained with "stabilized" clocks, which work on the same principle.

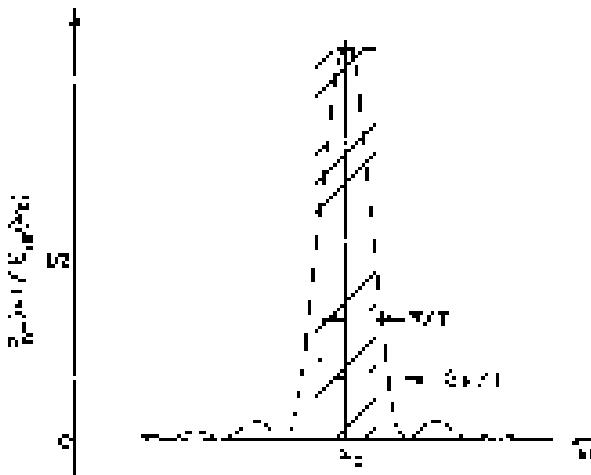


Fig. 9-4. Transition probability for the emitted radiation as a function of the frequency.

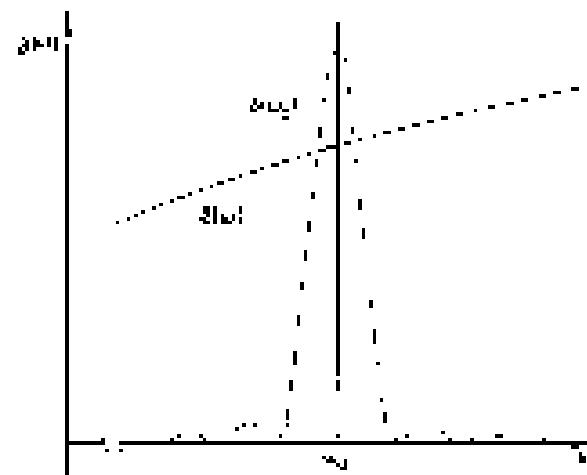


Fig. 9-5. The expected intensity $I(\omega)$ versus approached by the value of ω_0 .

9-6. The absorption of Light

The last result does appear to be more general approach than the one under review. You have noted the behavior of a molecule under the influence of an electric field, whether this field was constant or varying or not. So we could be simply defining a beam of light — microwave frequencies, at the molecule and set it to the geometry of emitting or absorbing. Our equations apply equally well to this case, but let's rewrite the $\langle \sin \theta \rangle$ in terms of the intensity I rather than $\langle E_x \rangle$ and the electric field. This is done in terms γ of the average energy these photons have per second, taken from chapter 27 of Volume II, we can write

$$I = \omega c^2 \gamma / 8\pi c = \frac{1}{2} \omega^2 (\epsilon \times \theta_{max} - \theta_{min})$$

(The maximum value of θ is $2\pi/3$.) The transition probability now becomes,

$$M\Gamma \propto M_0 = \sqrt{\frac{1}{4\pi c^2 \gamma^2}} \frac{\omega^2 \sin [\theta_0 - \omega_0 t] \partial \theta}{\omega - \omega_0/2}, \quad (9.55)$$

*Using the formula $\Gamma_0 = \omega_0 \sin \theta_0/\omega_0 = \gamma$.

Obviously the light striking on such a system is not easily monochromatic. It is, therefore, interesting to take one more path—in that is, to calculate the transition probability when the light has frequency ω_0 per unit frequency interval, covering a broad range rather than $\Delta\omega$. Then the probability of going from $|0\rangle$ to $|M\rangle$ will become an integral:

$$P(0 \rightarrow M) = 2\pi \frac{\omega^2}{4\pi c^2 \hbar \epsilon_0} T \int_{-\infty}^{\infty} d\omega \frac{e^{-i(\omega - \omega_0)t/\hbar}}{(\omega - \omega_0)^2 + \Gamma^2} \rho_M \quad (9.51)$$

In general, ρ_M will vary much more slowly with ω than the sinus parameter here. The raw sinusoid might appear, as shown in Fig. 9.6. In such case, we can let $\rho_M(\omega)$ by its value ρ_M at the center of the sharp resonance curve and take the width of the integral. What replaces is just the integral under the curve of Fig. 9.5, which is, as we have seen, also equal to $2\pi/T$. We get the result that

$$P(0 \rightarrow M) = 4\pi^2 \frac{\omega^2}{c^2 \hbar \epsilon_0 \Gamma^2} \sin^2 \theta \quad (9.52)$$

This is an important result, because it is the whole theory of the absorption of light by any atomic or atomic system. Although we began by considering a state in which state M had a higher energy than state 0 , the sum of our arguments depends on that fact. Equation 9.52 still holds if the state $|M\rangle$ has a lower energy than the state $|0\rangle$ (Section 9.5). It represents the probability for a transition with the absorption of energy from the incident electromagnetic wave. The absorption of light by any atomic system always involves the amplitude for a transition in an oscillating electric field between two states separated by an energy $E = \hbar\omega_0$. For one particular case, it is always worked out in just the way we have done here and given in expression like Eq. (9.52). We, therefore, summarize the following features of this result. First, the probability is proportional to ω^2 . In other words, this is a constant probability per unit time that transitions will occur. Second, this probability is proportional to the intensity of the light incident on the system. Thirdly, the transition probability is proportional to $\sin^2 \theta$, where you remember, as defined, the shift in energy due to the electric field \mathbf{E} . Because of this, ω_0 that appeared in Eqs. (9.3b) and (9.3b) as the coupling term, is responsible for the transition between the otherwise stationary states $|0\rangle$ and $|M\rangle$. In other words, the small t we have been considering, t is the so-called transition time $t = \hbar/\omega_0$ by the coupling interaction term which connects the states $|0\rangle$ and $|M\rangle$. In the general case, we would have this t being replaced by the matrix element $\langle M | H | 0 \rangle$ (see Section 9.6).

In Volume 3 (Section 4.2) I talked about the relations among light absorption, induced emission, and spontaneous emission in terms of the Einstein A -and- G -coefficients. Here, we have to use the quantum mechanical procedure for computing these coefficients. What we have is that $\langle 0 | \rightarrow M \rangle$ for our transition in unit of seconds is proportional precisely to the absorption coefficient A_{0M} of the Tisza radiation theory. For the spontaneous and stimulated transitions, it is too difficult for anyone to calculate. We have taken the matrix element $\langle M | H | 0 \rangle$ as the value that is to be gotten from separation. For simple atomic systems, the A_{0M} which belongs to any particular transition can be calculated from the definition

$$A_{0M} = \langle M | \partial / \partial t | 0 \rangle = H_{0M}, \quad (9.53)$$

where H_{0M} is the matrix element of the Hamiltonian which includes the effects of a weak electric field. This H_{0M} which we in this way is called the electric dipole moment of system. The quantum mechanical theory of the absorption and emission of light is, therefore, reduced to the calculation of the matrix elements for particular atomic systems.

Our study of a single two-state system has led us to an understanding of the general problem of the absorption and emission of light.

Other Two-State Systems

10-1 The hydrogen molecular ion

In the last two chapters we discussed some aspects of the one-magneton theory under the approximation that it can be considered as a two-state system. It is, of course, necessary to re-examine quantum mechanics in terms of realistic vibrational frequencies, and to do so for each of three states of hydrogen must be analyzed in terms of two magneton theories, one because of the spin flip of the nitrogen atom. Here we are going to consider other examples of systems which, as far as the one-magneton theory goes, can be considered as two-state systems. Lots of things will not happen in these cases, but in other states, some of which we will see later, they will now be able to be taken into account. But instead of our examples we will see this in more detail by just thinking about two states.

Now we will move to dealing with two-state systems, the H₂m region, or region I, as just like the one we had in the last chapter. When the Hamiltonian is independent of time, we know that there are two energy levels with different and equal frequencies. Generally, however, we start our analysis with one of these states which is not these stationary states, but since we may, perhaps, have some other simple classical theory. Then, the state in place of the system will be represented by a linear combination of these two states.

For convenience, we will combine the equations from Chapter 9, of the original theory of two states by ψ_1 and ψ_2 . Then the state ψ is represented by the linear combination

$$\psi = \alpha\psi_1 + \beta\psi_2 = \alpha\psi_1|\psi\rangle + |\psi\rangle\langle\psi_1| - \beta\psi_2|\psi\rangle. \quad (10.1)$$

The amplitudes ψ_1 and ψ_2 are denoted C_1 and C_2 respectively in the initial equations:

$$\partial_t \frac{\partial \psi}{\partial t} = \sum_i \partial_i \psi_i, \quad (10.2)$$

where ∂_i is any operator on the values ψ_1 and ψ_2 .

Within the terms of the ψ equation, ∂_i denotes expansion of the two-state combinations ψ_1 and ψ_2 (the stationary states), which we call

$$|\psi_1\rangle = (\psi_1^x e^{-iH_0t/\hbar} - i\psi_1^y) + (\psi_1^y e^{iH_0t/\hbar} + i\psi_1^z),$$

and the conjugate

$$\psi_1^x = \frac{H_{01} - H_{11}}{\hbar} + \sqrt{\left(\frac{H_{01} - H_{11}}{\hbar}\right)^2 - H_{11}H_{01}}, \quad (10.3)$$

$$\psi_1^y = \frac{H_{02} - H_{12}}{\hbar} + \sqrt{\left(\frac{H_{02} - H_{12}}{\hbar}\right)^2 - H_{12}H_{02}},$$

The ψ_1 and ψ_2 basis states of the ψ state have the same time dependence. The three states $|\psi\rangle$, $|\psi_1\rangle$ and $|\psi_2\rangle$ are very closely related; the stationary states are related to the original base states $|\psi_1\rangle$ and $|\psi_2\rangle$ by

$$|\psi\rangle = \alpha|\psi_1\rangle + \beta|\psi_2\rangle, \quad (10.4)$$

$$|\psi_1\rangle = |\psi\rangle\langle\psi_1| + |\psi\rangle\langle\psi_2|,$$

10-2 The hydrogen molecular ion

10-2 Nuclear forces

10-3 The hydrogen molecule

10-4 The boron molecule

10-5 Dyes

10-6 The Methionine a spin-half particle is a magnetic field

10-7 The spinning electron is a magnetic field

These are complex constants, which satisfy

$$|\alpha_1|^2 + |\alpha_2|^2 = 1$$

$$\frac{\alpha_1}{\alpha_2} = \frac{H_{11}}{H_{22}} \quad (10.5)$$

$$|\beta_1|^2 + |\beta_2|^2 = 1.$$

$$\frac{\beta_1}{\beta_2} = \frac{H_{12}}{H_{22}} = \frac{H_{11}}{H_{11}} \quad (10.6)$$

If H_{11} and H_{22} are equal—so that α_1 is equal to β_1 , and $H_{11} = H_{22} = -A$, that $E_1 = E_2 = A$, $\alpha_{11} = \beta_1 = A$, and the states $|E\rangle$ and $|H\rangle$ are particularly simple:

$$|E\rangle = \frac{1}{\sqrt{2}}(|1\rangle - |2\rangle) \quad |H\rangle = \frac{1}{\sqrt{2}}(|1\rangle + |2\rangle) \quad (10.7)$$

Now we will use these results to discuss a number of interesting examples taken from the fields of chemistry and physics. The first example is the oxygen molecule ion. A positively charged oxygen molecule contains two protons with one electron, occupying its very outer shell. If the two protons are far apart, what state would we expect for this species? The answer is pretty clear. The electron will stay close to one proton and leave a hydrogen atom. This leaves strong, and the other proton will remain alone as a positive ion. So, if the two protons are far apart, we can visualize one physical state in which the electron is "attached" to one of the protons. There is, though, another state of the molecule in that region in which the electron is near the other proton, and the first proton is the one that is un-ionized. We can take these two as our basis states and will call them $|1\rangle$ and $|2\rangle$. Using the rules of $\langle \cdot | \cdot \rangle$ given in Fig. 10-1, for instance, there are really many more than three electrons in the proton, because the ionization can occur at any one of the excited states of the oxygen atom. We are not interested in most varieties of wave functions; we will consider only the situation in which the hydrogen atom is in the lowest state, its ground state—and we will, for the moment, disregard spin of the electron. We can just suppose that for $|1\rangle$ the electron has no spin ("up") along the \hat{x} -axis.⁴

Now to perform an electron-for-a-hydrogen atom requires 13.6 electron-volt energy. Since each of the two protons of the oxygen molecular ion are far apart, it will require about this much energy—which is for a $1s$ proton something like a great deal of energy—to get the electron somewhere near the midpoint between the protons. So it is a surprising discovery, for the first time, to jump from one proton to the other. It is, however, in quantum mechanics, it is possible—but not very easily. There is some small amplitude for the electron to move from one proton to the other. As a first approximation, then, each of our basis states, $|1\rangle$ and $|2\rangle$ will have the energy E_0 , which is just the energy of one hydrogen atom plus one proton. We can take that the $\langle \cdot | \cdot \rangle$ matrix elements H_{11} and H_{22} are both approximately equal to A . The other matrix elements H_{12} and H_{21} , which are the coupling between the two orbitals of "proton," we will again set equal to zero. This is the same spirit as the one we followed in the last two chapters. If we disregard the fact that the electron can flip back and forth, we have been taken exactly the same energy. This energy will, however, be split in a two-energy levels by the coupling of the electron going back and forth—the symmetric probability of the transition, the quantity $\langle \text{spf} \rangle$. So our two energy levels of the system are $E_0 + A$ and $E_0 - A$, and the values of the other two different energies are given in Fig. 10-8.

⁴ This is considerably as large as current superconducting fields. We will discuss the nature of magnetic fields in the chapter four in this chapter, and they will illustrate them in the solid-state atom in Chapter 12.

Electrons, of course, we say that it is proton and a hydrogen ion are just moving near together, the electron will not stay on one of the protons but will flip back and forth between the two protons. If it reaches one of the protons, it will oscillate back and forth between them. It does this, giving a time-varying solution. In order to have the lower energy solution which does not vary with time, it is necessary to heat the system with certain amplitudes for the electrons to have random motion. Remember, thermal motion of electrons were not saying that there is an electric current and a photon. There is only one electron and it has its own amplitude. Fig. 10-3 corresponds to be in either position.

Now the amplitude of the electron which is due to the exchange of the outer core depends on the separation between the protons. As those are protons and together, the larger the amplitude. You remember that we talked in Chapter 10 that the coupling between the proton and the "polarization field" which is called dielectric. We have the same coupling here. The amplitude of the electron will go to zero as the distance between the protons goes over. Since the term δ_0 is proportional to $1/r$, where r is the large when the protons are very far apart, the separation of the energy levels will be yet larger. If the system is in the state of the energy $E_1 = E_0 + A$ increases with decreasing the separation between the two protons. This is because the energy of the electron is increasing due to the influence of the distance. These photons are becoming closer together, there is no problem from putting the photons together. The variation of the two energy levels with the distance between the two protons are shown in Fig. 10-4. We know, there is a mathematical explanation of the hydrogen atom and both the E_1 and E_2 are.

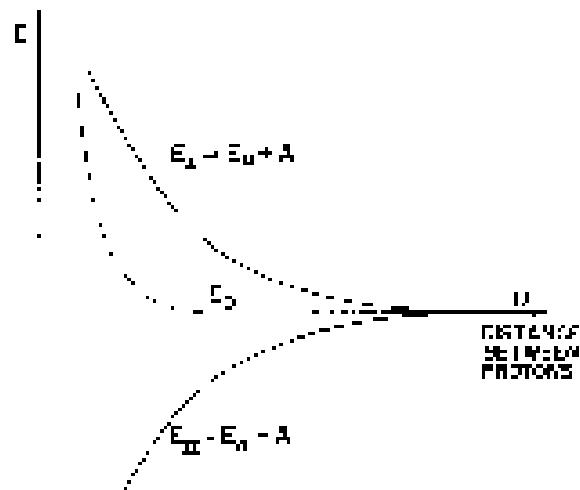


Fig. 10-4. The energies of the two energy levels in the H_2^+ ion as a function of the distance between the two protons.

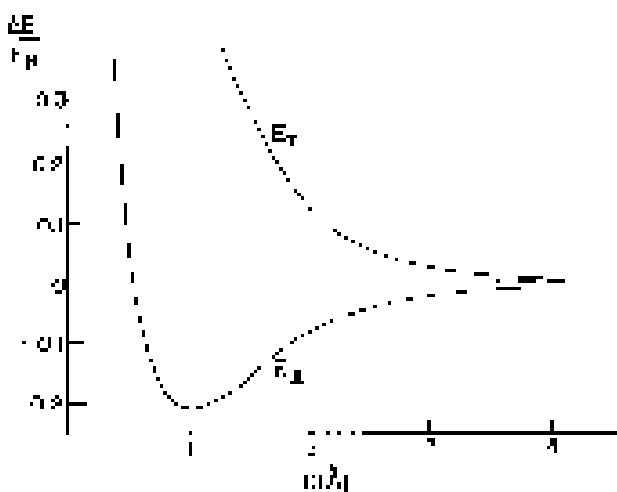


Fig. 10-5. The energy levels of the H_2^+ ion, finite and its separation is denoted. ($E_i = 10.6 \text{ eV}$)

We have, however, forgotten one thing. In addition to the above we have just described, there is also an electrostatic repulsive force between the two protons. When the two protons are far apart, as in Fig. 10-4, the "dielectric" or the central force or there is a negligible electrostatic force. At very short distance, however, the "proton" should begin to feel "proton" the electron distribution, that is it is closer to the proton on the average than to the electron. So there begins to be some sort of electrostatic energy which is, of course, positive. This energy which is in varies with the separation—should be written in Fig. 10-4 or 10-5, we speak, that comes up like the broken line curve in Fig. 10-2 when this happens, first starts low then it will increase a little, goes up, then it will decrease and then the final energy is from this E_1 . When we do this, the energies E_1 and E_2 will vary with the two-proton distance d as shown in Fig. 10-5. (In this figure, we have plotted the results of a more detailed calculation. The interproton distance

brighter than $1.5 \mu\text{J}/\text{cm}^2$, and the excess energy can strip off a hydrogen atom's electron, in which the binding energy of the hydrogen atom, the so-called "ionizing" energy, is 13.6 eV . We see that the state $|1\rangle$ has a minimum-energy point. This will be the equilibrium configuration—minimum-energy condition for BeH_2 ion. The energy of this point is finite, i.e., the sum of a two-electron potential energy and a zero-point energy. The system is bound. A single electron adds to make the two electrons repel each other. A chemist would call this a "molecular hybrid."

This kind of chemical bonding is also often called "electron-mechanical resonance." By analogy with the heat capacity problem we have described before, this name really sounds more mysterious, but it is only a "resonance" if you start from an atomic or ionic state for your basis states, as we did, and if you pick the same H_0 you would find the resonance state, the $|S\rangle$.

We can see at once the very why such a system should have a lower-energy, but a positive, total hydrogen atom. Let's think about an electron and two protons with some fixed, but not too large separation. You remember that with a single proton the electron is represented by a set of the uncertainty principle. It needs to balance between having a low Coulomb potential energy and not being confined in a too small space, which would make the known energy decrease of the uncertainty relation disappear. Now, if there are two protons, there is more room, so the electron can have a less electrical energy. It can spread out, lowering its kinetic energy without increasing its potential energy. The result is a lower energy than a hydrogen atom. Then why does another state, $|1\rangle$, have a higher energy? Notice that this state is the rightmost of the states $|1\rangle$ and $|2\rangle$. Because of the symmetry of $|1\rangle$ and $|2\rangle$, the difference must have zero component, so therefore it cannot fully be seen in the ψ_{12} profile. This means that the electron is somewhat more confined, which adds to a larger energy.

We should say that the approximate treatment of the H_2 ion as a two-electron system breaks down pretty badly near the minimum point, since either of them are at the minimum of the curve of $E(\text{H}_2)$, and so will not give a good value for the total binding energy. For a full calculation, the energies of the two "hybrids" we've given in Fig. 16.11 cannot really explain the experimental quantity. A detailed treatment is needed.

Suggest we ask now what would happen if instead of two protons, we had two different projectiles, say, for example, one proton and one lithium projectile. In such a case, both projectiles still have a single positive charge. In such a case, the two terms R_{11} and R_{22} of the Hamiltonian would no longer be equal, they would be just being different. If it should happen that the difference $(R_{11} - R_{22})$ is an absolute value, much greater than $4\pi\epsilon^2/(M_1 + M_2)$, then the curve looks just very weak, as we see in the following way.

If we put $R_{11}, R_{22} = 1.1 \times 10^{-10}$ in Eq. (16.3) we get

$$E = \frac{R_{11} + R_{22}}{2} - \frac{R_{11} - R_{22}}{2} \sqrt{1 + \frac{4\pi^2\epsilon^2}{(M_1 + M_2)^2}}$$

With $R_{11} + R_{22}$ much greater than $4\pi\epsilon^2/(M_1 + M_2)$ this expression is very nearly equal to

$$\frac{1}{2}(R_{11} + R_{22})$$

The two energies are then

$$\begin{aligned} E_1 &= R_{11} + \frac{\epsilon^2}{(M_1 + M_2)^2} \\ E_{22} &= R_{22} + \frac{\epsilon^2}{(M_1 + M_2)^2} \end{aligned} \quad (16.2)$$

The difference vanishes just the energies R_{11} and R_{22} of the isolated atoms, plus apart only slightly by the fly-by amplitude α .

The energy difference $E_1 - E_{22}$ is

$$(R_{11} - R_{22}) + \frac{2\pi^2\epsilon^2}{(M_1 + M_2)^2}$$

The distance separation from the Rb atom of the electron is no longer small (≈ 2.1 , it is smaller by the factor $A/(R_{H2}) = B_{H2}$), which we are now taking to be much less than one. As in the dependence of $E_0 = 1/r_0$ on the separation of the two nuclei is much smaller than for the H_2 case. It is also reduced by the factor $A/(R_{H2}) = B_{H2}$. We see that the binding of atoms with different nuclear charges is generally very weak.

In our theory of the H₂ ion we have discovered an explanation for the decrease in the ΔE due to the fact that the two protons involved in effect act as active 'bonds' between the two protons which can be present even when the protons are at large separations. The attractive force comes from the reduced energy of the system due to the possibility of the electron jumping from one proton to the other. In such a jump the system changes from the configuration hydrogen atom, proton to the configuration (proton, hydrogen atom) or switches back. We can write the process symbolically as

$$(H, p) \leftrightarrow (p, H)$$

The energy shift due to this process is proportional to the amplitude of the two electrons whose energy is $-kE$ (the binding energy of the hydrogen atom) and goes from one proton to the other.

For large distances R between the two protons, the interaction potential energy of the electron is nearly zero over most of the space it must go when it makes its jump. In this case then the electron moves nearly like a free particle in empty space but with a negative energy. We know from in Chapter 3 (Eqn (3.1)) that the energy with some constant of proportionality to get them one plus the absolute distance energy is proportional to

$$\frac{p^2}{r} - \frac{1}{r}$$

where p is the momentum corresponding to the infinite energy. In the present case (using the non-relativistic formula) p is given by

$$\frac{p^2}{2m} = -kE \quad (10.10)$$

This gives a finite bound in infinite numbers.

$$p = \sqrt{2mkE}$$

(the other sign for the radial part is unimportant).

We should expect then, that the amplitude A for the H₂ ion will vary as

$$A \sim \frac{\sqrt{-2k^2 E^2 m}}{R} \quad (10.11)$$

at large separation R between the two protons. The energy shift due to the electron hopping is proportional to A , so there is a force pulling the two protons together which is proportional to large R (to the derivative of (10.10) with respect to R).

Finally, as an example, we should remark that the two-proton interaction potential is still too exact since which gives a dependence of the energy on R . We have neglected it until now because it is usually a low-order perturbation. The exception is just for R at very large distances where the energy of the exchange term is not suppressed exponentially to very small values. The new effect we are thinking of is the electrostatic attraction of the nuclei by the hydrogen atoms which remain static in the same way that the nuclei do not move in our theory. The two protons make an electric field E varying as $1/R^2$ at the two hydrogen atoms. The nuclei becomes polarized, taking on net induced electric moments proportional to R . The energy of the system is E_0 which is proportional to $1/R^2$ or $1/R^3$. So there is a $-kE$ in the energy of the system which decreases with red double power of the distance. It is a correction to E_0 . This energy called well

distance more slowly than the with β given by (10.10). At some large separation R , it becomes the only remaining component, because the interaction of energy with β is zero, therefore the energy decreasing term. Even though the electrostatic term has the same sign as the bare α' the bare α' the force is attractive, so the energy is negative and smaller for the two stationary states, whereas the different contributions from β gives opposite signs for the two stationary states.

10-2 Nuclear forces

We have seen that the system of a hydrogen atom has a total *negative* energy of interaction due to the exchange of the single electron, which makes it large enough to form a H^- .

$$\frac{e^2 \alpha'}{R} \quad (10.11)$$

such a $\alpha' = -\sqrt{2\pi}/(m_e R)$. (Remember why this is an *exchange* of a "virial" electron between atoms—the electron has to jump across a barrier where it would lose a tiny bit energy. More specifically, a "virial exchange" means that the production involves a *symmetric* exchange, interchanging between an *interacting* state and a *noninteracting* state.)

Now we might ask the following question: Could it be that there is between a $n=1$ state of particles p and $n=2$ another origin? What about, for example, the nuclear force between a neutron and a proton, or between two protons? Is it also due to *exchange*? The name of Yukawa is associated with the force he wrote, the nuclear force is due to a similar exchange effect—only, in his case, due to the virtual exchange, not of an *exotic*, but of a new particle, which is neither a "meson," nor a "baryon." Today, we would identify Yukawa's meson with the ρ meson (a "pion") produced by high-energy collisions of protons or deuterons.

Let's see, as an example, what kind of a force we would expect from the exchange of a positive pion (ρ^+) and $n=2$ between a neutron and the $n=1$ electron. Just as a hydrogen atom H^+ can go to a g^- state by giving up an electron:

$$H^+ + e^- \rightarrow g^- \quad (10.12)$$

a scalar ρ^+ can go into a neutral π^0 by giving up a π^0 meson:

$$\rho^+ \rightarrow \pi^+ + \pi^0 \quad (10.13)$$

So if we have a proton at r_1 and a neutron at r_2 separated by the distance R , the proton can "lose" its ρ^+ by emitting it, ρ^+ , which is then absorbed by the neutron in a "turning it into a proton." This is an energy of interaction α' of the two nucleon (plus pion) system which depends on the amplitude A for the pion exchange, just as we found for the electron exchange in the H^- case.

In the process (10.13), the energy of the ρ^+ is transferred to one of the pions by the ρ^+ scattering (antirelativistically), and emitting the rest energy $m_\rho^2 c^2$ of the electron, and the electron loses negative kinetic energy (imaginary contribution \rightarrow in Eq. (10.9)). In the nuclear process (10.12), the proton and neutron have a more rapid transfer, so they $\rightarrow \pi^0$ with zero gain in energy. The relation between the total energy E and the incompatibility (or δ) will be discussed.

$$E^2 = p_{\rho^+}^{2c^2} + m_\rho^2 c^2$$

Since E is scalar (and $m_\rho^2 c^2$ is symmetric with ρ), the ρ energy can be again imaginary:

$$y = i m_\rho c$$

Using the same arguments we gave for the amplitude that a bound electron would penetrate the barrier, the cross-section, as p varies, we would find that the rate of exchange amplitude \propto other quantity \propto $y^2 \propto 1/R^2$ (Eq. 10.10).

$$\frac{\sigma \propto \pi y^2}{y} \quad (10.14)$$

the interaction energy is proportional to λ , and so works in the same way. Yet the energy variation is the sum of the scattered momenta divided by mass two volumes. Indirectly, we obtain this same formula starting directly from the differential equation for the motion of a proton in free space (see Chapters 26, Vol. II, §2, §§4, 6).

We can, following the scheme of Feynman, discuss the interaction between two protons (or between two neutrons) which results from the exchange of a virtual pion (π^0). The basic process is now

$$p^+ + \bar{p}^+ \rightarrow \pi^0. \quad (16.1)$$

A pion and an electron virtual π^0 , but the pions and a proton. If we have two protons, proton No. 1 can emit a virtual π^0 which is absorbed by proton No. 2. At the end, we still have two protons, but this is scattered differently from the E_2 law. This is the J^P law in a different situation—the proton—electron coupling. We assume. Now we are assuming that a proton is emitted without carrying its charge e . Such processes are, in fact, observed in $p\bar{p}$ and $n\bar{n}$ collisions. The process is analogous to the ionization electron emits a photon and ends up still in electron:

$$e \rightarrow e + \text{photon}. \quad (16.2)$$

We do not feel the photons emitted by electrons, unless they are emitted so often they are detectable, and then emission does not change the "masses" of the electrons. Coming back to the two proton case, we can imagine energy which arises from the amplitude of the one-pion-exchange process, given which nucleon emits (with its existing momentum) a virtual pion field is absorbed. This amplitude is again conserved in (16.1), without the loss of the normal even. All the symmetries give an equal interaction energy for both nucleons. Since the nuclear charge symmetry of the Coulomb effect between neutron and proton, between proton and proton, between neutron and neutron are the same, we conclude that the sum of the charged and neutral parts should be the same. Experimentally, the cross-sections are nearly equal, and no real difference is found when one varies either their electric fields (see Chapters 10, 11, of Vol. II).

There are other kinds of particles, like Δ -mesons—which can be used to penetrate the nucleus. It is also possible to have pions to be exchanged at the same time. But all of these other exchanges “objects” have a total mass M , greater than the pion mass, and lead to terms in the exchange amplitude which vary as

$$\frac{1}{M^2} \sim \frac{1}{m^2},$$

These terms die out faster and faster as increasing M than the pion mass. Thus the pion theory, being the lowest mass terms, can for large enough values of R justify the one-pion-exchange theory. And, indeed, these experiments which involve nuclear interactions using at large distances, show us that the interaction comes from the one-pion-exchange theory.

The classical theory of electricity and magnetism, the so-called electrostatics is interaction and the radiation of light by an accelerated charge are closely related—the first one of the two well-known laws. Yet have seen in Chapter 10 that light can be represented as the wave propagation of the harmonic oscillations of the classical electromagnetic fields E and B . Alternatively, the quantum theory can be derived by it, starting with the quantum mechanics—another—so-called Dirac theory. We emphasize in Vol. II, §3, that the two otherwise prima facie always give identical predictions. Can the general point of view be carried through completely to include all electromagnetic effects? In particular, we want to describe the electromagnetic field precisely in terms of Dirac particles—that is, in terms of photons, what is the evidence here for this?

From the “quantum” point of view the two-body interaction between two electrons comes from the energy of a virtual photon. One electron emits a photon—say in beam α (Fig. 16.1), which goes over to the second electron, β , and is absorbed in the reverse of the interaction. The interaction energy is again given

by a formula like (10.14), but now with α replaced by the mass of the photon which is e^2 . So the virtual exchange of a photon between two electrons gives an interaction energy that varies simply inversely as R , the distance between the two electrons, just the normal Coulomb potential energy. In the particle¹⁷ theory of electromagnetism, the process of virtual photon exchange gives rise to all the quantum-mechanical interactions.

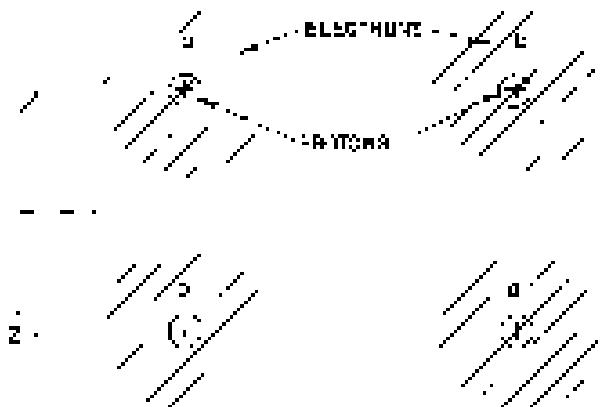


Fig. 10.4. A set of basis states for the H_2 molecule.

10-3 The hydrogen molecule

As our next two-particle system we'll look at the neutral hydrogen molecule. This is naturally more complicated than hydrogen because there are electrons again. We start by thinking of what happens when the two protons are well separated. Only one of the two electrons is odd. To keep track of them, well call one of them "electron 1" and the other "electron 2". We can again imagine two possibilities. One possibility is that "electron 1" is grouped with the first proton and "electron 2" is around the second, as shown in Fig. 10.4(1). We can equally well hydrogen atoms. We will call this state |1>. There is another, the possibility that "electron 1" is around the first proton and that "electron 2" is around the second. We call this state |2>. Now the symmetry of the situation. These two possibilities should be energy levels, at similar, but as we will see, the energy of one hydrogen atom. We could mention that there are many other possibilities. For example, "electron 1" might be near the first proton and "electron 2" might be in another state around the new center. We'll discuss that kind of case, since it will certainly have a large change from one of the large coherent oscillations between the two electrons. For your familiarity, we could have to include such cases, but we can get the essentials of the molecular binding by considering the two cases of Fig. 10.4. To do approach, let us call ψ_1 the wave function for state |1> and ψ_2 the wave function for state |2>, and let ψ be the total wave function:

$$|\psi\rangle = \sum_i |\psi_i\rangle |s_i\rangle$$

To proceed, we assume—as usual—that the wave amplitude A_{11} for the electron 1,1 state through the interacting atom and hydrogen nuclei. This "readily" exchange means that the energy of the ground state, as we have seen, is rather exact the system. As for the hydrogen molecule, i.e., the splitting, is very small when the distance between the hydrogen atoms. As the protons approach each other, the amplitude for the electron 1,1 goes back and forth decreases, so the splitting increases. The reason of this behavior is quite simple: that there's an attractive force which pulls the atoms together. Again the energy levels are when the protons are very close together, because of the coulomb repulsion. The zero point in total wave function would have energy which vary with the separation as shown in Fig. 10.5. At a separation of about 0.75 Å, the zero energy

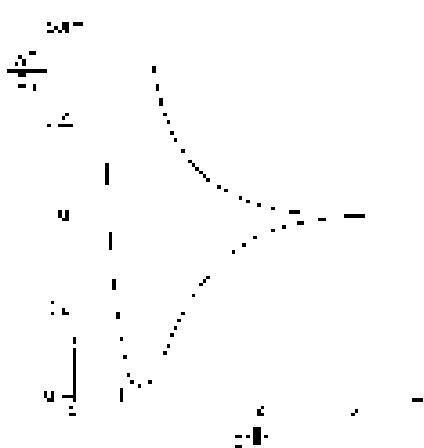


Fig. 10.5. The zero-point energy E_0 of the H_2 molecule for different distances. ($m_e = 0$; $R_0 = 1.2 \times 10^{-8} \text{ cm}$)

In *hydrogen combination*, this is the pre-reaction disease of the two hydrogen molecules.

Now you have probably seen thinking of an oxygen atom. What about the fact that the *w-electrons* we had in 1g₁ are now? We have been talking from previous chapters about 1g₁, "but there really is no such 1g₁—tell what is what." And we have said in Chapter 4 that the electrons which are fermion particles. If there are two electrons being run together by exchanging the electrons, the wave amplitudes will interfere with the same sign. This means that the wave of a *w-electron*, i.e., which, the sign of the amplitude must be +. We have just concluded, however, that the bound state of the hydrogen molecule would be $\psi = -\psi_1$.

$$\psi = \frac{1}{\sqrt{2}}(\psi_1 + \psi_2).$$

However, according to our rules of Chapter 4, this state is not allowed. It is *w-electron* which we get the wave

$$\frac{1}{\sqrt{2}}(\psi_1 - \psi_2),$$

and we get the same sign instead of the opposite one.

But hydrogen does not run if both electrons have the same spin. Let us then if both electrons have spin up (or both have spin down), the wave state for *w-electron* is

$$\psi = \frac{1}{\sqrt{2}}(\psi_1 - \psi_2)$$

for this state, an in-phase state of the two electrons gives

$$(\psi_1 - \psi_2)$$

which is $= \psi_1$, as required. So *w-electron* hydrogen comes pair to each other with their spins pointing in the same direction, they can go to the state $|1\rangle_1$ and not state $|1\rangle_2$. $|1\rangle_1$ under this state, E_1 is the lower energy state. The curve of energy versus r has no minimum. The two hydrogen will always repel and ψ_1^2 will form a molecule. So we conclude that the hydrogen molecule cannot exist with parallel electron spins. And that is right.

On the other hand, for a site M it is relatively easy to make fermion wave functions. In fact, if we interchange which location we call M and which we call N we get exactly the same state. We saw in Section 4.1 that if two Fermi particles are in the state $|1\rangle_1$, they annihilate instead of going $|2\rangle_1$, the bound hydrogen molecule and two free electrons, ψ_1^2 spin up are one with each other.

The whole story of the hydrogen molecule is really completed more completely if we want to include the problem of spin. But then no longer right to think of the molecule as a two-state system. It should really be looked at over all spin states—there are three possible spin arrangements for each *w-electron* states, $|1\rangle_1$ and $|2\rangle_1$ so we want to take things a little finer by neglecting the spins. Our first conclusions are, however, correct.

We find that the *w-electron* wave—the only fermion state of the H₂ molecule has the same character with some exception. The total spin angular momentum of the atom is zero. On the other hand two empty hydrogen atoms or hyper-pairing—*w-electron* is definitely unconstitutional, e.g., must have higher spin-bound state energy than the atoms alone will do. There is an interesting correlation between the atoms and the energies. It gives another illustration of something we mentioned before, which is that there appears to be an “in reaction” energy between two spins because the case of parallel spins has a higher energy than the opposite case. In a certain sense you could say that the spins like to react or interact in certain ways, giving an attractive potential in Fermi energy, but because there is a long-range Coulomb force, but because of the restricted principle,

We saw in Section 10.3 that the binding of two electrons by a single electron pair is likely to be stronger than that of two pairs by two electrons. Suppose that in bromoethane in Fig. 10.4 were replaced by say two ions with identical inner shells with a single innermost shell, and that the binding energies of an electron at the two sites are different. The energies of a pair of 1⁻ and 2⁻ would still be equal because in each of these sites we have one electron bound to each pair. Therefore, we design here by splitting proton bond 10.4 into two separate bonding is equivalent to the most extreme valency limit, i.e., having roughly twice as many H atoms as do the 'steps' of ten electrons. Although two atoms can be bonded together by valency sharing, it is relatively rare when we do this, and the result is very weak.

Finally, we want to mention that if the energy of attraction for an electron in one nucleus is much greater than to the other, then what we have will be different sharing rather than with a pair of ten electrons. Since one nucleus of C may be a positive and has a much stronger attraction for an electron than does bromine, it may seem bizarre that the total energy is still fairly low even when the electrons are stuck to a single electron pair nucleus A. This strong attraction may force the electrons to form the mutual repulsion of the two molecules. If it does, the lower energy state is as shown: large enough to bind both electrons at a binding distance to form a single compact molecule that is chemically inert. The molecule looks like a neutral ion with a positive end. This is, of course, what happens in a "ionic" molecule like NaCl. You can imagine all the gradients between covalent bonding and ionic bonding possible.

You can now begin to see why it is that much of the basis of chemistry can be clearly understood in terms of a quantum mechanical description.

10.4 The benzene molecule

Chemists have invented nice diagrams to represent complicated organic molecules. Now we are going to draw one of the most interesting of them—the benzene molecule shown in Fig. 10.6. It contains six carbon and six hydrogen atoms in a symmetrical ring. Each horizontal diagram represents a row of electrons, with spins opposite, along the covalent bond lines. Each hydrogen atom contains one electron and each carbon atom contains two lone electrons to make up the total of 12 electrons in each of the six covalent bonds. These are not enough bonds that are so tightly bound that they are not appreciably involved in the overall binding. So with six in the home representation, or w/ 12 electrons, we can interpret this as meaning that there are no paired electrons between the four pairs of carbon atoms.

This is a mystery about the benzene molecule. We know, however, that while energy alone should be required to form a branched compound, because the chlorine has saturated the energies of various compounds while needs pieces of training—such as those that cause the energy of resonance (and by a dynamic theory, are so far we can interpret, to double the total energy we should expect for the benzene

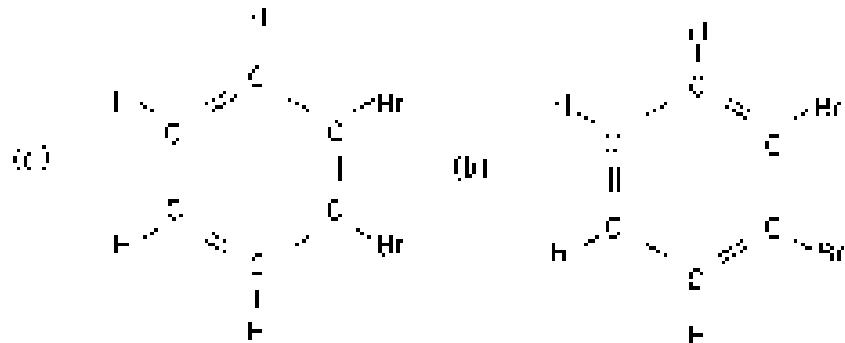


Fig. 10.7. Two possibilities of the benzene molecule. The two bonds could be satisfied by a single bond or by a double bond.

unstable. The actual energy of the pentadienyl cation, however, is much lower than we get by such a calculation; it is only slightly higher than we would expect from what is called the "vertical ionization potential,"¹ Using a double bond system which is not in such a trap is easily attacked chemically because it has a lower the high-energy— π —double bonds can be easily broken by the addition of other hydrogens. But at higher the ion is *very* permanent and hard to break. In fact, bonds between two π -systems have much lower energy than you would calculate from the bare potential.

Then there is another mystery. Suppose we replace two adjacent leg edges by horizontal ones to make ethylene-trimethylene. There are two ways to do this, as shown in Fig. 10-5. The structure could be (a) the separation of a double bond as shown in part (a) of the figure, or it could form the two ends of a single bond as in (b). One would think that these diisobutene-isomers should have different forms, were they real. There is only one such diisobutene!†

Now we want to know about these isomers—and perhaps you have already guessed this by noticing that one, the "ground state" of the hexene molecule is really a three-leg system. We need imagine that the double bonds would be in either of the two arrangements shown in Fig. 10-6. You may not really think so, but they should have the same energy. In fact, they should. And for this reason they must be analyzed as a two-state system. But with appropriate *orbital configuration of the whole set of electrons*, there is some possibility of that the vertical ionization potential of one arrangement to the other is so close that the electrons can move from one to the other.

As we have seen, the ground state triplet makes a mixed π -hexene energy is lower than you would calculate by looking separately at either of the two isomers in Fig. 10-5. Below them are two enthalpy states, one with a π -hexene and one with an energy below the enthalpy value. Recall why. In the hexene state (lowest energy), a hexene molecule is in the possibilities shown in Fig. 10-5, but it has the amplitude $\langle \psi_1 | \psi_2 \rangle$ which is each of the states shown. It is the only wave function involved in the chemistry of benzene at normal temperatures. Fortunately, the upper state also exists, because it is the hexene because has a strong absorption for ultraviolet light in the frequency $\nu = 142 \times 10^{12} \text{ cm}^{-1}$. You will remember this in ammonia, where this absorption, back and white, these subjects, the energy separation is in the microwave region. In benzene, the absorption is one and because they are much lighter, they find it easier to fly back and forth, which makes the oscillations very much longer. The result is that the energy difference is much larger, about 1.5 eV, which is the energy of an ultraviolet photon.

What happens? We submit our request! Again the two "models" (a) and (b) in Fig. 10-5 represent the two stable and close energy isomers. The only difference is that the two have states which with enough energy cause transitions. The lowest energy π -hexene does not involve a π -hexene combination of two states, but with unequal amplitudes. The amplitude for state (b) might have a large probability like $\sqrt{2}/3$, say, whereas state (a) might have a small probability

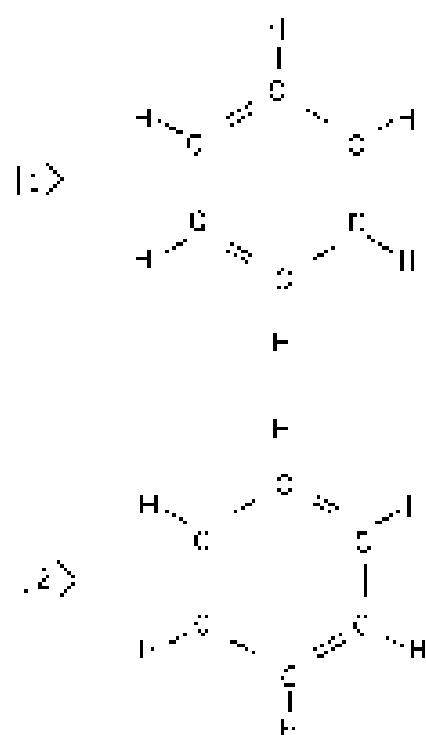


Fig. 10-6. A set of four states for the benzene molecule.

† We are over-simplifying a little. Originally, the extreme changes that these could bring to some of the hexene isomers, would cause the electrons on adjacent carbon atoms (or two others) to swap with the electrons on adjacent carbon atoms (or two others) and so forth. Then since the hexene species in nature are *resonance structures*, however, they would only these forms—that is why we find all the endo-ene-alkanes.

‡ If not too long and not too intense. A emission of ultraviolet light would very easily is the loss of electrons and return to hexene, but not the displacement of the electron between the two atoms. The hexene is then definitely substituted, as in δ,γ furan. Eq. (9) for the probability of a transfer of the triple bond to the π -hexene ligand is (law 16d). If these were the only states, the chance of a transfer would have to be about 1 in 10000. As it happens there are several hexene states with the same base state (such as three hexene states in the hexene shown) and the ratio of any state of hexene are slightly different from the two states found. The ratio of hexene states (at the transition) is not known, but we can assume it is proportional to the absorption of ultraviolet light.

$\psi_1/\sqrt{2}$. We can't say this state without more information, but note: the two energies E_1 and E_2 , are no longer equal, but the simple fact C_1 and C_2 no longer have equal weight. This makes, of course, this one of the two possibilities—the figure is more likely than the other—but the electrons are much more except ψ_1 from the $\psi_1 + \psi_2$ and $\psi_1 - \psi_2$ blocks. So here, the center state has different amplitudes like ψ_1 and ψ_2 with possibly our loss of a higher energy. There is only one lowest state, and so we see the naive idea of "local charge" bands would be right.

10-5 Dyes

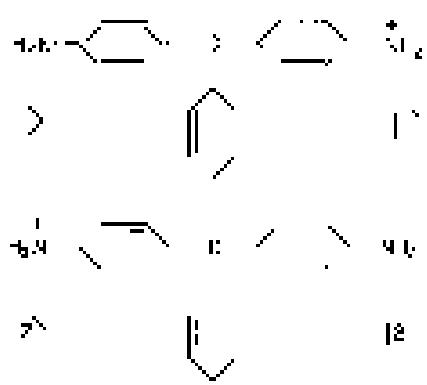


Fig. 10-2. Four resonance states for the molecule of the dye molecule.

We will give you one more chemical example of the resonance phenomenon this time on a larger molecular scale. It has to do with the dyes, dyes. Many dyes—let's see, most artificial dyes, I mean—have one characteristic: they have a kind of symmetry. Figure 10-2 shows an ion of a particular type called magenta, which has a purplish red color. The molecule has three ring structures, two of which are benzene rings. The third is not exactly the same as a benzene ring but not exactly nonbenzenoid, the ring. This figure shows two equilibrium stationary points, and we would guess that they should have equal energies E_1 . There is a certain amplitude that all the electrons can "tip" from one condition to the other, without the position of the "center" (position of the opposite end) with so many electrons involved, the flipping amplitude is considerably less than it is in the case of benzene, and the difference is in energy between the two is very, very small. Then, unfortunately, there are two stationary states ψ_1 and ψ_2 which are the sum and difference combinations of the two low states shown in the figure. The energy separation of ψ_1 and ψ_2 is enough to be equal to the energy of a photon in the visible region. There comes light on the molecule, and it is a very strong absorption at one frequency, and it appears to be highly colored. That's what it looks like.

An other interesting feature of such a dye molecule is that in the two case states shown, the center electric charge is located at different places. As a result, the molecule should be a rough testbed for a static electric field. We can get a similar effect in the ammonia molecule. To carry on the analysis by using exactly the same mathematics, provides work on the numbers E_1 and E_2 . Generally, these are obtained by putting experimental data. If we make measurements with many atoms, it is often possible to guess what will happen with some related dye molecule. Because of the large shift in the position of the center of electric charge, the value of μ in formula (10-20) is large, and this results in a high probability for absorbing light of the characteristic frequency ν_{ch} . Therefore, it is not only colored but very strongly so; it can absorb λ visible at about ϵ of light.

The rule of hopping—any, therefore, ψ is very sensitive to the complete structure of the molecule. By changing ψ , the energy splitting and with it the color of the dye can be changed. Also, the molecule does not have to be perfectly symmetrical. We know, even if the molecule has no mirror axis with slight modifications, even if there is some small asymmetry present, ψ_1 , ψ_2 can get some modification of the values by introducing slight asymmetries in the molecule. The example, an even important one, methylene green, is very green or blue-green, but has two of the cyano groups replaced by CH_3 . It's a different color because the ψ_1 is shifted and the flip flop ψ_2 is changed.

10-6 The Hamiltonian of a spin one-half particle in a magnetic field

Now we would like to discuss a two-state system involving an object of spin one-half. Some of you may well say, has been another electron example, but using it again may help to make some ideas perhaps a little clearer. We can think of it, electron, as just an $s = 1/2$ system. Although we will be talking in this chapter about "one-electron," when we find a $s = 1/2$ system for any spin one-half particle, suppose we choose for our base states ψ_1 and ψ_2 the states of ψ_1 and ψ_2 occupied at the electron spin $s = 1/2$ and $-1/2$.

In the previous section, we saw that the wave functions for two electrons in a magnetic field along the \hat{z} -axis were called $|+ \rangle$ and $|-\rangle$. In earlier chapters, we kept the notation of the chapter consistent, though, so we call this "up" and "down". From the time-symmetric state $|2\rangle$, the "up" and "down" refer to the angular momentum in the \hat{z} -direction.

Any possible value ϕ for the electron can be expressed as in Eq. (7.10), say giving the amplitude C_1 for the electron to be in state $|+\rangle$, and the amplitude C_2 that it is in state $|-\rangle$. To treat this problem, we will need to know the Hamiltonian for the two-state system that we have chosen to represent a magnetostatic. We begin with the contribution of a magnetic field in the \hat{x} -direction.

Remember that the \hat{x} -axis is the direction of the component B_x . From the definition of the two basis states (and the spins parallel and antiparallel to B_x) we know right away they are already stationary states with a definite energy in the \hat{x} -magnetic field. State $|+\rangle$ corresponds to an energy equal to $-e\hbar\omega$ and state $|-\rangle$ to $+e\hbar\omega$. The hamiltonian must be very simple in this case since C_1 or amplitude to have state $|+\rangle$ is not affected by C_2 , and vice versa.

$$\begin{aligned} h_1 \frac{dC_1}{dt} &= C_1 E_1 = -e\hbar\omega C_1, \\ h_2 \frac{dC_2}{dt} &= B_x C_2 = +e\hbar\omega C_2. \end{aligned} \quad (7.19)$$

In this special case, the Hamiltonian is

$$\begin{aligned} H_{1x} &= -e\hbar\omega, & H_{2x} &= 0, \\ H_{2x} &= 0, & H_{1x} &= +e\hbar\omega. \end{aligned} \quad (7.20)$$

So we know what the hamiltonian is due to magnetostatic in the \hat{x} -direction, and we know the energies of the stationary states.

Now suppose the field is not in the \hat{x} -direction. What is the hamiltonian? How are the basis elements changed if the field is not in the \hat{x} -direction? We are going to make an assumption here, the \hat{x} -is-a-kind-of-supposition principle for the form of the hamiltonian. More specifically, we want to assume that if two magnetic fields are superposed, the total is the hamiltonian simply added unless one field B_x has a plus B_x and the other has B_y . For a plus B_x then the total is H_x and H_y ; the basis is simply the same. This is sensible because if we consider only terms in the \hat{x} -direction, it's evident no other B_x are included. So let's assume that H is just $H_x + H_y$ the field B . That's all we need to be able to find the H_{ij} for any magnetic field.

Suppose we have a constant field B . We could take B to be along \hat{x} , its direction, and we would have found two stationary states with energy $-e\hbar\omega$ and $+e\hbar\omega$ respectively. Now if B is at a different orientation θ , it's not the \hat{x} -axis. Our assumption of the energy levels will be sufficient but, however, not quite the "full" form H ,

$$E_1 = -e\hbar^2\omega_x^2 + \eta_1^2 + \eta_2^2 \quad (7.21)$$

and

$$E_2 = +e\hbar^2\omega_x^2 + \eta_1^2 + \eta_2^2.$$

The rest of the proof is easy. We have here the formulas for the energies. We want a hamiltonian where η_1 , η_2 , and ω_x are what will give the new energies as calculated in our general formula of Eq. (7.10). The problem lies the θ -direction. First notice that the energy splitting is consistent with an average potential energy. Taking Eq. (7.21) we immediately find the expected

$$H_{12} = -H_{21}$$

which checks with what we already know about η_1 and η_2 are both zero.

[†]We are calling the resulting η_1 and η_2 "average" or energy of the magnetic moment instead of the $\sqrt{\eta_1^2 + \eta_2^2}$ to keep things simple, especially for the first part of the proof.

in that case $H_{11} = -\mu B_1$ and $H_{22} = \mu B_2$. Now if we equate the averages of $\langle H_{ij} \rangle$ (0.2.16) to what we argue from Eq. (10.10) we have

$$\left(\frac{n}{2} - \frac{H_{12}}{2}\right)^2 = H_{11}^2 = \epsilon^2(B_1^2 + B_2^2 + B_3^2) \quad (10.20)$$

(we have also made use of the fact that $H_{12} = H_{21}$, so that $H_{12}H_{21}$ can also be written as H_{12}^2). Again for the special case of a field in the x -direction, this gives

$$\epsilon^2(B_1^2 + B_2^2) = H_{11}^2 = \epsilon^2 B_1^2.$$

Thus B_1 , H_{11} must be zero in this special case, which means that H_{11} cannot have any terms in B_2 . (Remember we have said that all terms in H_{11} lie in B_1 , B_2 and B_3 .)

So far then, we have determined that H_{11} and H_{22} have to be $\pm B_1$, while H_{12} and H_{21} are not. We can make a simple guess as will satisfy Eq. (10.20) if we do this:

$$\begin{aligned} H_{11} &= \pm \mu B_1 \\ H_{22} &= \pm \mu B_1 \\ \text{and} \quad H_{12} &= \epsilon^2(B_1^2 + B_2^2) \end{aligned} \quad (10.21)$$

And it turns out (and that's not only my way to do it) that

"Well" you say "What's not linear in B ? So (10.21) gives $H_{12} = \mu(B_1^2 + B_2^2)$ " Not necessarily. There is another possibility which is linear in B ,

$$H_{12} = \mu(B_1 + B_2).$$

There are, in this, several other possibilities too, but given μ , we can always

$$H_{12} = \mu(B_1 + B_2),$$

where ± 1 is some arbitrary sign. Which sign should there should be? Well, as you can see you can change either sign and my phase you want, and the physics does not care, always be the same. So the choice is a matter of convention. I hope already we have chosen to use the minus sign and we take $\epsilon^2 = -1$. We might as well follow suit and write

$$H_{11} = -\mu B_1 + iB_2, \quad H_{22} = -\mu B_1 - iB_2.$$

Finally, all these quantities are related in a consistent manner, and of the arbitrary choices we made in Chapter 6,

to generate them based for an electron in an arbitrary magnetic field, is, then

$$\begin{aligned} H_{11} &= -\mu B_1, \quad H_{22} = -\mu B_1 + iB_2, \\ H_{12} &= -\mu(B_1 + iB_2), \quad H_{21} = -\mu B_2. \end{aligned} \quad (10.22)$$

And the equations for the amplitudes C_1 and C_2 are

$$\begin{aligned} \frac{dC_1}{dt} &= -iB_2 C_1 - (\mu - i\mu) C_2, \\ \frac{dC_2}{dt} &= -i\mu B_2 + iB_1 C_1 - \mu C_2. \end{aligned} \quad (10.23)$$

So we have discovered the "equations of motion for the spin state" of an electron in a magnetic field. We can test them by making some physical measurements, but the real test of any Hamiltonian is to see if specific gas problems in agreement with experiments. According to the tests that have been made, these equations are right. In fact although we made our elements only be constant, fitting the hydrogen we find evidence is enough for magnetic fields which vary with time. So we can now use Eq. (10.23) to look at all kinds of interesting problems.

10-7 The spinning electron in a magnetic field

Example: Let's say the electron has a constant field in the \hat{x} -direction. There are just the two stationary states with energies $E_{\pm}\mu_B$. Suppose we add a small field in the \hat{z} -direction. Then the equations look like our old two-state problem. We get the Landau levels again, more, and new energy levels are split a time for it is open. Now let's let the component of the field vary with time—say, as $B = B_0 \sin \omega t$. The equations are then the same as we had when we put an oscillating electric field in the \hat{x} -direction (discussed in chapter 9). You can work out the details in the same way. You will get the result that the oscillating field causes transitions from the $+$ state to the $-$ state—and vice versa. When the horizontal field oscillates with the natural frequency $\omega_0 = 2eB_0/\hbar$, this gives the maximum instantaneous energy of the magnetic resonance frequency ω_0 described in chapter 15 of Fermi's book.

It is also possible to make a magnet which does a spin precessing system. A Stern-Gerlach apparatus is used to produce a beam of particles following \hat{x} or \hat{y} , the \hat{z} -direction, which are to interact with an external magnetic field. The oscillating fields in the cavity can couple with the magnetic moment and induce oscillations which give energy to the cavity.

Now back to our preceding question. Suppose we have a magnetic field. At other points in the direction where polarization is \hat{z} and azimuthal angle is ϕ , as in Fig. 10-3. Suppose additionally that there is a current which has been produced with by spin oscillating along this field. What are the amplitudes C_1 and C_2 for such an electron? In other words, taking the sine of the electron ψ , we want to write

$$\psi = C_1 e^{iE_1 t} + C_2 e^{iE_2 t},$$

where E_1 and E_2 are

$$E_1 = E_0(\rho), \quad E_2 = \omega^2 |\psi|,$$

(here by $| \psi |$ and $| \psi |^2$ we mean the same thing we used to call $| + \rangle$ and $| - \rangle$ for real wave functions, see).

The answer to this question is obtained by generalizing to two-state systems. First we know that since the electron's spin is $\pm \hbar/2$ and it is in a stationary state with energy $E_0 = \pm \mu_B$, therefore both C_1 and C_2 must vary according to $e^{iE_0 t} (\propto \cos)$, and their coefficients a_1 and a_2 are given by (10.21), namely,

$$\frac{a_1}{a_2} = \frac{H_{12}}{E_0 - H_{11}}, \quad (10.24)$$

An added condition is that a_1 and a_2 should be normalized so that $|a_1|^2 + |a_2|^2 = 1$. We can take H_{12} and H_{11} from (10.22) using

$$B_x = B_0 \cos \theta, \quad B_y = B_0 \sin \theta \cos \phi, \quad B_z = B_0 \sin \theta \sin \phi.$$

So we have

$$\begin{aligned} H_{11} &= -\mu_B B_0 \cos \theta \\ H_{12} &= -\mu_B B_0 \sin \theta \cos \phi + i \mu_B B_0 \sin \theta \sin \phi. \end{aligned} \quad (10.25)$$

The last factor in the second equation is, incidentally, $i^2 = -1$ (a simpler writing).

$$H_{12} = -\mu_B \sin \theta \sin \phi. \quad (10.26)$$

Using these in the expression in Eq. (10.16) and canceling out the numbers after the decimal point, we get

$$\frac{a_1}{a_2} = \frac{\sin \theta \sin \phi}{-\cos \theta}. \quad (10.27)$$

With this ratio and the normalization condition, we can find both a_1 and a_2 . This is not hard, for we can make a selection with a little trick. Notice that

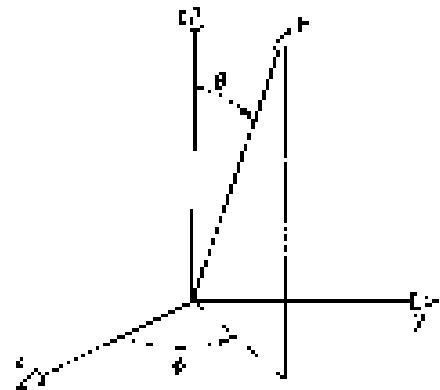


Fig. 10-3. The magnetic field \vec{B} in a rotating frame. The field \vec{B} is defined by the polar angles θ and ϕ in the coordinate system $x-y-z$.

$\hat{S}_z = \cos\theta = 2 \sin^2(\theta/2)$, and $\hat{C}_{12} = \hat{C}_1 \sin(\theta/2) \cos(\theta/2)$. Then Eq. (10.27) is equivalent to

$$\frac{\alpha_1}{\alpha_2} = \frac{\cos^2(\theta/2)}{\sin^2(\theta/2)}. \quad (10.29)$$

Several possibilities exist:

$$\alpha_1 = \cos \frac{\theta}{2} e^{i\phi}, \quad \alpha_2 = \sin \frac{\theta}{2}. \quad (10.29)$$

Now, Eq. (10.28) and this makes

$$|\beta_1|^2 = |\alpha_1|^2 = 1.$$

At this stage, multiplying by $e^{i\phi}$ is allowed, because arbitrary phase factors don't change anything. I myself personally prefer to make Eqs. (10.28) and (10.29) look like multiplying both by $e^{i\theta/2}$. So the form usually seen is

$$\alpha_1 = \cos \frac{\theta}{2} e^{-i\phi/2}, \quad \alpha_2 = \sin \frac{\theta}{2} e^{-i\phi/2}, \quad (10.30)$$

and this is the answer to our question. The spin vector and its components along the x -axis and y -axis are given up at this stage, because we know that its spin $\frac{1}{2}$ along the z -axis and α_1, α_2 amplitudes C_1 and C_2 are just α_1 and α_2 times $e^{i\phi/2}$.

Now we notice an interesting thing—the wrong half of the information has disappeared from us, see Eq. (10.29). This is clearly the same as multiplying that agrees to zero. The message that we have to search is general: the question of how to represent a particle whose spin is $\frac{1}{2}$ is an arbitrary task. The amplitudes of (10.30) are the projection amplitudes for spin $\frac{1}{2}$ particles propagating to the projective σ_z spin states we gave in Chapter 6 (Eqn. (6.32)). To spin $\frac{1}{2}$ particles, we can now give the amplitudes for several basis of spin $\frac{1}{2}$ particles in an elliptical state, particular Stern-Gerlach filter.

For $\sigma_z = +\frac{1}{2}$ say, we start with an input state $\psi_{in} = \alpha_1 |+\rangle + \alpha_2 |-\rangle$ and the spin does not. If $\sigma_z = +\frac{1}{2}$, represents state $+\frac{1}{2}$ spin up channel ψ_{out} , which makes the polar angles S and ϕ well defined, then in the middle part of the sum, we have

$$(+\epsilon) + \epsilon^* = \cos \frac{\theta}{2} e^{-i\phi/2} + (-\epsilon) |+\rangle = \sin \frac{\theta}{2} e^{-i\phi/2}. \quad (10.31)$$

This result is equivalent to what we began in the sum, Eq. (6.32), by partly ignoring the parameter ϕ . If you need to skip Chapter 6, you won't have lost anything important!

As in the Example 8.3 back again at the start of this section, a number of times. Suppose that we consider the following problem. We start with an electron whose spin is in some given direction, i.e., the σ_z in a magnetic field in the z -direction (Eq. 10.29 numbers), and then we want to know what is the σ_x direction. Again let's remember the rule by the Pauli combination $\psi_{out} = C_1 \psi_{in} - i C_2 \psi_{in}$ for this problem. However, the states of σ_z are unique and always have the form $|+\rangle$ and $|-\rangle$, so C_1 and C_2 only vary in phase. We know that

$$C_1(0) = C_1(0)e^{-i\phi/2} = C_1(0)e^{i\phi/2},$$

and

$$C_2(0) = C_2(0)e^{-i\phi/2} = C_2(0)e^{i\phi/2}.$$

Now initially we set the initial spin up σ_z was set in a given direction. This means that initially C_1 and C_2 are two numbers given by Eq. (10.29). After we wait a period of time T , the new C_1 and C_2 are the same two numbers multiplied respectively by $e^{i\phi/2T}$ and $-e^{i\phi/2T}$. What value is $\phi/2T$? It's $\pi/2$. That's why it's exactly the same as the angle θ we've struggled hard to find the sum of $C_1 \hat{S}_x T \psi_{in} + C_2 \hat{S}_y T \psi_{in}$, the angle of the last but one amplitude. This means that all the terms of the sum

cancel.

2. the same spin represents an electron having spin direction which differs from the original direction, say by a rotation about the \hat{z} -axis through the angle $\theta = \pi/2$. As we discussed in problem 10.1, we can also say the direction in the spin space is the angular vector $(\sin \theta, 0, \cos \theta)$ in the x - y - z axis. This result we discuss in several cases previously in the beginning and right now more. Now we have additional examples and accurate quantum mechanical description of the propagation of atomic magnetism.

It is interesting that the systems that we see here have no joint use. For the spinning electron in a magnetic field can be applied to any two-state system. This means that by making a quantum field analogy to the spinning electron, any particular two-state system can be solved by just putting $\psi = f(\theta, t)$ where $t = \tau/\omega_0$ so that the sum of energy so that $H_{\text{tot}} = H_{\text{ext}}$ is equal to zero or that $H_{\text{tot}} = -H_{\text{ext}}$. Thus any two-state problem is basically the same as the electron in a magnetic field. All you have to do is write $\psi = f(\theta)$, with $\partial/\partial\theta$ and $-\partial/\partial\theta = \partial\theta/\partial t$. Right? So similar what the physics is going to be—any number methods, or whatever you can think, it has a corresponding effect on problem. So it was not easy to make up problem to answer, we have solved all the state problems.

Now we have the general solution to the problem. Suppose you have come close to start with that this spin "is" in one direction and you get a constant field at last point in some other direction. You just need to understand that if the spin of ψ with the same angular velocity and equal to a constant times the vector B (initially $\psi = \psi_0 e^{i\omega_0 t}$) as you write with the θ , you have to make the use of the relation to keep ψ parallel with B , and keep changing the speed of rotation so that it is always perpendicular to the field and B , see Fig. 10-11. If you look along Ch. 20, will the ψ with certain final orientation of the spin axis, and the angle $\theta_1, \theta_2, \theta_3$ are just given by the angle between ψ and B ? Just to make this simple. You see it's another geometry problem to keep track of when you do the propagation resulting. Although it's easy to see what's involved, this geometry problem (of finding the norm of ψ in a relation with a varying angular velocity vector) is not easy to solve especially in the general case. Anyways, we can do it without, by given condition in any particular problem. In the next chapter we will talk more about the mathematical techniques for handling the important issue of a spin-orbit coupling and, therefore, for handling n -state systems in general.

21

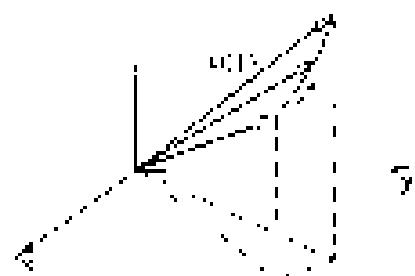


Fig. 10-11. The spin vector ψ in a magnetic field B as a function of the frequency ω_0 . The initial value is ψ_0 .

More Two-State Systems

II-1 The Pauli spin matrices

We continue our discussion of two-state systems. At the end of the last chapter we were talking about a spin and all particles in a magnetic field. We described the spin state by giving its amplitude C_1 , that the z-component of spin angular momentum is $|1/2\rangle$ and the anisotropy $\langle \hat{S}_z \rangle$ was $-h/2$. In earlier chapters we have called these basis states $|+\rangle$ and $|-\rangle$. We will now go back to that notation, although we may occasionally find it convenient to use $|+\rangle$ or $|-\rangle$, and $|+\rangle$ or $|-\rangle$ interchangeably.

We saw in the last chapter that when a spin interacts with a magnetic moment μ is in a magnetic field $\mathbf{B} = (B_x, B_y, B_z)$, the amplitudes $C_+(t)$ and $C_-(t)$ are connected by the following differential equations:

$$\begin{aligned} i\hbar \frac{dC_+}{dt} &= -i(B_x C_+ + (B_y - iB_z)C_-), \\ i\hbar \frac{dC_-}{dt} &= -i((B_x + iB_y)C_+ - B_z C_-). \end{aligned} \quad (II.1)$$

In other words, the Hamiltonian matrix H_{ij} is

$$\begin{aligned} H_{11} &= -iB_y & H_{12} &= i(B_x - iB_z) \\ H_{21} &= -i(B_x + iB_y) & H_{22} &= -iB_z. \end{aligned} \quad (II.2)$$

And $(II.1)$ and $(II.2)$ are of course the same as

$$i\hbar \frac{dC_i}{dt} = \sum_j H_{ij} C_j, \quad (II.3)$$

where i had ± 1 take on the values $i = +1$ and $-i = -1$ (Fig. 1 and 2).

The two-state system of the electron spin is so important that it is very useful to know exactly how we are writing things. We will now make a little mathematical digression to show you how people usually write the equations of a two-state system. It is done this way: first, note that each term of the Hamiltonian is proportional to μ and to some component of \mathbf{B} ; we can then "cancel μ " and write that:

$$H_{ij} = -i(\epsilon_{ij} B_x + \alpha_i B_y + \beta_{ij} B_z) \quad (II.4)$$

where ϵ_{ij} is now physical (here, ϵ is obtained from means m), the coefficient is α_i , β_{ij} , and ϵ_{ij} . There are $N \times N = 12$ of them, can be figured out so $(II.4)$ is identical with $(II.2)$.

Let's see what they have to do. We start with $\alpha_i = 0$; B_x appears only in H_{11} and H_{22} , everything will be R.K. If

$$\begin{aligned} \epsilon_{11} &= 1, & \epsilon_{22} &= 0, \\ \alpha_{11} &= 0, & \alpha_{22} &= -1. \end{aligned}$$

We often write the matrix H_{ij} as a table like this:

$$H_{ij} = \begin{pmatrix} \epsilon_{11} & H_{12} \\ H_{21} & \epsilon_{22} \end{pmatrix}.$$

II-1 The Pauli spin matrices

II-2 The spin matrices and operators

II-3 The solution of the two-state equations

II-4 The polarization vector of the photon

II-5 The natural Raman effect

II-6 Generalization to N-state systems

Review Chapter 23, Vol. I, Physics 102

† The second article, as printed on the 2nd reading of the book. It is close to my original notes; however, I have not checked it.

For the Hamiltonian we can substitute back in the magnetic field \mathbf{B} . This is interesting:

$$H_0 = \frac{1}{2} \begin{pmatrix} -\omega_B & \omega(B_x - iB_y) \\ \omega(B_x + iB_y) & +\omega_B \end{pmatrix}$$

In the same way we can write the coefficients σ_i^x as the matrix

$$\sigma_i^x = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (1.3)$$

Multiplying with the coefficients of H_0 , we get that the terms of σ_i have to be

$$\sigma_{11}^x = 0, \quad \sigma_{12}^x = 0,$$

$$\sigma_{21}^x = 0, \quad \sigma_{22}^x = 0.$$

Or, in shorthand,

$$\sigma_i^x = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \quad (1.4)$$

Finally, looking at B_y , we get

$$\tau^2 - \omega_B - \sigma_{12}^y = 0,$$

$$\sigma_{12}^y = 0, \quad \sigma_{21}^y = 0;$$

or

$$\sigma_i^y = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \quad (1.5)$$

With these three signs missing, Eqs. (1.12) and (1.13) are identical. To have room for the subscripts $(+)$ and $(-)$, we have shown which signs go with which component of \mathbf{B} in \mathbf{B}_1 , \mathbf{B}_2 , \mathbf{B}_3 , and \mathbf{B}_4 respectively. Usually, however, one finds the signs omitted.

It's easy to verify that they are there. And the σ_i 's can be written as (Table 1.1). Table 1.1 (1.1) is written:

$$\sigma_i^z = -\sigma_1 \sigma_2 \sigma_3 + \sigma_2 \sigma_3 \sigma_1 + \sigma_3 \sigma_1 \sigma_2. \quad (1.6)$$

Because the sigma matrices are so important, they are used all the time by the professionals. We have put them together in Table 1.1. (Anyone who is going to work in quantum mechanics has to memorize them. They are also in the book just mentioned over the physicist who invented them.)

In the text we have included an "easy" system matrix which is useful if we want to calculate the energy of a system with two spin degrees of freedom, or if we want to choose a different zero energy. For such situations we take $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ to be the first equation (1.11) and \mathbf{E}_4, \mathbf{C} to be the second equation. We can include this in the new table (in) if we define the *spin matrix* " Γ " as follows:

$$\Gamma = \sigma_i^z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (1.7)$$

and reverse Eq. (1.8) to

$$H = \mathbf{E}_1 + \mu_B \mathbf{B}_1 - \mu_B \mathbf{B}_2 + \mu_B \mathbf{B}_3. \quad (1.8)$$

Usually, the unknown Γ is not given but like E_0 is automatically understood implicitly. In which case, the equations simply

$$H = \mathbf{E}_1 + \mu_B \mathbf{B}_1 - \mu_B \mathbf{B}_2 + \mu_B \mathbf{B}_3. \quad (1.11)$$

One reason the spin matrices are useful is that expressed by the matrix Γ we can be interested in terms of them. A vector \mathbf{v} and its conjugate \mathbf{v}' has four numbers in it, say,

$$\mathbf{v} = \begin{pmatrix} v_1 & v_2 \\ v_3 & v_4 \end{pmatrix}.$$

M can always be written as a linear combination of Pauli matrices. For example,

$$M = a \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + b \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + c \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + d \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

There are many ways of doing it, but one special way is to say that M is a certain amount of σ_x , plus a certain amount of σ_y , and so on. That is:

$$M = aI + b\sigma_x + c\sigma_y + d\sigma_z$$

where the "coefficients" a , b , c , and d may be positive or complex numbers.

Since any two-by-two matrix can be represented in terms of the Pauli matrices and the identity matrix, we have all that we need for any two-state system. We know what the two-state system—its wavefunction evolution, the measured energy, the Hamiltonian, everything can be written in terms of the sigma's. Although the sigma's seem to have a particular significance in the physical problem of magnetism in a magnetic field, they can also be thought of as just useful matrices, which can be used for any two-state problem.

For instance, in the case of blocking Larmor motion, and a photon can be thought of as interacting with either of two states. We say the photon (photon or electron) is a two-state system. In this case, two states with respect to the charge. When blocked in the $|+\rangle$ state, the $|-\rangle$ state can represent the photon and the $|+\rangle$ state can represent the hydrogen. People say that you can obtain two "hydrogen-like" states.

Since we will be using the properties of the "algebra" of the common mathematics of two-state systems, let's review quickly the common rules of matrix algebra. By the "sum" of any two square matrices we mean what was obvious in Eq. (II.2). In general, if we "add" two matrices A and B , the "sum" C means that each term C_{ij} is given by

$$C_{ij} = A_{ij} + B_{ij}$$

Each term of C is the sum of the terms in the same cells of A and B .

In Section 4-5 we have already mentioned that the rule of a sum is "zero sum." This concept will be useful calculating with the sigma matrices. In general, the "product" of two matrices A and B (in that order) is defined to be a matrix C whose elements are

$$C_{ij} = \sum_k A_{ik}B_{kj} \quad (\text{II.12})$$

The pattern of products of terms taken in pairs from the i th row of A and the j th column of B . If row indices are written out in tabular form as in Fig. II-1, there is a good "system" for getting the terms of the product matrix. Suppose you are calculating C_{12} . You run your left index finger along the second row of A and your right index finger down the first column of B , multiplying out pair and pairing as you go. The bars used to indicate zero to do it in two digits.

$\begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}$	$\begin{pmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ B_{21} & B_{22} & B_{23} & B_{24} \\ B_{31} & B_{32} & B_{33} & B_{34} \\ B_{41} & B_{42} & B_{43} & B_{44} \end{pmatrix}$	$\begin{pmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \\ C_{31} & C_{32} & C_{33} & C_{34} \\ C_{41} & C_{42} & C_{43} & C_{44} \end{pmatrix}$
$\overbrace{\hspace{10em}}$	$\overbrace{\hspace{10em}}$	$\overbrace{\hspace{10em}}$

$$C_{12} = \sum_k A_{1k}B_{k2}$$

$$= A_{11}B_{12} + A_{12}B_{22} + A_{13}B_{32} + A_{14}B_{42}$$

Fig. II-1. Multiplying two matrices.

It's a bit more complicated for two-system matrices. For instance, if we multiply σ_x times σ_z , we get

$$\sigma_x^2 = \sigma_x \cdot \sigma_z = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which is just the unit matrix \mathbf{I} . Or, for another example, it's worth noting

$$\sigma_x \sigma_y = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

Looking at Table 11-1, you see that the product is just i times the unit matrix—just as multiplying two numbers each term of the matrix. Since the products of the sigma taken over and over are important, it's well to remember. We considered them all in Table 11-1, but now we're interested in how they fit in a theory.

There's another very important and interesting point about these σ matrices. We can imagine, if we wish, that the three matrices σ_x , σ_y , and σ_z are analogous to the three components of a vector—it's sometimes called the "magnetic vector," or "angular momentum vector," or "vector-matrix." It is three different numbers, and hence associated with position, or, x , y , and z . With it, we can write the Hamiltonian of the system in a nice form which works in any coordinate system:

$$H = -\mu \mathbf{B} \cdot \mathbf{R}. \quad (11.12)$$

Although we'll see why this has three entries in the representation in which "up" and "down" states have different spin numbers, and a particle's spin σ —we could agree on which the matrix would look like in some other representation. Although it takes a lot of algebra, you can show that they change, among themselves like the components of a vector. (We won't, however, carry around proving it right now. You can check it if you want.) You can use it in different coordinate systems, as though it is a vector.

You remember that the H is related to energy in quantum mechanics. It is, at best, just equal to the energy in the simple situation where there is only one state. Even for two-state systems of the electron spin, where we write the Hamiltonian as in Eq. (11.12), it's like saying that the classical term is for the energy of a charged magnet with magnetic moment μ in a magnetic field B . Classically, we would get

$$E = -\mu \mathbf{B} \cdot \mathbf{R}. \quad (11.13)$$

where \mathbf{R} is the position of the magnet and \mathbf{B} is an external field. We can imagine that Eq. (11.13) can be converted to (11.12) if we replace the classical energy by the Hamiltonian and the classical \mathbf{B} by the \mathbf{B} from Eq. (11.1). Then, after this simple formal substitution, we'll be just the result as a magnetic dipole. It is sometimes said that magnetism in classical physics does not respond to magnetic quantum mechanics. It is really more correct to say that the Hamiltonian which corresponds to the energy and the quantity that can be related to energy has a corresponding matrix.

For example, the magnetic moment can be defined classically by saying that the energy in an external field B is $-\mu \cdot \mathbf{B}$. This defines the magnetic moment μ . Then we look at the formula for the momentum of a test particle in an object in a magnet field and try to identify whatever the charges are and correspond to the various components in the classical formula. That's the trick by which quantum力学和磁矩的对应关系可以被建立起来。

You may say, if you want to understand what a classical vector is, try to do it, and maybe you will discover something that won't break your head. But that's not the idea. They are two *separate* quantum mechanics, a different kind of a theory in different parts of the world. It just happens that there are certain correspondences which are really more than mere coincidences, which things to remember with. That is, you remember Eq. (11.12) because you learned classical physics, 11-4

then if you remember the correspondence principle you have a headache, or something like Eq. (11.13). Of course you're using the quantum mechanics, and the classical mechanics is only an approximation. So there is no mystery in the fact that in classical mechanics there is an equation of motion, or an "acceleration law," which are truly the same fundamental. To reconstruct the original object from the shadow is not possible in any general way, but the shadow does help you to remember what the object looks like. Equation (11.13) is the truth, and Eq. (11.10) is just a guess. Because we learn classical mechanics first, we would like to be able to get the quantum formula from it, but there is no automatic scheme for doing that. We must always go back to the real work and discover the correct quantum mechanical equations. When they come out looking like something in classical physics, we are in luck.

In the shadow concept each hope turns and appears to you to be elaborating the actual trouble down the relation of classical physics to quantum physics. Please notice the circumlocuted sentence of a professor who has really taught quantum mechanics to students who he didn't know them to represent, and they were in graduate school! That boy always seemed to be hoping that somehow symmetry in different ways could be seen to follow as a logical consequence of classical mechanics which they had learned three nights years before. (Perhaps they wanted to make it easier to learn something new.) Then he mentioned the classical formula, Eq. (11.14), in your old notes again, and then with something that it was far too quaint to expect you would be so unwilling to take new symmetric formula, Eq. (11.13), as the basic one.

11-2 The spin matrices no operators

While we are on the subject of undetermined numbers, you would like to be able to still consider ways of writing things—ways which is used very often because it is so simple. It comes directly from the "scatter" introduced in Chapter 8. If we have a system, it is a state $|S\rangle$ which varies with time, we can—as we did in Eq. (8.7)—write the amplitude that the system would be in the state $|S'\rangle$ at $t = \infty$ as

$$\langle S'|S\rangle = \delta(S, S') = \sum_i \langle S|S_i\rangle \delta(S_i, S').$$

The matrix elements $\langle S|S_i\rangle = \delta(S_i, S)$ is the amplitude that the state $|S\rangle$ will be converted into the state $|S_i\rangle$ in the time interval of Δt . We can define this by writing

$$S_i(t) = \delta(S_i, S) = C_i - \int_{-\infty}^t R_i(t') dt'$$

and we expect that the amplitudes $C_i(t) = \langle S|S_i(t)\rangle$ are related by the differential equations

$$\partial_t \frac{dC_i}{dt} = \sum_j R_{ij} C_j. \quad (11.15)$$

If we substitute the amplitudes C_i explicitly, the equation is then appears as

$$\partial_t \frac{d}{dt} \langle S|S\rangle = \sum_i \partial_t \langle S|S_i\rangle \delta(S_i, S). \quad (11.16)$$

Now the matrix elements R_{ij} are also amplitudes which we can write as $\langle S|R_j|S\rangle$ (as differential operators having time rates).

$$\partial_t \frac{d}{dt} \langle S|S\rangle = \sum_i \langle S|R_i|S_i\rangle \delta(S_i, S). \quad (11.17)$$

Please note that $-\partial_t \langle S|R_j|S\rangle$ is the amplitude that under the physical transformation described by R_j a state $|S\rangle$ will change the name of, "generate" the state $|S_j\rangle$. (All of this is implied in the discussion of Section 8.4.)

Now following analysis of Section 9.2, we can drop out the common term $\psi \sim \text{Eq. (11.7)}$ since it is Euler angle θ —and write the equation simply as

$$\Psi \frac{d}{dt} \psi = \sum_i H_i(j) |j\rangle \langle j| \psi. \quad (11.18)$$

Or, using one step further, we can do so much: let us also write

$$A \frac{d}{dt} \psi = H | \psi \rangle. \quad (11.19)$$

In Chapter 9 we pointed out that when things are working the way they *H* in Eq. (11.19) or $H | \psi \rangle$ is called an operator. From now on we will put the line of text “*H* is an operator” to remind you that *H* is an operator and not just a constant. Then we write $\partial_t \psi$. All eight of the two equations (11.18) and (11.19) were originally given using $\psi \sim \text{Eq. (11.7)}$. Eq. (11.19), we can think about it in a different way. For instance, we would observe Eq. (11.18); in this way: “The time derivative of the state vector $|\psi\rangle$ is equal to what you get by operating with the Hamiltonian operator *H* on each basis state, multiplying by the amplitude $\langle j| \psi\rangle$; the $|j\rangle$ is in the state j , and summing over all $|j\rangle$.” Eq. (11.19) describes this way: “The time derivative (change) of a state $|\psi\rangle$ is equal to what you get if you operate with the Hamiltonian *H* on the basis vector $|\psi\rangle$.” Let just a short-hand way of saying what is in Eq. (11.19); now as you will see, it can be a great convenience:

If we wish, we can ignore the “derivation” idea, one more step. Equation (11.19) is true for any wave $|\psi\rangle$. Also on left-hand side $\partial_t \psi$, is also an operator —it’s the equation “differentiable by t and multiply by $i\hbar$ ”. Therefore, Eq. (11.19) can also be thought of as a connection between operators—an operator equation

$$i\hbar \frac{d}{dt} = \hat{H}.$$

The Hamiltonian operator (with $i\hbar$ constant) produces the same results as does when acting on any $|\psi\rangle$. Remember that this equation $i\hbar \frac{d}{dt} = \hat{H}$ (11.19) is *not* a statement that the \hat{H} operator is just the identical operator as $i\hbar \frac{d}{dt}$. It is just $i\hbar \frac{d}{dt}$ the fundamental law of nature—the law of motion—in a quantum system.

Just to get some practice with these ideas, we will show you another way we could get to Eq. (11.18). You know that we can write any state $|\psi\rangle$ as its components into states $|k\rangle$ and $|l\rangle$ (see Eq. (3.5)),

$$|\psi\rangle = \sum_k c_k |k\rangle |\psi_k\rangle. \quad (11.20)$$

so does $|\psi\rangle$ change with time? Well, just take its derivative:

$$\frac{d}{dt} |\psi\rangle = \frac{i\hbar}{\hbar} \sum_k c_k \frac{d}{dt} |k\rangle |\psi_k\rangle. \quad (11.21)$$

Now, before states $|k\rangle$ do not change with time because we are talking about as definite fixed k and the amplitudes c_k are numbers which may vary. So Eq. (11.21) becomes

$$\frac{d}{dt} |\psi\rangle = \sum_k c_k \frac{i\hbar}{\hbar} \frac{d}{dt} |k\rangle |\psi_k\rangle. \quad (11.22)$$

Since we know $i\hbar \frac{d}{dt} |k\rangle$ from Eq. (11.16) we get

$$\begin{aligned} \frac{d}{dt} |\psi\rangle &= \frac{i\hbar}{\hbar} \sum_k c_k \sum_j H_{kj}(j) |j\rangle \psi_j \\ &= \frac{i\hbar}{\hbar} \sum_j c_j |j\rangle H_{jj}(\psi_j) = \frac{i\hbar}{\hbar} \sum_j c_j H_{jj}(\psi_j) |j\rangle \psi_j. \end{aligned}$$

This is Eq. (11.18) all over again.

So we have many ways of writing or the hamiltonian. We can think of the σ 's as "spins" (they're really a bunch of numbers), we can think of the "impurities" of $H[\psi]$, or we can think of the "matrix" H , or we can think of the operator \hat{H} . It all means the same thing.

Now let's go back to our two state system. If we write the Hamiltonian in terms of the sigma operators (and suitable normalized coefficients like α_i etc.), we can directly solve (1.19) for an amplitude $\langle \psi_1 | \psi \rangle$ etc. (as shown in the signature). If you use the operator idea, we can write the equation of motion of wave $|\psi\rangle$ in a magnetic field as

$$\partial_t \frac{\delta}{\delta t} |\psi\rangle = -i(\vec{\sigma}_x \partial_x + \vec{\sigma}_y \partial_y + \vec{\sigma}_z \partial_z) |\psi\rangle \quad (1.20)$$

When we want to "solve" such an equation, we will normally have to expand $|\psi\rangle$ in terms of basis vectors (just as we have to find the expansion of space vector when we want operate (1.19)). So we will usually want to put Eq. (1.20) in the so-called expanded form:

$$\partial_t \frac{\delta}{\delta t} |\psi\rangle = -\mu \sum_i (\vec{\sigma}_x i_x + \vec{\sigma}_y i_y + \vec{\sigma}_z i_z) |\psi\rangle \quad (1.21)$$

Now you will see why one operation is so useful. To use Eq. (1.21) we need to know which components of the σ -operators work on each of the basis states. Let's do that. Suppose we have $|\pm\rangle$ it is a spin vector. So, but what? Well, let's multiply it on the left by $\langle \pm|$, we have

$$\langle \pm | \vec{\sigma}_x | \pm \rangle = \sigma_{x,\pm} = 1$$

Using Table 1.1-1, we see from $\sigma_{x,\pm}$:

$$\langle + | \vec{\sigma}_x | + \rangle = 1 \quad (1.22)$$

Now let's multiply $\langle \pm | \pm \rangle$ on the left by $\langle \mp |$. We get

$$\begin{aligned} \langle \pm | \vec{\sigma}_x | \pm \rangle &= \sigma_{x,\pm} = 1 \\ \langle - | \vec{\sigma}_x | - \rangle &= 0 \end{aligned} \quad (1.23)$$

Here is why one basis vector that satisfies both (1.22) and (1.23), it is $|\pm\rangle$. We know then that

$$\vec{\sigma}_x | \pm \rangle = \pm \langle \pm | \vec{\sigma}_x | \pm \rangle \quad (1.24)$$

By this kind of argument, you can easily prove that all the properties of the sigma operators can be described in the *signature notation* by the set of rules given in Table 1.1-1.

If we have products of $\vec{\sigma}_x$, and if we they appear in a products of $\vec{\sigma}_x$, then when we expand it using the $\sigma_{x,\pm}$ notation, you carry out first the operations with the operators which is further to the right. For instance, by $\vec{\sigma}_x \vec{\sigma}_y \rightarrow \cdot$ we are to understand $\vec{\sigma}_x \vec{\sigma}_y = 0$. From Table 1.1-1, we get $\sigma_{x,\pm} \sigma_{y,\pm} = \pm \delta_{xy}$, so

$$\vec{\sigma}_x \vec{\sigma}_y = \langle \pm | \vec{\sigma}_x | \pm \rangle \langle \pm | \vec{\sigma}_y | \pm \rangle \quad (1.25)$$

Now my number (like i , just i , etc.) might be greater or smaller than one. In this material, an Eq. (1.25) is the general

$$\langle \pm | \vec{\sigma}_x | \pm \rangle \langle \pm | \vec{\sigma}_y | \pm \rangle = \langle \pm | \vec{\sigma}_y | \pm \rangle \langle \pm | \vec{\sigma}_x | \pm \rangle$$

If you do the same thing for $\vec{\sigma}_x \vec{\sigma}_y^{-1} \rightarrow \cdot$, you will find that

$$\langle \pm | \vec{\sigma}_x | \pm \rangle \langle \pm | \vec{\sigma}_y^{-1} | \pm \rangle = \langle \pm | \vec{\sigma}_y^{-1} | \pm \rangle \langle \pm | \vec{\sigma}_x | \pm \rangle$$

Looking at Table 1.1-1, you see that $\vec{\sigma}_x \vec{\sigma}_y$ carried by $\langle \pm | \vec{\sigma}_y | \pm \rangle$ gives us what you get if you swap $\vec{\sigma}_x$ with $\vec{\sigma}_y$ and multiply by $-\delta_{xy}$. We can, therefore, say,

Table 1.1-2
Properties of the *signature*

σ_x	$\langle \pm \vec{\sigma}_x \pm \rangle = 1$
σ_y	$\langle \pm \vec{\sigma}_y \pm \rangle = 0$
σ_z	$\langle \pm \vec{\sigma}_z \pm \rangle = \pm 1$
σ_x^{-1}	$\langle \pm \vec{\sigma}_x^{-1} \pm \rangle = -1$
σ_y^{-1}	$\langle \pm \vec{\sigma}_y^{-1} \pm \rangle = 0$
σ_z^{-1}	$\langle \pm \vec{\sigma}_z^{-1} \pm \rangle = 1$

Let's suppose $\hat{A}_0 = \hat{B}_0$, it's identical with the operator \hat{M}_z , now write this statement as an operator equation:

$$\hat{A}\hat{C}_0 = \hat{B}_0. \quad (11.29)$$

Notice that this equation is identical with one of the matrix equations of Table 11.2. So again we see the correspondence between the matrix and Schrödinger picture of view. Furthermore the equations of Table 11.2 can, therefore, also be considered as equations about the same operators. You can check and may the initial follow from Table 11.2. It is true when working with these things, how to keep track of whether a quantity like \hat{A} is an operator or a number. All the quantities are the same either way as Table 11.2 is the right operators, or the right numbers as you wish.

11-3 The solution of the two-state equations

We can now write any two-state equation in various forms. For example, since we

$$\begin{aligned} i\hbar \frac{d\hat{C}_0}{dt} &= \sum_j H_{0j} \hat{C}_j \\ \text{or} \quad i\hbar \frac{d\hat{C}_0}{dt} &= H_0 \hat{C}_0. \end{aligned} \quad (11.30)$$

The last term means *no coupling*. For a spin one-half particle in a magnetic field, the Hamiltonian H is given by Eq. (11.6) or by Eq. (11.1).

If the field is in the \hat{z} -direction, then, as we have seen several times by now—the solution is $C_0 = C_0(\cos \omega t, \sin \omega t)$, whatever it is, processes around the \hat{z} -axis (just as if you were to turn the physical object we rotate it). In the moment the \hat{z} -axis is at an angle θ relative to twice the magnetic field times $\mu\hbar$. The same is true as before for a magnetic field in any other direction because the process is independent of the coordinate system. Now take a situation where the magnetic field varies from B_0 to B_1 in a smooth, elicited way, then we can analyze our situation in the following way. Suppose you start with the spin in the \hat{z} -direction and you have an x -magnetic field. The spin starts to turn. Then if the field is turned off, the spin stops turning. Now if the field is turned on again processes about \hat{x} , and so on. Now suppose we let the fields vary in time, you can figure out what the final state is—along which axis it will point. Then you can refer this state back to the original $| + \rangle$ and $| - \rangle$ with respect to \hat{x} by using the projection formulas we saw in Chapter 10 (in Chapt. 6). If the state ends up with its spin in the direction (B_x, B_y) , it will have an up-amplitude $\cos(\theta/2)e^{-i\phi/2}$ and a down-amplitude $\sin(\theta/2)e^{+i\phi/2}$. That's pretty good. Let's work this out of the solution of the differential equations.

The solution just described is sufficiently general to take care of any reasonable system. Let's take an example of two numbers making up—including the effects of an electric field. If we describe the system in terms of the states $| + \rangle$ and $| - \rangle$ the equations look like this:

$$i\hbar \frac{d\hat{C}_0}{dt} = -\Gamma_1 \hat{C}_1 - \omega \hat{C}_{11}, \quad (11.31)$$

$$i\hbar \frac{d\hat{C}_{11}}{dt} = -\Delta \hat{C}_{11} - \gamma \hat{C}_{11}.$$

You say, "But, I remember there was some A_0 in there." Well, we have added the ω in to energy to make the \hat{E}_0 , etc. etc. (You can always do that by changing the \hat{A} 's amplitudes by $e^{i\omega t}$ some time t , $e^{-i\omega t}$ later, and get rid of any constant energy.) Now if our expanding equations always have the same solutions, then we really can't care to divide twice... If we look at these equations and look at Eq. (11.1) then we can make the following identification. Let's call $| + \rangle$ the state $| + \rangle$ and $| - \rangle$ the state $| - \rangle$. What does we mean this? we are thinking of a wavefunction in space so that $| + \rangle$ and $| - \rangle$ has anything to do with the space. It is purely artificial.

We have an *arbitrary* spin; then we might "call the antinodal electron representation in space," or something—a three-dimensional "Vigint"—in which being " $\hat{\sigma}_z$ " corresponds to "having the molecule in the state $|z\rangle$ " and being " $\hat{\sigma}_{\pm}$ " doing just only basis movements having a coordinate in the states $|+y\rangle$, $|+x\rangle$. Then, the equations will be identical as follows. First of all, you see that the Hamiltonian can be written in terms of the sigma matrices as

$$H = \frac{1}{2}m + \mu\hat{\sigma}_z \quad (1.13)$$

(m , perhaps, is often zero, $\mu\hat{\sigma}_z$ in Eq. (1.13) corresponds to $-i$ in Eq. (1.22), and $\mu\hat{\sigma}_z$ corresponds to $-i\omega$). In our "model" above, then, we have a current i flowing along the z-direction. If we have an electric field E which is changing with time, then we have a $\partial E/\partial t$ field along the z-direction which varies in proportion. So the current of an electron is a constant field with a constant component in the z-direction and at oscillating frequency ω (or resistance to oscillating frequency ω) due to the derivative of the current with respect to the oscillating electric field. Unfortunately, we do not have the time to go any further into the details of this correspondence, or a work-out of the technical details. We only wished to make the four basic assumptions of Fermi's theory become analogous to a spin one-half object precessing in a magnetic field.

11-4 The polarization states of the photon

There are a number of other interesting systems which are interesting to study, and the first one we would like to look at is the photon. To describe a photon we must first give its vector momentum. For a monochromatic frequency, it is represented by the momentum, so we don't have to say also what the frequency is. After all, though, we can have a property called the polarization. Imagine then that there is a photon coming at you with a definite momentum in two axes (which we'll take the same thing out of this discussion to start with) in a linear wave of momentum space. Then there are two directions of polarization. In the classical theory, light can be described as having an electric field which oscillates periodically in an electric field with amplitude vertically (for instance); if the two kinds of light are called rectangular and unpolarized light, the light can also be polarized to some other direction, which we call up from the unpolarization as a final—the z-direction and down in the y-direction. Or if you take the x and the y components out of given by 90° you get an electric field that rotates, the light is elliptically polarized. (There is just a quick reminder of the coherent theory of polarized light that we studied in Chapter 11, Vol. I.)

Now, however, suppose we have a single photon—just one. There is no electric field that we can discuss in the sym way—all we have is one photon. But a photon has a very important of the classical parameters of polarization. There must be at least two different kinds of photons. At first you might think that there are infinitely many, after all, the electric waves can point in all sorts of directions. We can, however, describe the polarization of a photon as a two-state system. A photon can have the same state as the state $|x\rangle$ or $|y\rangle$; we call the polarization state of each one of these photons to be an $\alpha\beta$ state which obviously is a polarized light. On the other hand, by $\beta\beta$ we mean the polarization state of each of the photons in a unpolarized beam. And we can take $|x\rangle$ and $|y\rangle$ as our basis states of a photon of given energy, in passing along you in p ; we will call this polarization. So there are two base states, $|x\rangle$ and $|y\rangle$ and they are all that are needed to cover nearly photons at all.

For example, if we have a piece of polarized set with its axis to pass light parallel in x , at $\alpha\beta$ and the radiation and we send in a photon which we know is in basis $\beta|1\rangle$, it will be absorbed by the photon. If we send in a photon of which we know is in the state $|x\rangle$, it will come right through as $|x\rangle$. If we take a glass of water which takes a beam of polarized light and splits it into an $|x\rangle$ beam and an y beam, that piece of water is the complete top of a Stern-Gerlach apparatus which splits a beam of electrons into the two states $|x\rangle$ and $|y\rangle$. So every-

ding we did before with particles and Stern-Gerlach apparatuses, we can do again with light and pieces of calcite. And when light goes through a piece of polarized it "samples" both axes to decide how much energy has gone along each axis or the axis of the polaroid π' to distinguish it from the axes of our base states. See Fig. 11-2. A photon that comes out will be in the state $|x\rangle$. However, any state can be represented as a linear combination of these states, so it is enough for the combination $|y\rangle$, here.

$$|x\rangle = \cos \theta |x\rangle + \sin \theta |y\rangle \quad (11.31)$$

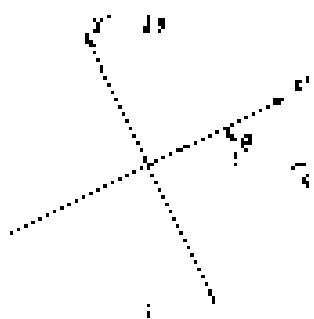


Fig. 11-2. Two polaroids, one light coming to the measurement vector of the other.

That is, if a photon comes through a piece of polaroid set at the angle θ (with respect to $|x\rangle$), it can still be resolved into $|x\rangle$ and $|y\rangle$ beams. In a case of calcite, for example. Or, you can, if you wish, just analyze it in $|x\rangle$ and $|y\rangle$ components in your imagination. Either way, you will find the amplitude due to $|x\rangle$ to be in the $|x\rangle$ state and the amplitude due to $|y\rangle$ to be in the $|y\rangle$ state.

Now we ask this question: Suppose a photon is polarized in the x -direction by a piece of calcite set at the angle θ and before the polaroid of the angle $\theta/2$ —as in Fig. 11-3, what will happen? With what probability will it get through? The answer is the following. After it goes through the first polaroid, it is definitely in the state $|x\rangle$. The second polaroid will let the photon carry on if it is in the state $|x\rangle$; that is, if it is the state $|x\rangle$, because $|x\rangle$ has total probability that it evolution appears to be in the state $|x\rangle$. We ask that equality from the absolute equation of amplitude $(x|x)$ that is present in the state $|x\rangle$ is also in the state $|x\rangle$. Whereas in $|y\rangle$? Inserting Eq. (11.31) by $|x\rangle$ to get

$$(x|x') = \cos^2 \theta (x|x) + \sin^2 (x|y).$$

Because $(y|y) = 0$, from the physics—it may mean if $|y\rangle$ and $|x\rangle$ are base states and $(x|x) = 1$. So we get

$$(x|x') = \cos^2 \theta,$$

and the probability is $\cos^2 \theta$. So, for example, if the first polaroid is set at 30° , a photon will get through $1/4$ of the time, and $1/4$ of the time it will fail the selection by being absorbed again.

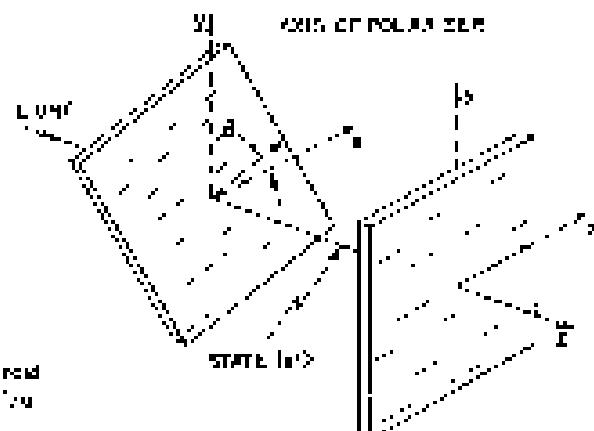


Fig. 11-3. Two stages of selection with angle $\theta/2$ between pieces of calcite.

Now let us see what happens classically in the same situation. We work now a beam of light with a double "tail" which is passing in some wave in another "sec transposed". After it goes through the first polaroid, the electric field is oscillating in the x -direction with a size E_0 ; we would draw the field as an oscillating vector with a peak value E_0 , in a diagonal like Fig. 11-4. Now when the light arrives at the second polaroid, only the component $E_0 \cos \theta$ of the field is left, gets through. The intensity is proportional to the square of the field and therefore $I = E_0^2 \cos^2 \theta$. So the energy passing through is $\cos^2 \theta$ smaller than the energy which was entering the first polaroid.

The classical picture and the quantum picture give similar results. If you were to throw 10 billion photons, the survival probability and the average probability of each one going through is, say, 5/8, you would expect 3/8 of 10 billion would pass through. Likewise, the energy that they would carry would be 5/8 of the energy that you are expected to pass through. The classical theory says nothing about the statistics of the photons; it simply says that the energy carried through will be precisely 5/8 of the energy which you were sending in. That is, it's more improbable if there is only one photon. There is no such thing as 5/8 of a photon. It is either off there, or it isn't there. In Quantum mechanics, it is off the $\pm 5/8$ of the time. The relation of the two theories is clear.

What about the other kinds of polarization? For example, right-hand circular polarization? In the classical theory, right-hand circular polarization has equal complex components x and y , which are 90° out of phase. In the quantum theory, a right-hand circularly polarized (RHC) photon has equal amplitudes to be polarized $|x\rangle$ or $|y\rangle$, and the amplitude is $\sqrt{2}/2$ out of phase. Calling a RHC photon state R and a LHC photon state L , we can write (see Vol. 1, Section 11.1)

$$\begin{aligned} |R\rangle &= \frac{1}{\sqrt{2}} (|x\rangle + i|y\rangle), \\ |L\rangle &= \frac{1}{\sqrt{2}} (|x\rangle - i|y\rangle). \end{aligned} \quad (11.24)$$

The $1/\sqrt{2}$ is put in to get normalized states. With these states you can calculate any linear or interference effects you want, using the laws of quantum theory. If you want, you can also choose $|R\rangle$ and $|L\rangle$ as base states and represent every thing in terms of them. You only need to show that $\langle R|R\rangle = 0$, which you can do by taking the conjugate form of the first equation above (see Eq. (8.12)) and multiplying it by the other. You can resolve light into x - and y -polarizations, or the x' - and y' -polarized basis, in the right and left polarizations as a basis.

Just as an example, let's try to turn our formulae around. Can we express the state $|x\rangle$ as a combination of light and left? Yes, here it is:

$$\begin{aligned} |x\rangle &= \frac{1}{\sqrt{2}} (|R\rangle - |L\rangle), \\ |y\rangle &= -\frac{i}{\sqrt{2}} (|R\rangle - |L\rangle). \end{aligned} \quad (11.25)$$

Proof. Add and subtract the two equations in (11.24). It is easy to go from one basis to the other.

One curious point has to be made, though. The photon is right-circularly-polarized, it shouldn't have anything to do with the x and y axes. If we were to look at the same thing from a coordinate system turned at some angle about the direction of flight, the light would still be right-circularly polarized—and similarly for left. The right and left circularly polarized light are the same for any such rotation; the definition is independent of the choice of the x -direction (except that the phase difference is given by i , but that doesn't take any axes to define it). Much easier than a proof, isn't it? Just a matter of what you call the x , and the y , and x' together you call. Did I get which direction of x ? I think "and" and "..." and "..." don't always mean the same thing, but if you put them back together again and get x , we can answer the question in general by writing out the state $|R\rangle$, which represents a photon RHC polarized in the frame x', y' . In that frame, you would write

$$|R'\rangle = \frac{1}{\sqrt{2}} (|x'\rangle + i|y'\rangle).$$

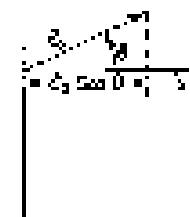


Fig. 11.4. The classical probability of the electric vector E .

How does each state look in the frame x, y ? Just substitute $\sqrt{2}$ from Eq.(11.32) and the corresponding $|y\rangle$ —we didn't write it down—but it is $(-\sin\theta, \cos\theta)$. Then

$$\begin{aligned} R_0 &= \frac{1}{\sqrt{2}} (\cos\theta |x\rangle + \sin\theta |y\rangle) = \cos\theta |x\rangle - i\sin\theta |y\rangle, \\ &= \frac{1}{\sqrt{2}} (\cos\theta + i\sin\theta) |x\rangle + i(\cos\theta - i\sin\theta) |y\rangle \\ &= \frac{1}{\sqrt{2}} (|x\rangle + i|y\rangle) (\cos\theta + i\sin\theta) \end{aligned}$$

The first term is $e^{-i\theta}|R\rangle$, and the second is $e^{+i\theta}|R\rangle$; our result is then

$$|R\rangle = e^{-i\theta}|R\rangle + e^{+i\theta}|R\rangle \quad (11.35)$$

The states $|R\rangle$ and $|R\rangle$ are the same except for the phase factor $e^{-i\theta}$. If you work out the sum $|R\rangle + |R\rangle$, you get that

$$|L\rangle = e^{-i\theta}|L\rangle \quad (11.36)$$

Now comes another surprise. If we add $|R\rangle$ and $|L\rangle$, we get something different from what we get when we add $|R\rangle$ and $|R\rangle$. For instance, the expectation value is $\langle L | (11.36), \cos\theta$ of $|L\rangle$ and $|L\rangle$ but a typical calculation is the sum $\langle L | (11.36)$ of the phase-shifted $|L\rangle$ backwards and $|L\rangle$ with -90° forward. That is just what we would get from the sum $|R\rangle + |R\rangle$ and $|L\rangle + |L\rangle$ in the second, angle $\theta = 90^\circ$ and east's right. An equalization of charges that is the same as the polarization in the original frame. So it is not exactly true that a stationary polarized photon looks the same for any set of axes. In short, these phase relations of the right and left circularly polarized components break off the polarization.

11.5 The nuclear K-meson

We will now describe a two-state system in the world of three-body particles—a system for which quantum mechanics gives a most remarkable prediction. To describe it completely would involve us in a lot of detail about particle systems, so we will, reluctantly, leave that out entirely. We can only give a outline of how a certain discovery was made—to show you the kind of reasoning that was involved. It begins with the discovery by Gell-Mann and Okubo of the octet of baryon states and of a new law of conservation of strangeness. It was with Gell-Mann and Pais we are studying the consequences of this new rule that they came up with the really interesting and remarkable phenomenon we are going to describe. First though, we have to tell you a little about "strangeness".

We must begin with what are called the strong interactions of nuclear particles. These are the interactions which are responsible for the strong nuclear force—as distinct, for example, from the relatively weaker electromagnetic interactions. The interactions are “strong” in the sense that if two particles get close enough to the point where $r \approx 10^{-15} \text{ m}$ with about a billion proton exchanges, either to nuclei or very nearby

to a similar system, we find (in Chapter 8) that a spin-orbit coupling when we repeat the construction above, the laws—then we get the “three nucleons” $|N\rangle$. In particular, when we work down at Section 5.7 on the $|+-+$ and $+--$ states of a quark-antiquark system’s wavefunction. The answer is a gluon loop which has however, no “color” here.

We can tell that the nature of this coupling is different from what we expect since at this point in our development, we argue that your spin-orbit coupling will, Section 11.6. If you are sufficiently curious, you may wish to come back to it later. We give it here because it is a beautiful example—taken from recent work in high-energy physics—of how one can start with our formulation of the quantum mechanics of two-particle systems

The nuclear particles themselves were to called a "weak interaction." By effect of this change can happen, such as beta decay, but also very slowly on a nuclear scale—the weak interactions are feeble, many orders of magnitude weaker than the strong interactions and good much weaker than electromagnetic interactions.

When the strong interactions were being studied with the big accelerators, people were surprised to find that certain things that "should" happen—had were reported to happen—but did not occur. For instance, in some interactions a particle of a certain sign the not appear when it was expected. Fermi and his team noticed that many of these predicted happenings could be explained if made by inventing a new conservation law: the conservation of strangeness. They proposed that there was a new kind of attribute associated with each particle—which they called its "strangeness" number—and that in any strong interaction, the "quality of strangeness" is conserved.

Suppose, for instance, that a high-energy negative K-meson with, say, an energy of many GeV collides with a proton. Most of the interaction may focus many other particles—mesons, baryons, lepton particles. Some particles are of the "strange" type indicated in Table 2-2 of Vol I. It is observed, however, that only certain combinations appear, and never others. Two certain conservation laws were already known to apply. First, energy and momentum are always conserved. The total energy and momentum of the system must be the same as before the event. Second, there is the conservation of electric charge which says that the total charge of the outgoing particles must be equal to the total charge carried by the original particles. In our example of a K-meson and a proton colliding, together, the following two cases do occur:

$$\begin{aligned} \text{K}^- &+ p \rightarrow p + K^- + \pi^+ + \pi^0 \\ \text{or} \quad K^- &+ p \rightarrow \Sigma^- + \pi^+ \end{aligned} \quad (11.3F)$$

We should never get:

$$K^- + p \rightarrow p + K^+ + \pi^0 \quad \text{or} \quad K^- + p \rightarrow \Lambda_0 + \pi^+ \quad (11.3G)$$

because of the conservation of charge. It was also known that the number of baryons is conserved. The number of baryons must be equal to the number of baryons in. For this law, an exception of a baryon is possible as mentioned before. This means that we can—and do—see

$$\begin{aligned} \text{K}^- + p &\rightarrow K^0 + \pi^0 \\ \text{or} \quad K^- + p &\rightarrow \bar{\nu} + \pi^0 + p + \bar{p} \end{aligned} \quad (11.4A)$$

(where $\bar{\nu}$ is the antineutrino, which carries a negative charge). But we never see

$$\begin{aligned} \text{K}^- + p &\rightarrow K^- + \pi^0 \\ \text{or} \quad K^- + p &\rightarrow K^- + \bar{e} \end{aligned} \quad (11.4B)$$

even when there is plenty of energy, because baryons would not be conserved.

These laws, however, do not explain the strange fact that the following reactions—which do not immediately appear, to be especially different from some of those in (11.3A) or (11.4B)—are also never observed:

$$\begin{aligned} \text{K}^- + p &\rightarrow \bar{\nu} + K^0 + K^0 \\ \text{or} \quad K^- + p &\rightarrow \gamma + \pi^0 \quad (11.4C) \\ \text{K}^- + p &\rightarrow \Lambda^0 + K^0 \end{aligned}$$

The explanation is the conservation of strangeness. With each particle goes a number—the strangeness S , and there is a law that in any strong interaction, the

Table 14-4

ANSWER *pass sentence(s) of the s⁺ and the t⁺s*

and a range of other molecular-level categories that exist. The postural anticipations (β_1 , β_2 , γ and δ are omitted from \mathbf{B}_1^*) and the transitions ($\pi = \pi^1$, π^- ; all have the same geneses) combine to make \mathbf{K}' and \mathbf{R}' because both have 217 geneses + 1; and \mathbf{K}' and \mathbf{T}' (the anti- \mathbf{K}') if the α^1 and α^2 genes are in \mathbf{K}' , \mathbf{K}' has 220 geneses + 1. There is also a gene with strangeness = 1—the hyper-kidneyed hairpin—and perhaps others as yet unknown. We have listed most of these components in Table 11-4.

Let's see how the strangeness conservation works in some of the reactions we have written down. Consider with a K⁻ and a proton, we have a total strangeness of $f = -1 + 0 = -1$. The conservation of strangeness says that the strangeness of products after the reaction must also add to -1 . That is to say, that is to let the reactivities of π^+ , π^- and π^0 be a , b , c . But in the reactions of (1), (2) the strangeness of the right hand side is zero in each case. Such reactions can't conserve strangeness and strangeness. Nobody knows. Nobody knows why not. We just have to go along with it.

Now let's look at the following reaction in the π^+ channel. You might, for instance, put a Λ^0 source plus a deuteron. Key words: two neutral particles. Now what would S do you add? Since the Λ strangeness is +1/2, the π^+ and Λ^0 have a strangeness of -1 and +1/2. In the first spectator reaction, the strangeness must change. So if particle one has strangeness = +1/2, it must be the Λ^0 . The reaction is



三

$S \cdot S + 2 = 1$!!! (incorrect)

If $\mathbf{t} \in \mathbb{R}^4$ and there is a slot of the \mathbf{b}^k , the component on the right would be $-\mathbf{t}^k$ — which means that not permit, since the string term on the left side is zero. On the other hand, a \mathbf{R}^4 can be produced in $\mathcal{O}(N^2)$ iterations, with \mathbf{t} .



$$\beta = 0, \quad \delta = 2, \quad \alpha = \pm 1 = \pm 1$$

9



$S = +1 + \dots$ $\theta = -1$

You may be thinking "What's all this stuff, because how do you know whether β is R^1 or R^2 ? They look exactly the same." They are superpositions of each other, so they have exactly the same mass, and both have zero electric charge.



Fig. II-5. High-energy events as seen in a hydrogen bubble chamber. (a) A p -nucleus interaction with a hydrogen atom produces a \bar{K} particle and a π^+ meson. Both particles decay in the chamber. (b) A \bar{K}^0 meson interacts with a proton producing a π^+ meson and a K^- particle which then decays. The decay products are shown. Their thermal trajectories are indicated by light dashed lines.

How do you distinguish them? By the particle they produce. For example, a \bar{K}^0 interacting with a nucleon becomes a π^+ and a K^- , see Fig. II-6:

$$\bar{K}^0 \rightarrow \pi^+ + K^-$$

and a K^0 meson. There is no way a K^0 can produce a particle which interacts with ordinary matter (protons and neutrons). So the experimental distinction between the K^0 and the \bar{K}^0 would depend on one of the two possibilities will not interfere with each other.

One of the predictions of the strangeness theory is that this will not occur. You can have high-energy nuclear \bar{s} -particle production with a nucleon. You know that the neutral kaons do; going into other states or maybe will never produce a \bar{s} . The experiment didn't do something like this. You can also have a \bar{K}^0 meson turn into a large hydrogen bubble chamber. A p -track appears, but sometimes else a pair of tracks appear (a proton and a π^+) indicating that a \bar{s} -particle has disappeared¹ (see Fig. II-5). Then you know that the \bar{K}^0 somewhere must have come from s .

You can, however, figure out where it is going by using the conservation law of momentum and energy. The weak interaction law, by disintegrating into two charged particles, as shown in Fig. II-6(a), so the \bar{K}^0 goes there, doing its best interaction with all the hydrogen nuclei (protons), scattering perhaps some other particles. The prediction of the strangeness theory is that it will always produce a \bar{s} -particle in a simple reaction like this.

$$\bar{K}^0 \rightarrow p \rightarrow \bar{K}^0 + \pi^+$$

Enough of K^0 mesons yet? This is the only s -channel \bar{K}^0 meson we have in the event sketched in Fig. II-5(b), in which the \bar{K}^0 is far because it decays into a K^0 meson. That's the first part of our story. That's the conservation of strangeness.

The conservation of strangeness is, however, not perfect. There are very slow disintegrations of the strange quarks (the s -quarks taking a long time, 10^{-10} to 10^{-12} second in which the s -quarks are not destroyed). These involve the "slow" decays. For example, the K^0 disintegrates into a Λ^0 and a $\bar{\Lambda}^0$ mesons ($\sim 7 \times 10^{-12}$

¹ Except, of course, if you measure the K^0 or other particles with a low enough resolution of $\frac{1}{2}$. We can think of situations in which the s channel might go to another s -channel strange particle.

² The first s -particle decays slowly since each interaction is so long it need not be completed. The decay products are either a p and a π^+ , or a n and a μ^0 . The lifetime is 2.7×10^{-10} sec.

³ An important limit for strange interconversion is given in Fig. II-7(a).

with a lifetime of 10^{-10} secnd. That was, in fact, the way K-mesons were first seen. Notice that this decay reaction

$$K^0 \rightarrow \pi^+ + \pi^-$$

does not conserve strangeness, so it cannot go "fict" by the strong interaction; it can only go through the weak decay process.

Now the K^0 also decays in the same way— $K^0 \rightarrow \pi^+ + \pi^-$ —and also with the same lifetime

$$K^0 \rightarrow \pi^+ + \pi^-.$$

Again we have a weak decay because it does not conserve strangeness. There's a principle that for any reaction, there is the corresponding reaction with "barred" indices by "antimatter" and vice versa. Since the K^0 is the antiparticle of the \bar{K}^0 , it should decay into the antiparticles of the π^+ and π^- ; but the antiparticle of a π^- is the π^+ (Or, if you prefer, vice versa.) It turns out that for the processes it doesn't matter which one you call "barred." So as a consequence of the weak decay, the K^0 and \bar{K}^0 can go into the same final products. When "barred" through their decays— π^+ in a bubble chamber, they look like the same particle. Only their strong interactions are different!

At last we are ready to describe the work of Gell-Mann and Pais. They discovered that since the K^0 and the \bar{K}^0 each contain two states of two σ -mesons, there must be some coupling between a K^0 state and its \bar{K}^0 , and also that a K^0 can turn into a \bar{K}^0 . Writing the reactions as new decay mechanisms, we would have

$$K^0 \rightarrow \pi^+ - \pi^- \rightarrow \bar{K}^0. \quad (1.43)$$

These reactions imply that there is some amplitude per unit time, say $-i/4t$ times $(K^0)^\dagger \Psi | \bar{K}^0 \rangle$, that a K^0 will turn into a \bar{K}^0 through the weak interaction responsible for this decay into two π -mesons. And there is the corresponding amplitude $(\bar{K}^0)^\dagger \Psi | K^0 \rangle$ for the reverse process. Because matter and antimatter behave in exactly the same way, these two amplitudes are numerically equal; we'll call them both A :

$$(K^0)^\dagger \Psi | \bar{K}^0 \rangle = (K^0)^\dagger \Psi | K^0 \rangle = A. \quad (1.44)$$

Kerr and Gell-Mann and Pais—here is an interesting situation. When people have been calling two distinct states of the world—the K^0 and the \bar{K}^0 —should really be considered as one two-state system, because there is an amplitude to go from one state to the other. For a complete treatment, our world, of course, has to deal with more than one particle, because there are four states of π 's, and so on. But since they were mainly interested in the reaction of K^0 and \bar{K}^0 , they did not have to complicate things and could take the approximation of a two-particle system. The other states were taken into account... to the extent that their effects appeared implicitly in the amplitudes of Eq. (1.44).

Accordingly, Gell-Mann and Pais analyzed the neutral particle as a two-state system. They began by choosing as their two basis states the states $| K^0 \rangle$ and $| \bar{K}^0 \rangle$. (From here on, the story goes very much as it did for the hydrogen molecule.) Any state $|\psi\rangle$ of the neutral K-particle could then be described by giving the amplitudes C_1 if you're in your first state, C_2 if you're in your second state:

$$C_1 = (K^0)^\dagger \Psi \langle \cdot \cdot \cdot | K^0 \rangle, \quad C_2 = (\bar{K}^0)^\dagger \Psi \langle \cdot \cdot \cdot | \bar{K}^0 \rangle. \quad (1.45)$$

The next step was to write the Hamiltonian equations for this two-state system. If there were no coupling between the K^0 and the \bar{K}^0 , the equations would be simply

$$\begin{aligned} i \frac{dC_1}{dt} &= E_1 C_1, \\ i \frac{dC_2}{dt} &= E_2 C_2. \end{aligned} \quad (1.46)$$

but since there is the amplitude $\langle K^0 | W | \bar{K}^0 \rangle$ for the K^0 to form a \bar{K}^0 , there would be the additional term

$$\langle K^0 | W | \bar{K}^0 \rangle C_1 = 4C_1$$

added to the right-hand side of the last equation. And similarly, the term $4C_{-}$ should be inserted in the equation for the case of change ΔC_{-} .

The last result. When the two-pion state is taken into account, the additional amplitude for the K^0 to form itself through ΔC_0 is

$$K^0 \rightarrow \pi^+ + \pi^- - K^0.$$

The additional amplitude, which we would write $\langle \pi^+ | W | K^0 \rangle$, is just equal to the amplitude $\langle K^0 | W | \bar{K}^0 \rangle$ since the amplitude to go from an \bar{K}^0 to a pair of mesons is identical for the K^0 and the \bar{K}^0 . If you wish, the argument can be carried out in detail like this. Just write

$$\langle \pi^+ | W | K^0 \rangle = \langle \bar{K}^0 | W | \pi^+ \rangle \langle \pi^- | W | K^0 \rangle$$

$$(147) \quad \langle K^0 | W | \bar{K}^0 \rangle = \langle K^0 | W | \pi^+ \rangle \langle \pi^- | W | K^0 \rangle.$$

Because of the symmetry of the two-pion state

$$\langle \pi^+ | W | K^0 \rangle = \langle 2\pi^- | W | K^0 \rangle,$$

and also

$$\langle K^0 | W | \bar{K}^0 \rangle = \langle K^0 | W | \pi^+ \rangle$$

It then follows that $\langle K^0 | W | \bar{K}^0 \rangle = \langle K^0 | W | K^0 \rangle$, and therefore $\langle \bar{K}^0 | W | K^0 \rangle = \langle K^0 | W | \bar{K}^0 \rangle$, so we will write ΔC_0 . Again, there are two additional amplitudes $\langle K^0 | W | \Delta C_0$ and $\langle K^0 | W | \Delta C_{-}$ having to do with which should be included in the Hamiltonian equations. The first goes to form K^0 or be equivalent to the exchange of ΔC_0 , and the second gives a new term ΔC_{-} in the equation for ΔC_{-} . Reasoning the way, Gell-Mann and Pais concluded that the three-pion vertex amplitudes for the $K^0 \bar{K}^0$ were unchanged by

$$\begin{aligned} iS \frac{\partial \Gamma_0}{\partial t} &= E_0 C_0 + 4C_0 = 4C_0, \\ iS \frac{\partial C_0}{\partial t} &= E_0 C_0 - 4C_0 = 4C_{-}. \end{aligned} \quad (147)$$

Now it is now evident concerning we have said in earlier chapters, that the amplitudes for $\langle \pi^+ | W | \bar{K}^0 \rangle$ and $\langle \bar{K}^0 | W | K^0 \rangle$ which are the source of odd-order, non-hermite complex conjugate. This was true when we were dealing with particles the did not decay. But if there is some decay, and even, therefore, because other — the two amplitudes are not necessarily complex conjugate. So the equality of (147) does not mean that the amplitudes are real numbers; they are in fact complex numbers. The coefficient 4 is, therefore, complex; and we can't just integrate it into the energy E_0 .

Having played often with diagrammatic and with our hermiticity, let the Hamiltonian equations of (147) prove that there was another pair of side terms enough to fit the need to represent the 3-particle system and which would make especially simple its behavior. They said, "Let's take the sum and difference of those two equations. Also, let's measure all the amplitudes from E_0 , and the units for

¹ We are making a simplification here. The γ source we have many states and a moving in various directions of the π mesons, but we should take the right-hand side of the equation after summing up all the γ source amplitudes. The amplitude would still work in the same way, we see.

energy and time that make $\alpha = 1/\gamma$? (That's where modern theoretical physics always goes.) I don't change the physics but makes the equations look a bit simpler from my point of view.

$$i \frac{d}{dt} (K_+ - C_+) = \mathcal{H}(K_+ - C_+), \quad i \frac{d}{dt} (K_- - C_-) = 0. \quad (1.46)$$

It is apparent that the combinations of amplitudes $(C_+ - C_-)$ and $(C_+ + C_-)$ act independently from each other (so a corresponding $i\mathcal{H}$ comes to the secondary states we have been studying earlier). So they conjecture that it would be more convenient to use a different representation for the K-particle. They define new variables:

$$|K_0\rangle = \frac{1}{\sqrt{2}} (|K^+\rangle - |K^-\rangle), \quad |K_1\rangle = \frac{1}{\sqrt{2}} (|K^+\rangle + |K^-\rangle) \quad (1.47)$$

They said instead of thinking of the K^+ and K^- mesons, we can equally well think in terms of the two "particles" (that is, baryons) K_1 and K_2 . (These correspond, of course, to the states we have already called $|C\rangle$ and $|A\rangle$. We are not changing α , it is just because we want some follow the definition of the original K -mesons and the way you will see in physics soon, etc.)

Now Goldstein and Pais didn't do all this just to get different names for the particles – there is some strange new physics in it. Suppose that C_+ and C_- are the amplitudes that come about $\psi(t)$ will be $C_0 \in K_0$, $C_1 \in K_1$:

$$C_0 = \langle K_0 | \psi \rangle, \quad C_1 = \langle K_1 | \psi \rangle.$$

From the equations in (1.46):

$$C_1 = \frac{i}{\sqrt{2}} (C_+ + C_-), \quad C_0 = \frac{i}{\sqrt{2}} (C_+ - C_-). \quad (1.48)$$

Then, the Eqs. (1.47) become

$$i \frac{dC_0}{dt} = 2\mathcal{H}C_0, \quad i \frac{dC_1}{dt} = 0. \quad (1.49)$$

The solutions are:

$$C_0(t) = C_0(0)e^{-i\mathcal{H}t}, \quad C_1(t) = C_1(0), \quad (1.50)$$

where, of course, $C_1(0)$ and $C_2(0)$ are the amplitudes at $t = 0$.

They also note say that if a real K -particle starts out in the state $|K_0\rangle$: $C_0(0) = 1$, then $C_1(0) = 0$, the amplitude at the time zero

$$C_0(0) = e^{-i\mathcal{H}t}, \quad C_1(0) = 0.$$

Remember up the \mathcal{H} is a complex number, so it is convenient to take $\mathcal{H} = \alpha - i\beta$. (Since the imaginary part of \mathcal{H} corresponds to energy loss, we will take $\beta > 0$.) With this substitution, $C_0(t)$ reads

$$C_0(t) = C_0(0)e^{-(\alpha - i\beta)t}. \quad (1.51)$$

The probability of finding a K_0 particle at t is the absolute square of this amplitude, $|e^{-(\alpha - i\beta)t}|^2$. And from Eq. (1.51), the probability of finding the K_1 state at that time is zero. That means that if you make a K-particle in the state $|K_0\rangle$, the probability of finding it in the same state decreases exponentially with time but you will never find it in the $|K_1\rangle$ state (except for a tiny quantum fluctuation). This agrees with the idea $\text{Rate} = 1/2\beta$ which is, experimentally, 10^{-11} sec. We made previous to the video we said that β was complex.

On the other hand, Fig. (1.51) says that if we make a K-particle completely in the K_1 state, it stays that way forever. Well, that's not really true. It's inferred large intervals of time, the interval between, say, the 800 times slower than the

recognition, many we have discussed. So there are some terms small terms we have left out in our approach, i.e., the so long as we are considering only the reaction $K^+ \rightarrow K^+$, the basis "correct".

Now in this history of Gell-Mann and Pais they were interested whether baryons when a π^+ -particle is produced into π^+ , which is a way in reaction. Since it can then have a probability of ~ 1 , it must be produced in the K^0 state. But $K^0 \rightarrow K^+$ is neither a K_1 nor a K_2 but a mixture. The initial conditions are

$$C_1(0) = 1, \quad C_2(0) = 0.$$

But that means—from Eq. (11.29)—that

$$C_1(0) = \frac{1}{\sqrt{2}}, \quad C_2(0) = \frac{1}{\sqrt{2}}.$$

And from Eq. (11.30), that

$$C_1(0) = \frac{1}{\sqrt{2}} e^{-imv/\omega}, \quad C_2(0) = \frac{1}{\sqrt{2}}. \quad (11.31)$$

Now remember that K_1 and K_2 are total contributions of K^0 and K^+ [in Fig. (11.24); the amplitudes were both taken as that $1/2 = 0$ the K^0 particle carried zero quantum by interference leaving only a K^+ alone]. But the K_1 's show coherence with time, and the K_2 's are also now. After $t = 0$ the interference of C_1 and C_2 will give little amplitudes for $\pi^+ + K^0$ and K^+ .

What does all this mean? Let's go back and think of the experiment we sketched in Fig. 11-2. A π^+ -beam has produced a K^0 particle and a K^+ meson which is moving along through the hydrogen in the chamber. As it goes along there is some small interaction between it and with hydrogen nuclei. At first, we thought that elementary considerations would prevent the K particle from taking a π^+ as such an interaction. Now however, we see that that is not right. For although our K particle starts out as a K^0 which cannot make a π^+ , it does not stay that way. After a while, there is some amplitude that it will have changed to the K^+ state. We can, therefore, according to what we have learned along the K -particle track. The content of this happens is given by the amplitude C_2 , which when we use Eq. (11.29) backwards relate to C_1 and C_3 . The relations is

$$C_2 = \frac{1}{\sqrt{2}} (C_1 - C_3) = (\pi^{+} v^{-\omega} - 1). \quad (11.32)$$

As the K nucleus goes along, the probability that it will "act like" a K^+ is equal to $|C_2|^2$, which is

$$|C_2|^2 = 2(1 - v^{-2\omega}) + 2v^{-2\omega} \cos \omega. \quad (11.33)$$

A complicated and strange result!

This then is the remarkable prediction of Gell-Mann and Pais, when a π^+ is produced, the chances that it will turn into a K^+ —as it can subsequently be found to produce a π^0 -meson with time according to Eq. (11.32). The prediction came from using only their logic and the basic principles of the quantum mechanics—were no knowledge at all of the inner workings of the K -particle. Since nobody knew anything about the inner machinery, that is as far as Gell-Mann and Pais could go. They could not give any theoretical values for v and ω . And actually has been able to do so to this date. They were able to give a value of ω calculated from the experimental decay rate of decay into $\pi^+ + \pi^0$ ($2\omega = 0.7$ sec), but they could say nothing about v .

We have plotted the function of Eq. (11.33) for two values of ω in Fig. 11-2. You can see that the $|C_2|^2$ depends very much on the value of ω . There is an K^0 probability at first—the K_1 builds up. If ω is large, the probability would have

large oscillations. It is to recall what will happen in the case of the double slit - the probability will just rise & settle to $\frac{1}{2}$.

Now, typically, in Raman effect we will be traveling at a constant speed near the speed of light - the curves of Fig. 11-6 then can represent the probability distribution of observing a E^2 wave (typical energy of photon) distributions. You can see why the position is unmeasurable quantum. You choose a single source and instead of just choosing one, determine n modes. Simultaneously in this situation, another measurement can tell you the total energy of a particle. This changes the probability - of measuring an effect varies in a strange way as it goes along. There is nothing else; it's like it is tame. And the most remarkable prediction was made solely by applying quantum mechanics of equilibrium.

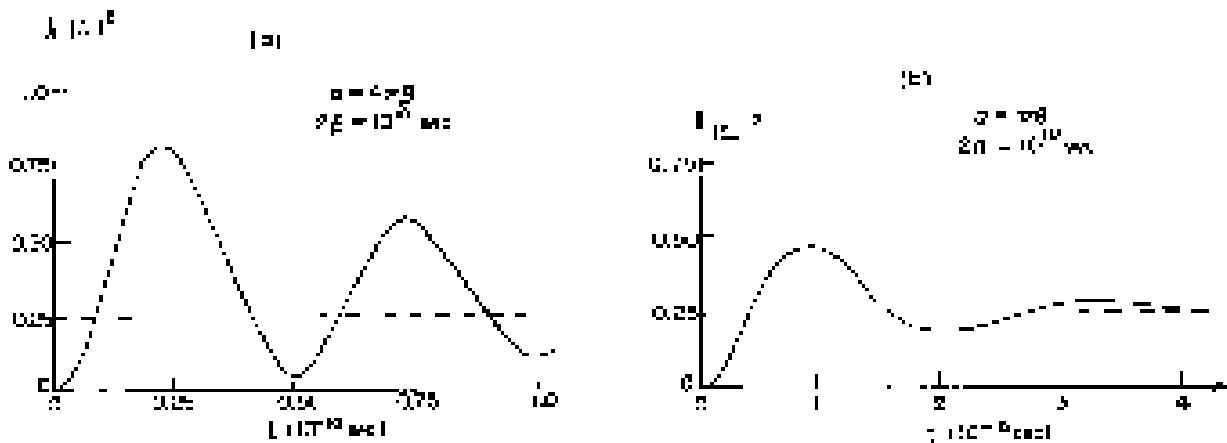


Fig. 11-6. The function of Eq. 11-10 for (a) $\alpha = \beta$, (b) $\alpha = 4\beta$.
 $\text{With } \beta = 10^{-10} \text{ sec}.$

If there is any place where we have a chance, or if the basic principles of quantum mechanics in the present way access the superposition of simple, the weak or these? it is α/β . Despite the fact that this effect has been predicted now for several years, there is no experimental confirmation; that is very clear. These calculations might tell us what value of α/β the experiments find that the effect really occurs. They indicate that it between 30 and 40. Let's hope to experimentally. It would be very nice, of course, to do some exactly to see if the principle of superposition really works in such a mysterious world as the of the strange particles, with unknown reasons for the decay, and unknown reasons for the disappearance.

The paradoxes we have just described is not a characteristic of the way quantum mechanics is being used today in the search for a fundamental theory of the strange particles. All the complicated theories that you may hear about are not much and no less than this kind of elements of nature than using the principle of superposition and other principles of quantum mechanics of the level. Some people claim that they have theories by which it is possible to understand the α/β ratio. I agree, but I say that these theories are completely useless. For instance, the theory that attempts to calculate α/β given the ρ tells us that the value of ρ should be infinite. The set of equations with which they originally start involves two components. They pass from the last ρ 's back to a λ' , and so on. When all worked out, it does indeed produce a pair of equations like the ones we have here, but because there are an infinite number of terms of $\rho = \lambda'$, depending on their number n , averaging over all the possibilities gives an α/β which is infinite. Our nature's ρ is not infinite, so the fundamental theory is not unique. It is not quite remarkable that the phenomena which can be predicted at all can be well by the strange particles come from the principles of quantum mechanics of the level to which you are learning them now.

11.4 Generalization to N-state systems

We have finished with all non-perturbative systems we wanted to talk about! The following sections we will go on to study systems with more states. The connection to N-state systems of the form we have worked out for two states is pretty straightforward. It goes like this:

If a system has N distinct states, we can represent any state $\psi(t)$ as a linear combination of any set of basis states ϕ_i , where $i = 1, 2, 3, \dots, N$:

$$\psi(t) = \sum_{i=1}^N c_i(t) \phi_i(t). \quad (1.57)$$

The coefficients $c_i(t)$ are the amplitudes $\phi_i(t)$. The behavior of these amplitudes c_i with time is governed by the equations

$$i\hbar \frac{dc_i(t)}{dt} = \sum_j H_{ij} c_j \quad (1.58)$$

where the energy matrix H_{ij} describes the physics of the problem. It looks the same as for two states. Only now, there are N energy eigenvalues and basis states, and the energy matrix is H_{ij} , $i, j = 1, \dots, N$. The Hamiltonian is an N by N matrix with N^2 elements. As before, $H_{ii} = E_i$, so it's a diagonal matrix, and the diagonal elements H_{ii} are real numbers.

We have found a general solution to the EEs of a two state system, so for the energy matrix we can just plug in equation (1.58). It is also not difficult to solve Eq. (1.58) for an N state system when H is not time-dependent. Again, we begin by looking for a possible solution in which the amplitudes all have the same time dependence. We try

$$c_i = e^{i\omega t} C_i \quad (1.59)$$

When these C_i 's are substituted into (1.58), the derivatives $i\hbar \omega/c_i$ become just $-i\hbar \omega C_i$. Canceling this common exponential factor from (1.58), we get

$$\dot{E} c_i = \sum_j H_{ij} c_j. \quad (1.60)$$

This is a set of N linear algebraic equations for the N unknowns c_1, c_2, \dots, c_N , and there is a solution only if you are lucky—only if the determinant of the coefficient matrix is non-zero. But this is not necessary to be that sophisticated: you can just start to solve the equations one way you want, and you will find that they can be solved only for certain values of C . (Remember that C is the only adjustable thing we have in the equations.)

If you want to be formal, however, you can write Eq. (1.60) as

$$\sum_j (H_{ij} - \omega_j) c_j = 0. \quad (1.61)$$

Then you can use the rule—if you know ω_j —that no solutions will have a value other than these values of ω for which

$$\text{Det}(H_{ij} - \omega_j E) = 0. \quad (1.62)$$

Each term of the determinant is just H_{ij} , except that E is subtracted from every diagonal element. That is, it looks more just

$$\text{Det} \begin{pmatrix} H_{11} - E & H_{12} & H_{13} & \dots \\ H_{21} & H_{22} - E & H_{23} & \dots \\ H_{31} & H_{32} & H_{33} - E & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix} = 0. \quad (1.63)$$

This is, of course, just a special way of writing an algebraic equation for \mathcal{E} which is the sum of a broad of products of all the terms taken in certain ways. These processes will give us the picture of \mathcal{E} up to S^2 .

So we have an N -order polynomial equal to zero, and there are, in general, N roots. (We must remember, however, that some of them may be multiple roots, meaning that two or more roots are equal.) Let's call the N roots

$$E_1, E_2, E_3, \dots, E_n, \dots, E_N. \quad (1.64)$$

(We might also represent each with Roman numeral, or let n take on the values $1, 2, \dots, N$. It may be that some of these energies are equal, say $E_1 = E_{n'}$, but we will still choose to call them different names.

The equations in Eq. (1.64) are one equation for each value of k . It can contain one of the E 's, say E_1 , and (1.64) will solve for the n , from which belongs to the energy E_1 . We will call this set $\{\alpha\}$.

Using these α 's in Eq. (1.64), we have the complete solution, but the other $N-1$ equations remain to be solved. In taking n values for the remaining $N-1$ of the N energy states at $k=0$, we are using

$$C(n) = b_n n! e^{-\beta E_n},$$

and

$$C(0) = c(0). \quad (1.65)$$

The complete definite energy state $|E_n\rangle$ can then be written as

$$|\psi_n(k)\rangle = \sum_{\alpha} \text{coeff}_{\alpha} e^{-\beta E_n} |E_n\rangle, \quad (1.66)$$

The state has one, of course, the only exception of the definite energy states, because the time dependence factor is one. Then they are constant vectors which can be used as a new basis as I have thought.

Let's do this using b_n and the α 's, as you can easily see—that when operated on by the Hamiltonian operator \hat{H} it gives just E_n times the same state:

$$\hat{H}|\psi_n\rangle = E_n |\psi_n\rangle. \quad (1.67)$$

The energy E_n is, then, a number which is a characteristic of the Hamiltonian operator \hat{H} . As we have seen, a Hamiltonian is Q^* . In general, one cannot obtain its eigenvalues. In the mathematics world they would be called the "eigenvalues proper" of the Q^* or H_0 . Physicists usually call them the "eigenvalues" of \hat{H} ("Eigen" is the German word for "characteristic" or "proper"). These are eigenvalues of \hat{H} . In other words, for each energy E_n is the state of definite energy, which we have called the "definite energy". Physicists usually call the states of "eigenstates of \hat{H} ". Each eigenstate corresponds to a particular representation.

Now, generally, the states $|\alpha\rangle$ —of which there are N —can't be independent bases. For this to be true, all of the states must be orthogonal, meaning that the overlap of them, say $|\alpha\rangle$ and $|\beta\rangle$,

$$\langle \alpha | \beta \rangle = 0. \quad (1.68)$$

This will be true automatically if all the energies are different. Also, we can multiply all the α 's by a suitable factor so that the states are normalized. So that we need that

$$\langle \alpha | \alpha \rangle = 1 \quad (1.69)$$

for all n .

When it happens that \hat{H} (Eq. 1.67) accidentally has two (or more) roots with the same energy, there are some easier complications. First, there are still two $\langle \alpha | \alpha \rangle$ and $\langle \beta | \beta \rangle$ which go with the two equal energies for the states $|\alpha\rangle$ and $|\beta\rangle$.

now and be orthogonal. Suppose you performed the normal procedure and find two stationary states with equal energies—let's call them $|v\rangle$ and $|w\rangle$. Then it will not necessarily be so that they are orthogonal—“you’re kidding.”

$$\langle v | v \rangle = 1.$$

It is, however, true that you can end up two new states, which we will call $|v'\rangle$ and $|w'\rangle$, that have the same energies and are also orthogonal, so that

$$\langle v' | v' \rangle = 1. \quad (11.69)$$

You can do this by making $|v'\rangle$ and $|w'\rangle$ a suitable linear combination of $|v\rangle$ and $|w\rangle$ with the ~~constraint~~ chosen to make it come out right. Eq. (11.70) is true. It is always convenient to do this. You will probably encounter this situation so that you can always associate two proper energy states $|n\rangle > n$, all orthogonal.

We would like, for fun, to prove the other two of the ordinary states have different energies, they are indeed orthogonal. For the state $|0\rangle$ with the energy E_0 , we have that

$$\hat{H}_0 |0\rangle = E_0 |0\rangle. \quad (11.70)$$

This operator equation really means the $\langle 0 | \hat{H}_0 | 0 \rangle$ is EQUAL TO zero numbers. Finding the missing part $\langle 0 | \hat{H}_0 | 0 \rangle$ means the same as

$$\sum_i \langle 0 | \hat{H}_0 | i \rangle \langle i | 0 \rangle = E_0 \langle 0 | 0 \rangle. \quad (11.71)$$

If we take the complex conjugate of this equation, we get

$$\sum_i \langle 0 | \hat{H}_0 | i \rangle^* \langle i | 0 \rangle^* = E_0^* \langle 0 | 0 \rangle^*. \quad (11.72)$$

Remember that $\langle 0 | \hat{H}_0 | i \rangle^*$ is a complex number. The magnitude of a complex number is the same as its real part, so (11.72) can now be written as

$$\sum_i \langle 0 | \hat{H}_0 | i \rangle^* \langle i | \hat{H}^* | 0 \rangle = E_0^* \langle 0 | 0 \rangle. \quad (11.73)$$

Now, this equation is valid for any i in its “closed form” is

$$\langle 0 | \hat{H}^* | 0 \rangle = E_0^* \langle 0 | 0 \rangle, \quad (11.74)$$

which is just identical to Eq. (11.71).

Now we can easily prove that E_0 is a real number. We multiply Eq. (11.73) by $\langle 0 |$ to get

$$\langle 0 | \hat{H}^* | 0 \rangle = E_0. \quad (11.75)$$

Since $\langle 0 | 0 \rangle = 1$, then we multiply Eq. (11.75) on the left by $| 0 \rangle$ to get

$$\langle 0 | \hat{H} | 0 \rangle = E_0^2. \quad (11.76)$$

Comparing (11.76) with (11.73) it is clear that

$$E_0 = E_0^*, \quad (11.77)$$

which shows that E_0 is real. We can make the same for E_n in Eq. (11.75).

Finally we can easily to show that two different energy states are orthogonal. Let $|m\rangle$ and $|n\rangle$ be any two of the definite energy basis states. Using Eq. (11.75) for the state $|m\rangle$ and multiplying by $\langle m |$, we get that

$$\langle m | \hat{H} | n \rangle = E_{mn} \langle m | n \rangle.$$

But if we multiply (11.71) by $|m\rangle$, we get

$$\langle m | j | n \rangle = E_n |m|n\rangle$$

Since the left sides of these two equations are equal, the right sides are also:

$$E_m |m\rangle |n\rangle = E_n |m\rangle |n\rangle \quad (11.79)$$

If $E_m = E_n$, the equation does not tell us anything. But if the energies of the two states $|m\rangle$ and $|n\rangle$ are different ($E_m \neq E_n$), Eq. (11.79) says that $|m\rangle |n\rangle$ must be zero, as we wanted to prove. The two states are consequently orthogonal so long as E_m and E_n are numerically different.

The Hyperfine Splitting in Hydrogen

12-1 Base states for a system with two spin one-half particles

In this chapter we take up the "hyperfine splitting" of hydrogen. Because it is a physically interesting example of what we can already do with quantum mechanics, it can exemplify a more far-reaching idea, and it will illustrate all the methods of quantum mechanics as applied to slightly more complicated systems. It is enough more complicated that once you see how this one is treated you can perhaps predict the generalization to all kinds of systems.

As you know, the hydrogen atom consists of an electron moving in the neighborhood of the proton, where it can exist in any one of a number of discrete energy levels, in addition to half the pattern of lines of the exact one-electron. The first excited state, for example, lies just at a Rydberg, or about 10 eV above its ground state. But even the so-called excited only one hydrogen carries really a single definite energy state because of the spin of the electron and the proton. These spins are responsible for the hyperfine splitting. The energy levels, while split, all the energy levels are present pretty much here.

The electron can have its spin either "up" or "down" and the proton can also have its spin either "up" or "down." That is, there are four possible spin states for every electronic condition of the atom. That is, when people say "the ground state" of hydrogen they really mean the "true ground state," and the first excited state. The four spin states do not all have exactly the same energy; there are slight shifts from the energies we would expect with no spin. The shifts are, however, much much smaller than the 10 eV shift from the ground state to the next state above. As a consequence, each electronic state has no energy splitting into very many energy levels—the so-called hyperfine splitting.

The energy differences among the four spin states is what we want to calculate in this chapter. The hyperfine splitting is due to the interaction of the magnetic moments of the electron and proton, which gives a slightly different magnetic energy for each spin state. These energy shifts are only about one billionth of an electron volt, and are small compared with 10 eV's. It is because of this tiny gap that we can think about the ground state of hydrogen as a "true state" system, without worrying about the fact that there are really many more states at higher energies. We are going to have something like a theory of the hyperfine structure of the ground state of the hydrogen atom.

For our purposes we cannot introduce any of the details about the structure of the electron and proton because that has all been worked out by the theory of the spin—it has worked itself out by getting to the ground state. We need know only that we have an electron and proton in the neighborhood of each other with some definite spatial relationship. In addition, they can have various different relative orientations of their spins. It is only the effect of the spins that we want to look into.

The first question we may have to answer is: What are the base states for the system? Now the question has been put incorrectly. There is no such thing as "the" base states because, of course, the set of base states you may choose is not unique. You can in always make out of three combinations of the sort. There are always many choices for the base states, and among them, any choice is equally good. So the question is not what is the base set, but what one do you use? Well, when choosing a set we wish for a measure of bases. It is always best to start with a base set which is physically reasonable. It may not be the solution

12-1 Base states for a system with two spin one-half particles

- 12-2 The Hamiltonian for the ground state of hydrogen
- 12-3 The energy levels
- 12-4 The Zeeman splitting
- 12-5 The states in a magnetic field
- 12-6 The propagation matrix for spin one

to any problem, or may not have any direct importance, but it will generally make it easier to understand what is going on.

We choose the following four base states:

State 1: The electron and proton are both spin "up."

State 2: The electron is "up" and the proton is "down."

State 3: The electron is "down" and the proton is "up."

State 4: The electron and proton are both "down."

We need a handy notation for these four states, so we'll represent them this way:

State 1: $|+\rangle$; electron up, proton up

State 2: $+ - \rangle$; electron up, proton down

State 3: $- + \rangle$; electron down, proton up

State 4: $- - \rangle$; electron down, proton down

(C2.1)

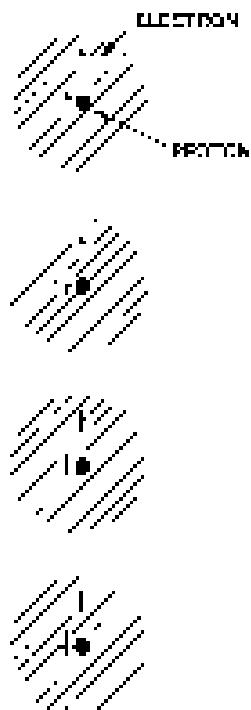


Fig. C2-1. As we see here under the ground state of the hydrogen atom,

You will have in mind either that the first plus sign indicates the electron and the second, is, the proton. For handy reference, we'll also communicate the notation $|+\rangle, |-\rangle, |+\rangle, |-\rangle$. Sometimes, it will also be convenient to use these codes |+, -, +, ->.

You may say, "But the particles interact, and maybe these aren't the right basis states. I propose as though you are considering the two particles independently." Yes, indeed! This is interesting, raises the problem, when is the *interaction* for the system. But the interaction is not involved in the question of how to describe the system. When we choose the base states are according to what other is happening. It may be that the two cannot stay in one of these two states, well, if it is stated this way. That's another question. That's the question: How do the amplitudes change with time in a particular (fixed) case? Describing the two states, we are just choosing our "unit vectors" for our description.

Well, we're in the subject, let's look at the general problem of finding a set of basis states when there is more than one particle. You know the basis states for a single particle. An electron, for example, is completely described in no life—not in our simplified cases, but in real life—by giving the amplitudes (or the coefficients of the following states:

, electron "up" with momentum p_1

$|+$

, electron "down" with momentum p_1

There are really two infinite sets of states, one state for each value of p_1 . There is to say that an electron state $|+\rangle$ is completely described if you know all the amplitudes

$$|+\rangle = |\psi_1^+(p_1)\rangle \quad \text{and} \quad |-\rangle = |\psi_1^-(p_1)\rangle$$

Where the $\psi_1^+(p_1)$ represents the component of angular momentum, along some axis—usually the z axis, and p_1 is its wave-momentum. This means, therefore, the two amplitudes for every possible momentum (a multi-infinite set of cases). That is all there is to describing a single particle.

When there are more than one particle, the base states can be written in a similar way. For instance, if there were an electron and a proton in a more complicated situation (for example, considering the base states of both of the following kind).

|an electron with spin "up" moving with momentum p_1 , and

a proton with spin "down" moving with momentum p_2 ,

And so on, for other spin combinations. If there are more than two particles, so much the same. So you see that the infinite number of base states is really very large. The only problem is, where is the Hamiltionian?

For our study of the ground state of hydrogen we won't need to use the full set of base states for the six spins momenta. We can specialize particular one.

moment shared by the proton and electron when we say "the spin, \vec{s} , is down." The details of the configuration, i.e., computes the all the momentum, how states can be calculated, etc., that is another problem. Now we are concerned only with the effects of the spin, so we will focus only the S_{tot} base states of (12.1). Our next problem is: What is the Hamiltonian for this up of spins?

12-2 The Hamiltonian for the ground state of hydrogen

We'll tell you in a moment when it is that we should remind you of one thing: any state can always be written as a linear combination of the base states. For any state $|\psi\rangle$ we can write

$$|\psi\rangle = \alpha_1 |+\rangle + \alpha_2 |-\rangle + \alpha_3 |+\rangle + \alpha_4 |-\rangle + \alpha_5 |+\rangle + \alpha_6 |-\rangle \quad (12.2)$$

Remember that the complex brackets are just complex numbers, so we can choose them in any way. Let's use C_i, where $i = 1, 2, 3, 4, 5$, and write Eq. (12.2) as

$$|\psi\rangle = \alpha_1 |+\rangle_C_1 + \alpha_2 |-\rangle_C_2 + \alpha_3 |+\rangle_C_3 + \alpha_4 |-\rangle_C_4 + \alpha_5 |+\rangle_C_5 + \alpha_6 |-\rangle_C_6 \quad (12.3)$$

By giving the four amplitudes, C_i , we completely describe the spin state $|\psi\rangle$. If these four amplitudes change with time, i.e., they will, the rate of change in time is given by the operator H . The problem is to find out H .

There is no general rule for writing down the Hamiltonian of an atomic system, and finding the right formula is much more difficult than finding a set of new states. We were able to tell you a general rule for writing a set of base states for my problem of a proton and an electron, but to come to the general Hamiltonian of such a combination is beyond us at this level. Instead, we will find you the Hamiltonian by guess by trial and error—and you will have to accept it as the correct one because the results will agree with the test of experimental measurement.

You will remember that in the last chapter we were able to describe the Hamiltonian of a single, spin-one-half particle by using the sigma operators—the various equivalent sigma operators. The properties of the operators are summarized in Table 12-1. These operators, which has just a momentum shift and stay at keeping track of the spin, is elements of the type $(|+\rangle, \sigma_x, |-\rangle)$ —useful for describing the behaviour of a single-particle of spin one-half. The question is: can we find an analogous device to describe a system with two spins? The answer is yes, very simply, as follows. We are looking which we will call "Sigma electron," which we represent by the sigma operator σ_y^e , and which has two components, σ_{y1}^e and σ_{y2}^e . We now know the condition that when the σ_y^e operator acts upon our six base states of the hydrogen atom, it acts only on the electron spin, and in exactly the same way as if the electric field of itself. Example: What is $\sigma_y^e |+-\rangle$? Since e_2 is an electron, "down" is $-$ times the run-around up state with the electron "up".

$$\sigma_y^e |+-\rangle = -|+\rangle |-\rangle$$

Other, σ_y^e acts on the combined state it will give the electron, but does nothing to the proton and multiplies the result by $-i\hbar$. Operations on the other states, σ_y^e would give

$$\begin{aligned}\sigma_y^e |++\rangle &= i\hbar |++\rangle \\ \sigma_y^e |+-\rangle &= -i\hbar |-\rangle \\ \sigma_y^e |--\rangle &= -i\hbar |--\rangle\end{aligned}$$

You remember that the operators σ_y^e work only on the first spin symbol—that is, on the electron spin.

Next we define the corresponding operator "Sigma proton" for the proton spin. Its three components σ_x^p , σ_y^p , σ_z^p act in the same way as σ_y^e only on the

Table 12-1

$\sigma_x +\rangle =$	$ +\rangle$
$\sigma_x -\rangle =$	$ -\rangle$
$\sigma_z +\rangle =$	$+ +\rangle$
$\sigma_z -\rangle =$	$+ -\rangle$
$\sigma_y +\rangle =$	$ +\rangle$
$\sigma_y -\rangle =$	$- +\rangle$

your spin. For example, if we have σ_z acting on each of the four basis states we get, always using Table 1.1 —

$$\begin{aligned}\sigma_z^1 | + \rangle &= | + \rangle, \\ \sigma_z^2 | + \rangle &= | + \rangle, \\ \sigma_z^3 | + \rangle &= | - \rangle, \\ \sigma_z^4 | - \rangle &= | + \rangle.\end{aligned}$$

As you can see, it's not very hard.

Now in the next general case we would have more complex things. For instance, we could have products of the two operators like $\sigma_x\sigma_y^2$. When we have such a product we do first what the operator on the right says, and then do what the other one does. For example, we would have this

$$\sigma_x\sigma_y^2 = -i(-i\sigma_x) = i(-i\sigma_x)(-i\sigma_y) = -i(-i\sigma_y) = -i\sigma_y.$$

Now, think these operators don't do anything on pure numbers—we have used this fact when we defined $\langle -| = (-D)$. However, that the operator "commutes" with pure numbers, so that a factor "can be moved through" the operator. You can practice by showing that the product $\sigma_x\sigma_y^2$ gives the following result for the four states:

$$\begin{aligned}\sigma_x\sigma_y^1 | + \rangle &= | + \rangle, \\ \sigma_x\sigma_y^2 | + \rangle &= | + \rangle, \\ \sigma_x\sigma_y^3 | + \rangle &= | - \rangle, \\ \sigma_x\sigma_y^4 | - \rangle &= | - \rangle.\end{aligned}$$

If we take τ_{-} the possible operators, using each state as operator only once, there are sixteen possibilities. Using commutativity we include also the "unit operator" 1. First, let's look at the $\sigma_x\sigma_y^2$. This has 16 states, $\tau_1, \tau_2, \dots, \tau_{16}$, which makes six. In addition, there are the 16 possible products of the form $\sigma_x\sigma_y^2$, which makes a total of 16. And then there is the unit operator which just leaves any given state alone. Sixteen in all.

Now, note that for a four-state system, the Hamiltonian matrix has to be a 16x16 matrix of coefficients. It will have zeros entries. It is easily demonstrated that any two $\sigma_x\sigma_y^2$ matrices will, therefore, be commutative. In particular, can be written as a linear combination of the sixteen double-spin operators corresponding to the set of operators we have just made up. Therefore, the commutation between a proton and an electron (i.e. involves only their spins) but you expect that the commutation operation can be written as a linear combination of the same 16 operators. The only question is, how?

Well, first we know that the interaction doesn't depend on our choice of coordinate system at all. If there is no central disturbance like a magnet, the basic interaction is only in direction of space, the Hamiltonian can't depend on the choice of the direction of the x , y , and z axes. That means that the Hamiltonian can't have a term like $\sigma_x\sigma_y^2$. It would be ridiculous, because then somehow with a different coordinate system would go off-center value.

The only possibilities are to come with the unit matrix, say a constant v (from Eq. 1.1), and some combination of the sigma's that does not depend on the coordinates—say, "intrinsic" combination. The only possible invariant combination of the vectors is the dot product, which is zero's to

$$\sigma_x\cdot\sigma_y^2 = \sigma_x\sigma_y^2 = \sigma_x\sigma_y^2 + \sigma_y\sigma_x^2. \quad (1.2.1)$$

This quantity is invariant, will depend to any relabeling of the coordinate system.

[†]For these particular operators you will notice I have got that the sequence of the operators don't matter.

So the only possibility for a Hamiltonian with the proper symmetry to give us a constant term in mutual is plus a constant times the dot product, say,

$$\hat{H} = E_0 + \vec{A} \cdot \vec{\sigma} \cdot \vec{B}_0. \quad (12.2)$$

That's our Hamiltonian. It's the only thing that it can be, by the symmetry of space, no long or short from external field. The constant term doesn't tell us much; it just depends at this level we choose a constant energies from. We must have, as well, $E_0 > 0$. The second term will be all we need to show we find the total binding of the hydrogen.

If you want to, you can think of the Hamiltonian in a different way. If the two are two magnets near each other with magnetic moments μ_1 and μ_2 , the mutual energy will depend on $\mu_1 \cdot \mu_2$, among other things. And, you remember, we found that the classical thing we call the energy is quantum mechanics is $\mu_1 \cdot \mu_2$. Similarly, when you calculate $\mu_1 \cdot \mu_2$ will give you minus in quantum mechanics to be $\mu_1 \cdot \mu_2$ (where μ_2 is the negative moment of the proton, which is about 10.0 times smaller than μ_1 , and has the opposite sign). So Eq. (12.2) says that the interaction energy is like the interaction between two magnets, only not quite, because the interaction of the two magnets depends on the total distance between them. But Eq. (12.2) could represent, in fact, a combination of an average interaction. The electron is moving all around itself, he says, and our theory which gives only the average interaction energy. All it says is the \vec{B}_0 is a vector representing in space for the electron and proton there is an energy proportional to the cosine of their angle between. So the magnetic moment is speaking classically. Such a classical evaluation picture may help you to understand where it comes from, but the important thing is that Eq. (12.2) is the correct quantum mechanical formula.

The order of magnitude of the classical interaction between two magnets would be the product of the two magnetic moments divided by the cube of the distance between them. The reason: he sees the electron and the proton in the hydrogen atom is speaking roughly, one has to stick to each other, or 3.5 separated. It is, therefore, possible to make a crude estimate that the constant should be proportional to the product of the two magnets divided $\mu_1 \cdot \mu_2$, divided by the cube of the separation. Such an estimate gives a number 1.0e-12, perhaps. Then $\vec{B}_0 > 0$ that I can probably estimate once you understand the complete quantum theory of the hydrogen atom—which we are not doing. It has, in fact, been calculated in seconds of about 20 years in one million. So, unlike the dipole constant d of the ammonia molecule, which couldn't be calculated at all well by a classical theory, for the hydrogen can be calculated from complete quantum theory. But never mind, we will for present purposes think of the \vec{B}_0 as a number which needs to be determined by experiment and analyze the physics of the situation.

Taking the Hamiltonian of Eq. (12.2) we can use . with the equation

$$AC = \sum_i B_i C_i \quad (12.3)$$

and out what the spin matrix elements the energy levels. To do that, we need to work in the mixed matrix element $\langle M_i | B_i | M_j \rangle = 0$, B_i is corresponding to each one of the four wave states in (12.1).

We begin by working in what $B|0\rangle$ is the zeroth of the four basis states for $\sigma = \pm \frac{1}{2}$,

$$B = (\vec{\sigma} - \vec{\mu}_1 \cdot \vec{\sigma}_1^0 - \vec{\mu}_2 \cdot \vec{\sigma}_2^0) = \vec{\sigma} = \vec{\sigma}_1^0 + \vec{\sigma}_2^0 = \sigma(\vec{\sigma}). \quad (12.4)$$

Using the method we described a little earlier, it's easy if you have memorized Table 12-1—just find what each pair of σ 's does on $|+\pm\rangle$. The answer is

$$\begin{aligned} \sigma\sigma^0_1 + \sigma^0_2 &= |+\pm\rangle \\ \sigma\sigma^0_2 - |\pm\rangle &= |-\pm\rangle, \\ \sigma\sigma^0_1 - |\pm\rangle &= |+\mp\rangle. \end{aligned} \quad (12.5)$$

be (12.7) becomes

$$\hat{H} = +; \quad A(- -) - - -; + + -) \cdot \mathbf{I} \mid 1 \rangle \lambda = (12.9)$$

Since our basis states are all orthogonal, this gives us immediately that

$$\begin{aligned} (- + \mid H \mid - -) &\sim (0 - + \mid 1 \mid -) = 0, \\ (- - \mid H \mid + +) &\sim A(- - + +) = 0 \quad (12.10) \\ (- + \mid H \mid + +) &\sim A(- + + +) = 0, \\ (- - \mid H \mid + +) &\sim A(- - + +) = 0. \end{aligned}$$

Assimilating each $A(- - + +) / \lambda^2$, we can easily write down the differential equation for the amplitudes C_1 :

$$\begin{aligned} dC_1 - H_{11}C_1 - H_{12}C_2 + H_{13}C_3 - H_{14}C_4 \\ = 0 \\ dC_1 = AC_1. \end{aligned} \quad (12.11)$$

Table 12.2

Spin operators for the hydrogen atom

σ_x^2	$\sim +;$	$- - -;$
σ_y^2	$\sim - +;$	$+ + -;$
σ_z^2	$\sim +;$	$- + -;$
$\sigma_y\sigma_z$	$\sim +;$	$+ + -;$
$\sigma_x\sigma_z$	$\sim +;$	$- - +;$
$\sigma_x\sigma_y$	$\sim + -;$	$- + -;$
$\sigma_y\sigma_x$	$\sim + -;$	$- + -;$
$\sigma_x\sigma_y\sigma_z$	$\sim + - +;$	$- + - +;$
$\sigma_y\sigma_x\sigma_z$	$\sim + - +;$	$- + - +;$
$\sigma_z\sigma_x\sigma_y$	$\sim + - +;$	$- + - +;$
$\sigma_z\sigma_y\sigma_x$	$\sim + - +;$	$- + - +;$

That's all! We get only the one term.

Now, since the rest of the Hamiltonian equations we have to solve through the same procedure for \hat{H} operating on the other states. Then, we will let you proceed by checking out all of the sigma products we have written down in Table 12.2. Then, we can use them to get:

$$\begin{aligned} i(- + -) - A(- + +) &= -C_1, \\ i(- - +) - A(2(+ -) - - +) &= -C_2, \\ i(- - -) - A(- - +) &= C_3. \end{aligned} \quad (12.12)$$

Then, multiplying each one of them on the left by all the other state amplitudes, we get the following Hamiltonian matrix, H_{ij} :

$$H_{ij} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -4 & 2A & 0 \\ 0 & 2A & -A & 0 \\ 0 & 0 & 0 & A \end{bmatrix}. \quad (12.13)$$

It means, of course, nothing more than that the differential equations for the four amplitudes C_i are

$$\begin{aligned} dC_1 &= AC_1, \\ dC_2 &= -4C_1 - 2AC_3, \\ dC_3 &= 2AC_2 - AC_4, \\ dC_4 &= AC_3. \end{aligned} \quad (12.14)$$

Before solving these equations we can't resist adding one about a clever rule due to Dirac: it will make you feel that you are really advanced although you don't read it for the work. We have seen from the equations (12.9) and (12.12) that

$$\begin{aligned} \sigma_x^2 \cdot \sigma_x^2 &= 1 = + + +, \\ \sigma_y^2 \cdot \sigma_y^2 &= - + - 2(+ -) = -A, \\ \sigma_z^2 \cdot \sigma_z^2 &= + + - 2(+ + -) = + + A, \\ \sigma_x^2 \cdot \sigma_y^2 &= - + - \end{aligned} \quad (12.15)$$

Look, we're doing this also with the first and last equations to

$$\begin{aligned} \sigma^x \cdot \sigma^y |+> &= \frac{i}{\hbar} (1 - 2) |+> + \frac{i}{\hbar} (-1) \\ \sigma^x \cdot \sigma^y |-> &= \frac{i}{\hbar} (-1 - 2) |-> + \frac{i}{\hbar} (-1) \end{aligned}$$

that they are quite similar. Now I invent a new operator, which I will call P_{exchange} , and which I ought to have the following properties:

$$\begin{aligned} P_{\text{exchange}} |+> &= |+> + |> \\ P_{\text{exchange}} |-> &= |-> + |> \\ P_{\text{exchange}} |+> &= |+> + |> \\ P_{\text{exchange}} |-> &= |-> + |>. \end{aligned}$$

All the operator does is interchange the spin directions of the two particles. Then I can write the whole set of equations in (12.13) as a single operator equation:

$$\sigma^x \cdot \sigma^y + 2P_{\text{exchange}} = 0. \quad (12.16)$$

(Don't the terms of $\sigma^x \cdot \sigma^y$ —the "spin exchange operator" gives a handy rule for labeling the states? You see, you can do everything now. The gates are open.)

12-3 The energy levels

Now we are ready to work out the energy levels of the ground state of hydrogen by solving the Hamiltonian equations (12.14). We want to find the energies of the stationary states. This means that we want to find those ψ 's, called ψ 's, for which each equation $H\psi = E\psi$ (ψ belongs to ψ) has the *exact* time dependence—namely, $e^{-iEt/\hbar}$. (I bet the state will have the energy $E = \hbar\omega_0$. So we want to set this, which

$$C_1 = a_1 e^{-iE_1 t/\hbar}, \quad (12.17)$$

where the other coefficients are still independent of time. To see whether we can get such simplitudes, we substitute (12.17) into Eq. (12.14) and see what happens. First multiply in Eq. (12.14) from left to right, and then cancel out the common exponential factor $e^{-iE_1 t/\hbar}$. (Remember why?) we get

$$\begin{aligned} Ee_1 &= Ae_1, \\ Ee_2 &= -Ae_2 - 2Ae_1, \\ Ce_3 &= 2Ce_2 - Ae_1, \\ Ee_4 &= Ae_4. \end{aligned} \quad (12.18)$$

which we have to solve for e_1, e_2, e_3 , and e_4 . Let's notice that the first equation is independent of e_2 too—(that means we can get this solution right away. If we choose $C = A$,

$$e_1 = 1, \quad e_2 = e_3 = e_4 = e_5 = 0$$

gives a solution. (If anyone taking all these's equal to zero gives a solution, but can't be since it all!) Let's call our first solution the state $|+\rangle$ †

$$|+\rangle = |+> + |+>. \quad (12.19)$$

Its energy is

$$E_1 = \hbar\omega_0.$$

† This operator is now called the "Pauli spin exchange operator."

↓ It is *not* really $1/\sqrt{12}$. We just say it's *close* to the state for the convenience of the numbers and $\hbar = 1$.

With two outgoing spin components, another solution from the last eq. is given in (12.18).

$$a_1 = a_2 = a_3 = 0, \quad a_4 = 1.$$

$$E = A$$

We'll see that this leads to $| E\rangle$:

$$| E \rangle = | A \rangle = | + - \rangle, \quad (12.20)$$

$$E_{\text{tot}} = A.$$

Now it gets a little harder. The two equations left in (12.18) are mixed up. But we've done it all before! Adding the two, we get:

$$E(a_1 + a_2) = A(a_1 + a_2). \quad (12.21)$$

Subtracting, we have:

$$(a_1 - a_2) = -A(a_1 - a_2). \quad (12.22)$$

By inspection and remembering symmetry—recall that there are two solutions:

$$\begin{aligned} \text{and} \quad & a_1 = a_2, \quad E = A \\ & a_2 = -a_1, \quad E = -A. \end{aligned} \quad (12.23)$$

They are mirror images of (2) and (3). Getting these states $| A \rangle$ and $| -A \rangle$, and putting in a factor $\sqrt{2}/2$ to make the states properly normalized, we have

$$\begin{aligned} | A \rangle &= \frac{1}{\sqrt{2}} (| + - \rangle + | - + \rangle) = \frac{1}{\sqrt{2}} (| + - \rangle + | - + \rangle), \\ | -A \rangle &= \frac{1}{\sqrt{2}} (| + - \rangle - | - + \rangle). \end{aligned} \quad (12.24)$$

$$E_{\text{tot}} = A$$

and

$$\begin{aligned} | A \rangle &= \frac{1}{\sqrt{2}} (| + - \rangle + | - + \rangle) = \frac{1}{\sqrt{2}} (| + - \rangle - | - + \rangle), \\ | -A \rangle &= -| A \rangle. \end{aligned} \quad (12.25)$$

$$E_{\text{tot}} = -A$$

We have found four stationary states and their energies. Hence, miraculously, that any four states are orthogonal, so they also can be used for basis states if desired. Our problem is completely solved.

Two of the states have the energy A , and the last two the energy $-A$. The average is zero, which means that, when we take $E_1 = 0$ in Eq. (12.5), we were measuring to measure all the energies from the average energy. We can draw the energy-level diagram for the ground state of hydrogen as shown in Fig. 12-2.

Now let's consider the energy between state $| A \rangle$ and any one of the others $\sim 4E$. An atom which happens to be just on the state $| A \rangle$ would fall from there to states $| B \rangle$ and emit light. Not optical light, because the energy is so big—it would emit a microwave quantum. Or, if we shine microwaves on hydrogen gas, we'd find an absorption of energy when an atom is state $| A \rangle$ with energy and we take one of the upper states—say only at the frequency $\omega = 4\pi/3$. The frequency has been measured experimentally; the best result obtained very recently is

$$\omega = \omega/2\pi = 11,430,406.771963 \pm 0.0001 \text{ cycles per second.} \quad (12.26)$$

The error is only one part in 10.5 billion! Probability is basic physical quantity is measured rather than time. It's one of the most remarkable accurate measurements in physics. The Russians were very happy that they could compute the energy to 12 decimal places in 1961, but in the meantime had been measured to 2 parts in 10^{11} —a million times more accurate than the theory. So the experimentalists are

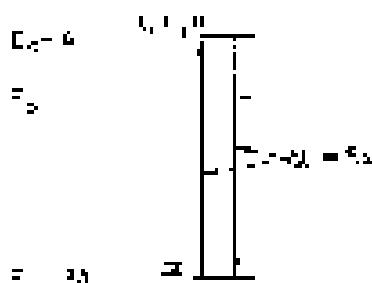


Fig. 12-2. Energy-level diagram for the ground state of atomic hydrogen.

way about 400 femtoseconds. In the theory of the ground state of the hydrogen atom you've seen pretty much everything. You, too, can just take your values out from experiment—that's what everybody does today in general.

You have probably heard before about the "Lyman-alpha" of hydrogen. That's the wavelength of the 121.6 nanometer spectral line between the 1s ground state and the 2s excited state. Radiation of this wavelength is emitted or absorbed by the atom of hydrogen gas in the jet device. So, with radio telescopes (radiation at 21 cm waves (or 1420 megacycles approximately); we can observe the reionization and the local ionization rate of atomic hydrogen gas. By measuring the intensity, we can determine the number of hydrogen. By measuring the frequency shift due to the Doppler effect, we can find out about the motion of the gas in the jet. This is one of the big advances of radio astronomy. So now we're talking about something that's very real—it is an observational program.

13.4 The Zeeman splitting

Although we have treated the problem of finding the energy levels of the hydrogen atom above, we would like to study this interesting system some more. In order to say anything more about it—for instance, in order to calculate the general width the hydrogen atom absorbs at various radio waves at 21 centimeters, we have to know what happens when the atom is disturbed. We have already provided for the atom moving in—say we lower the energy levels we want to find out what happens when the atom moves in a magnetic field. We can't calculate what happens when the atom is held in a radio wave. For the system you see in the magnetic field they're nothing to the levels except to move them off by some constant amount proportional to the square of the field—which is not the right thing because that won't change the energy difference. To do all the calculations that's important. So the next step is to write the Hamiltonian for a free atom in a situation in which the electrons in an external magnetic field.

What is it? Is the Hamiltonian? Well, just tell you the answer, because we can't give you any "secret" except to say that this is the way the atom works.

The Hamiltonian is

$$H = \frac{1}{2}m_e v^2 + p_{\phi}^2 + \mu_0 e \cdot B - \mu_0 \sigma \cdot B. \quad (122)$$

I have written it in three parts. The first term and v^2 represents the magnetic interaction between the electron and the proton. It is the same one that would be used if there were no magnetic field. This is the last term, however; the last and the influence of the magnetic field on the system. A is negligible. The effect of the nuclear magnetic field shows up in the last two terms. The second term, $-\mu_0 \sigma \cdot B$, is the energy the electron would have in the magnetic field if it were there alone.¹ In the same way, the last term, $\mu_0 \sigma^2 \cdot B$, would have been the energy of the proton alone. Classically, the energy of the two electrons together would be the sum of the two, and that works also quantum mechanically. In a magnetic field, the energy of the two electrons being magnetized is not the sum of the energy of interaction of the electron with its external field, since the presence of one field—but has to be expressed in terms of the dipole moments. In quantum mechanics these two don't add really the way they do, but the sum of the classical numbers is the energy. It's a way of doing things for verifying that the Hamiltonian works. Anyways, the current Hamiltonian is Eq. (122).

Now we are going back to the beginning and do the possible null observations. Much of the work is, however, done. We need only an additional term in the new form. Let's take a constant magnetic field B in the z-direction. Then we have to

¹ Remember that there is $-\mu_0 \sigma \cdot B$ to the energy of having when the moment is along the z-axis. The positive part of the magnetic moment is parallel to the spin and the negative part of it is opposite. So in Eq. (122) of the Hamiltonian there is a negative number.

Add to eq. (12.28) the second operator of the two new mass-scales; we call it \tilde{H}^2 :

$$\tilde{H}^2 = -(\mu_2 + \alpha_2)\delta.$$

Using Table 12-1, we get after simplification

$$\begin{aligned}\tilde{H}^2 | + - \rangle &= -(\mu_2 + \alpha_2)\delta, \\ \tilde{H}^2 | + + \rangle &= (\mu_2 + \alpha_2)\delta_1 = -\epsilon, \\ \tilde{H}^2 | - - \rangle &= -(-\alpha_2 + \lambda_2)\delta_1 = \epsilon, \\ \tilde{H}^2 | - + \rangle &= (\mu_2 + \alpha_2)\delta_1 = -\epsilon,\end{aligned}\quad (12.39)$$

thus very convenient! The \tilde{H}^2 operating on each state has given a value from each state. The matrix $(\tilde{H}^2)^{\dagger} \tilde{H}^2$ has, obviously, only diagonal elements; we can now add the coefficients in (12.29) to the corresponding diagonal terms of (12.13), and the Hamiltonian equation of (12.14) becomes

$$\begin{aligned}H(0, \beta) &= \{A + (\mu_2 + \alpha_2)\delta\}C_{11}, \\ H(0, \beta) &= \{A + (\mu_2 + \alpha_2)\delta\}C_2 + 2AC_3, \\ H(0, \beta) &= 2AC_3 + \{A + (\mu_2 + \alpha_2)\delta\}C_{22}, \\ H(0, \beta) &= \{A + (\mu_2 + \alpha_2)\delta\}C.\end{aligned}\quad (12.39)$$

The form of the equations is not quite arbitrary the coefficients. So long as δ doesn't vary with time, we can sum in just as we did before. Substitution, $C_2 = \alpha_2 C_1^{-1} C_{12}$, we get the modification of (12.38):

$$\begin{aligned}C_{11} &= A + (\mu_2 + \alpha_2)\delta\alpha_1, \\ E_{11} &= \{A + (\mu_2 + \alpha_2)\delta\}\alpha_2 = 2.1\alpha_2, \\ E_{22} &= 2\alpha_2 = \{A + (\mu_2 + \alpha_2)\delta\}\alpha_1, \\ E_C &= \{A + (\mu_2 + \alpha_2)\delta\}C.\end{aligned}\quad (12.40)$$

Equations (12.39) and fourth equations are β -independent of course so the same technique works again.

One solution to the above (12.40) when $\alpha_1 = 1, \alpha_2 = \alpha_3 = \alpha_4 = 0, \beta$

$$E = \{A + \beta\} | + + \rangle, \quad (12.41)$$

with

$$A = \beta = (\mu_2 + \alpha_2)\delta.$$

An other is

$$| D \rangle = | 0 \rangle = | + - \rangle,$$

with

$$E_D = A + (\mu_2 + \alpha_2)\delta. \quad (12.41)$$

A little more work is involved for the remaining two solutions because the coefficients of α_1 and α_2 are not proportional. But they are just like the previous case for the original α_1 and α_2 . Looking back at (12.30), we can make the following analogy (remembering that the levels I and II the α 's correspond to λ and β respectively):

$$\begin{aligned}H_{11} &= \beta + A = (\mu_2 + \alpha_2)\delta, \\ H_{12} &= 2\alpha_1, \\ H_{21} &= 2\alpha_1, \\ H_{22} &= \beta + A = (\mu_2 + \alpha_2)\delta.\end{aligned}\quad (12.42)$$

The energies are given by (12.27), which was

$$\epsilon_1 = \frac{H_{11}}{2} = \theta_{12} = \sqrt{\frac{4H_{11}}{3} + \frac{H_{22}(1 + H_{11}H_{22})}{3}}. \quad (12.43)$$

Making the substitutions from (12.7), the energy formula becomes

$$E = -A = \alpha(\mu_0 - \mu_0)B^2 + 4\mu_0.$$

Although the letters α are used to call the energies E_1 and E_{21} , and so on, in the problem calling them E_1 and E_{21} ,

$$\begin{aligned} E_{21} &= \alpha(1 + 2\sqrt{1 - \mu_0/\mu_0})B^2 + 4\mu_0, \\ E_{31} &= \alpha(1 + 2\sqrt{1 - \mu_0/\mu_0})B^2 + 4\mu_0. \end{aligned} \quad (12.8)$$

So we have found the energies of the four sub-energy states of a hydrogen atom in a constant magnetic field. Let's check on this by letting B go to zero, and seeing whether we get the same energies we find in the usual spectrum. You see that we do. For it is E_1 the energies E_1 , E_{21} , and E_{31} go to E_1 and E_{21} become $-4\mu_0$. From the labeling of the states you can see what we called them before. When we put in the magnetic field, though, all of the energy changes in a different way. Let's see how they do.

First, we have to remember that in the situation, μ_0 is negative, and that μ_0 is less negative than μ_0 , which is positive. So $\mu_0 + \mu_0$ and $\mu_0 - \mu_0$ are both negative numbers, and nearly equal. Let's call them μ_0 and $-\mu_0$.

$$a = -\mu_0 - \mu_0, \quad a' = (\mu_0 - \mu_0). \quad (12.9)$$

With a and a' two positive numbers, nearly equal to magnitude, one which is about one Bohr magneton, the energy levels are

$$\begin{aligned} E_1 &= 1 - \mu_0, \\ E_{21} &= 1 - \mu_0, \\ E_{31} &= 1 + 2\sqrt{1 - \mu_0/\mu_0}B^2, \\ E_{32} &= 1 + 2\sqrt{1 - \mu_0/\mu_0}B^2 + 4\mu_0. \end{aligned} \quad (12.17)$$

The energy E_1 starts at 1 and increases linearly with B , with the slope a . The

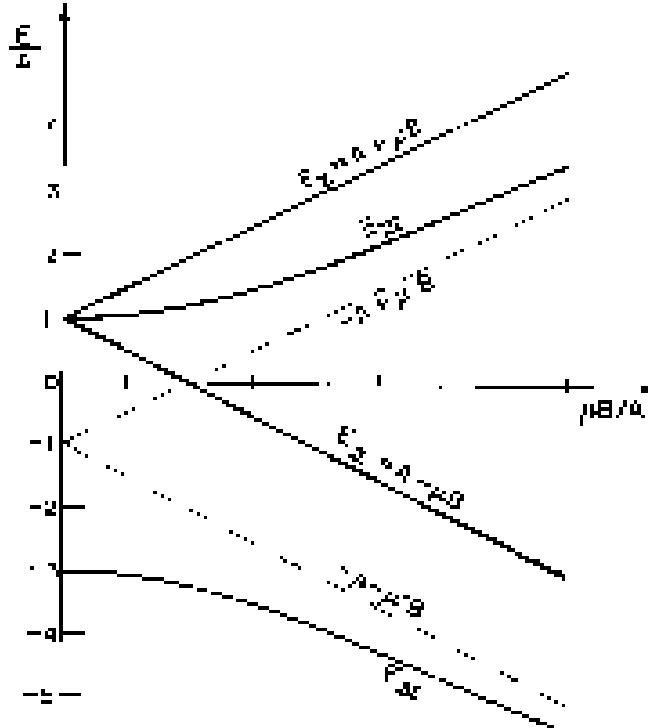


Fig. 12-1. The energy levels of the ground state of hydrogen in a magnetic field B .

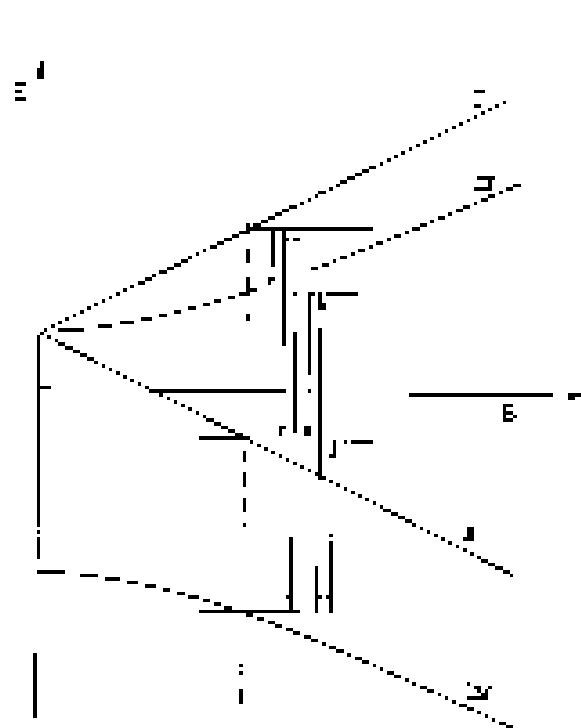


Fig. 12-2. Transitions between the levels of excited state energy levels of oxygen in a magnetic field B .

energy E_1 , which is at the lowest energy with increasing B . Its slope is $-g$. These two levels have with δ as above in Fig. 12-1. We show here in the figure the energies E_1 and E_0 . They have a linear B dependence. The small δ they depend on corresponds to 0 , or very near 0 , which is what we want. Then they begin to curve, and the slopes of the approach straight lines are slopes $-g$, which are exactly the slopes of E_1 and E_0 .

The shift of the energy levels due to a magnetic field is called the Zeeman effect. We can look at it in Fig. 12-1 below the $\psi = 0$ splitting in the ground state of hydrogen. When there is no magnetic field, we get just one ground line from the hyperfine structure of hydrogen. The interaction between spin and field by one of the electrons occurs here the same manner as in case of a photon whose frequency ω_0 is equal to $1/2$ times the energy difference ΔE . When the atom is in a magnetic field, it however has one new line. There are no transitions between only two of the four states, so if we calculate δ in each state, energy can be calculated as follows. It may see of the six transitions shown, by the vertical arrows in Fig. 12-4. Many of these transitions can be forced by the Rabi molecular beam techniques we learned in Section 12-1 (see Appendix).

We can do the transitions $2\psi_1 \rightarrow 1\psi_0$. The perturbing effect of disturbing ψ_1 is field that varies with time proportional to the steady strong field B_0 . Its job is to force a steady electric field on the hydrogen molecule. Only here, it is the magnetic field which couples with the molecular current and does not break. So the theory follows through in the same way that we worked it out for the atom. The ψ_1 is in the situation that a perturbing magnetic field δB comes in the ψ_1 plane. From this he can do oscillating field with δB . When you put this perturbing field as an oscillating term in the fluctuation, you get equations to solve. The amplitudes vary with time—just as it would for the atomic molecule. So you can calculate easily and accurately the probability of a transition from one state to another. And you find that it agrees with experiment.

12-5. The states in a magnetic field

We would like now to consider the energies of the curves in Fig. 12-3. In the first place, the energies in degeneracy are again relativistic and either in constant form or varying through the field. For $\mu_B = 1/2$ we can neglect the form the form of (12-3). This must change because

$$\begin{aligned} E_1 &= A + \mu B, & E_0 &= A - \mu B, \\ E_{0\perp} &= A + \mu' B & E_{1\perp} &= A - \mu' B \end{aligned} \quad (12-3)$$

This somewhat changes the bottom eight lines in Fig. 12-3. We can understand these energies physically in the following way. The nature of the stationary states in zero field is determined completely by the orientation of the two magnetic moments. The numbers of the last states $|+,-\rangle$ and $|-,+\rangle$ in the ordinary scales $|JM\rangle$ are A and $A - \mu B$ in interaction. In large enough field, however, the proton and electron m_e be influenced hardly at all by the effect of the other, and will just sit in one state in the external field. They $\rightarrow \psi_0$. There seem many lines—the plus spin will be down parallel to or opposite to the external magnetic field.

Suppose the electron spin is up— ψ_1 but ψ_0 may be field independent of $\pm \mu B$. The proton can still be either way. If the proton spin is also “up,” its energy is $A - \mu B$. The sum of the two is $A - \mu B - \mu B = A - 2\mu B$. That is just what we find for E_1 , which is the lowest energy occurring because $|+-\rangle = \psi_1$.

There is still the small additional term of $\delta \mu B$ in ψ_1 which represents the interaction energy of the proton and electron if their spins are parallel. We originally took δ as positive because, in ordinary we expect δ to be small, i.e., not too great. It is indeed so! On the other hand, the proton can have its spin down—then its energy is increased. If it is ψ_1 it gets $A - \mu B + \mu B = A$ and the energy $+ \mu B = \mu B^2 = \mu^2 B$. And the intermediate case of ψ_1 has $\delta \mu B$.

The sum neglects energy E_{ext} in (12.38). So the state $|M\rangle$ need not have zero because the state $= -\frac{1}{2}$.

Suppose now the electron spin is "down." Its energy in the external field is $\mu_B B$. If the proton is also "down," the two together have the energy $(m_1 + \mu_B)B = \mu_B$, and the interaction energy α —since their spins are parallel. That makes just the one for E_{int} in (12.38) and corresponds to the state $= + |M\rangle$, which is zero. Finally if the electron is "down" and the proton is "up," we get the energy $m_1 - \mu_B B = -\mu_B B$ for the interaction because the spins are opposite, which is just E_{ext} . And the state corresponds to $= - |M\rangle$.

"But wait a moment!" you are probably saying. "The states $|M\rangle$ and $|M'\rangle$ are NOT basis states $|+\rangle$ and $|-\rangle$; they are mixtures of the two. What's more, B is not $B = 0$, but we have not yet figured out what they are, for large B . When we used the results of (12.38) in the formulas of Chapter 9 to get the energies of the stationary states, we could substitute between the amplitudes that go with them. They come from Eq. (9.29), which is

$$\frac{C_1}{C_2} = \frac{E - E_{\text{ext}}}{\mu_B B}.$$

The ratio C_1/C_2 is of course, just $C_1 C_2^*$. Putting in the analogous quantities from (12.38), we get

$$\frac{C_1}{C_2} = \frac{E + A - (m_1 - \mu_B)B}{\gamma_A},$$

or

$$\frac{C_1}{C_2} = \frac{E + A + \mu_B B}{\gamma_A}. \quad (12.39)$$

What do E and A do to the hyperfine energy levels E_{HF} or E_{HF} ? For instance, for state $|+M\rangle$ we have

$$\left(\frac{C_1}{C_2}\right)_{+M} \sim e^{i\mu_B B} \quad (12.40)$$

So for large B the ground state $|+M\rangle$ (or $|-\rangle$) the atom has almost completely the state $|+M| + +\rangle$. Similarly, if we put E_{HF} zero (12.40) we get $(C_1/C_2)^*_{+M}$. $\propto e^{-i\mu_B B}$ for high-field state $|M'\rangle$ because just the state $|M| + +\rangle$ (you see that the coefficients in the linear combinations of our basis states which make up the stationary states depend on B). The state we call $|M'\rangle$ is a π -pseudostationary $= -\frac{1}{2}$ and $|-\rangle$, at very low fields, but shifts over to $= +\frac{1}{2}$ for large fields. Similarly, the state $|M'\rangle$ where the field is zero is at $\approx 90^\circ$ relative to the opposite energy $\propto |+\rangle$ (and $|-\rangle$), goes over into the state $|-\rangle$ when the spin is unoccupied by a strong external field.

We would also like to tell you about something particularly to what happens to your magnetic fields—there is one energy $= -14$ which does not change when you turn on a small magnetic field. And there is another energy $= -4$ which splits into three different energy levels when you turn on a small magnetic field. These two levels change very with B as shown in Fig. 12.5. Suppose you've somehow selected a bunch of hydrogen atoms which all have the energy $= -14$. If we put them through a Rabi-Clairaut experiment, with fields that are not too strong—you would find them just go straight through. (Since their energy doesn't depend on B , there is—according to the principle of virtual work—the force on them in a magnetic field is zero.) Suppose, on the other hand, we were to select a bunch of atoms with the energy $= -4$, and put the \perp through a Stern-Gerlach apparatus, say in X apparatus. (Again the field in the apparatus should not be so great that they disrupt the bunches of electrons, by which I mean a field going through the atoms that vary linearly with B .) We would find three bands. The states $|+M\rangle$ and $|M'\rangle$ get opposite forces—their energies vary linearly with B with the slopes $= \pm \alpha$. The others are like those of a dipole with $\alpha = 0$. But the state $|M\rangle$ goes straight through. So we are right back ... Chapter 5. A hydrogen atom with an energy $+4$ is a spinless particle. This energy isn't in a "pseudostate" for which $I = 1$, and it can be described with respect to some set of

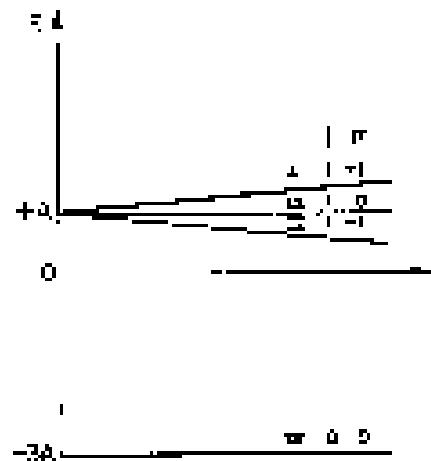


Fig. 12.5. Energy levels of the hydrogen atom for weak magnetic fields.

axes in space— $| +\text{S}\rangle$, $| -\text{S}\rangle$, $| +\text{S}\rangle$ and $| -\text{S}\rangle$ —and in Chapter 10, 3. On the other hand, since the hydrogen atom has the energy $E = \hbar^2/8\pi^2 r^3 m$ it represents a $\frac{1}{2}$ member which when applying $\hat{\mathbf{L}}_z$ will also have to be included in the basis. So we can project the states of hydrogen in zero magnetic field this way:

$$\begin{aligned} |\psi\rangle &= |\psi\rangle_{\text{S}} + |\psi\rangle_{\text{H}} \\ |\psi\rangle &= \left(\frac{1}{2} |\psi\rangle_{\text{S}} + \frac{1}{2} |\psi\rangle_{\text{H}} \right) + \left(\frac{1}{2} |\psi\rangle_{\text{S}} - \frac{1}{2} |\psi\rangle_{\text{H}} \right) \\ |\psi\rangle &= |\psi\rangle_{\text{S}} + \frac{1}{2} (\psi_{+} - \psi_{-}) \end{aligned} \quad (12.41)$$

$$|\psi\rangle = \frac{1}{2} (\psi_{+} + \psi_{-}) + \frac{1}{2} (\psi_{+} - \psi_{-}) \quad (12.42)$$

We have said in Chapter 10 of Vol. 1, II (Section 4) that for any particle its component of angular momentum along any axis can have only certain values, always \hbar apart. The component of angular momentum L_z will be $\hbar/2$ ($j = 1/2$) ($j = 1/2$, ..., $j = j/2$), where j is the spin of the particle (which can be an integer or half-integer). Although we neglected to say so, the same goes for the total angular momentum S .

Table 12.2
Zero-field states of the hydrogen atom

State $ m_S\rangle$	m_s	Orbit quantum no.
$ +\text{S}\rangle$	$+1/2$	1
$ 0\rangle$	0	2
$ -\text{S}\rangle$	$-1/2$	3
$ 0\rangle$	0	4

The m_s columns for one of the numbers $j = 1/2$, $j = 1/2$, ..., $j = j/2$. You will, therefore, see you're likely to find the $|+\text{S}\rangle$ ground state of $|y\rangle$ given by the vector $\hat{\mathbf{L}}_z$ (written below) and its partner called the “orthogonal” one (written vertically in front of it) and “angular momentum coupling” (see 10.4). Then, instead of our wave functions $|\psi_{+}\rangle$, $|\psi_{-}\rangle$ and $|\psi_{0}\rangle$, they will be $|\psi_{+}\rangle$, $|\psi_{-}\rangle$ and $|\psi_{0}\rangle$ which would give one linear superposition in zero field (12.41) and (12.42) as shown in Table 12.2. It's not very precise, it's all just a matter of notation.

12.6 The projecting matrix for spin one

We would like now to use our knowledge of the hydrogen atom in the zero-field case. We discussed in Chapter 5 that the state of the electron which has $j = 1/2$ is one of the two states $|+\text{S}\rangle$, $|-\text{S}\rangle$ with respect to the normalized operator $\hat{\mathbf{L}}_z$ and the $|+\text{S}\rangle$ state $|+\text{S}\rangle$ corresponds to $\hat{\mathbf{L}}_z = \hbar/2$ up to an amplitude. In both cases the $|+\text{S}\rangle$ state will be $|\psi_{+}\rangle$ with respect to the $\hat{\mathbf{L}}_z$ operator in space. There are three such amplitudes $|\psi_{+}\rangle$ which make up the projection matrix. In Section 10.4 we gave the form of this matrix for various orientations of $\hat{\mathbf{L}}$ with respect to $\hat{\mathbf{L}}_z$. Now we will see why, how and they can be derived.

In the hydrogen atom we have found a spin-one system which is made up of two spin-one-half particles. We have already worked out in Chap. 10 how to transform a spin-one-half impurity. Now we use this information to calculate the transformation for both one. The easiest way is worse. We have a system of a hydrogen atom with the energy E_{H} which has spin $j = 1/2$. Suppose we want it enough to transform it. From 5. we see that we know it is in one of the basis states with respect to $\hat{\mathbf{L}}_z$, say $|+\text{S}\rangle$. There is one amplitude a which will be in one of the basis states $|+\text{S}\rangle$ with respect to the $\hat{\mathbf{L}}_z$ operator. If we use the system of the S operator in the $|+\text{S}\rangle$ orientation, the $|+\text{S}\rangle$ state is $|\psi_{+}\rangle$ which we have been calling the state $= +\text{S}$ —the S spin number one, back his head along the axis of $\hat{\mathbf{L}}_z$. He will be selecting his states to whatever will be the S spin number. His “up” are “down” states for S electron and proton would be different from ours. His “parallel” they—which we can write $|+\text{S}\rangle$ —differ in the “up” S sense, i.e. $|+\text{S}\rangle$ state of the spin one particle. What we want is $|+\text{S}\rangle$ which is just another way of writing the amplitude $(|+\text{S}\rangle - |-\text{S}\rangle)$.

*This section does not impress Chapter 10, so I skip this section also.
12.11

We could do the superposition $|+\rangle = |+\rangle + |-\rangle$ in the following way. In our frame, the electron (the $|+\rangle$) state has spin “up”, \uparrow . This means that it has some amplitude $\langle +|+\rangle$ of being “up” in the frame, and some amplitude $\langle +|-\rangle$ of being “down” in the frame. Similarly, the proton $= |\psi_1\rangle + |\psi_2\rangle$ has spin “up” in our frame and the components $= |\psi_1\rangle$, and $= |\psi_2\rangle$ have spin “up” or spin “down” in the “prime” frame. So, as we are talking about the total particle, the amplitudes of each particle will be “up” together in the frame, in the product of the two amplitudes:

$$\langle +|\psi\rangle = \langle +| +\rangle + \langle +| -\rangle = \rho^2. \quad (12.45)$$

You may put the subscripts a and b to the amplitudes $\langle +| \psi_a \rangle$ to make it clear what we were doing – that they are both individual amplitudes for a single final particle, so they are really independent of each other. They are, in fact, just the amplitudes we have calculated $\langle +| \psi_a \rangle$ in Chapter 8, and which we used in the figures at the end of this chapter.

Now, however, we are going to go into much more detail. We have to handle all possible products of amplitudes $\langle +| \psi_a \rangle$, for a spin-one-half particle from which we have also called $= |\psi\rangle$. It is a reasonable idea, if they are completely独立的, that there should be three coefficients, α , β , γ , for the overall amplitude, as well as two for each different $\langle +| \psi_a \rangle$ for the two one-half amplitudes. To help you keep things straight, we can make the new notation in Tab. 12.4. We will continue to use the definition $= |\psi\rangle = |\psi_1\rangle + |\psi_2\rangle$, just as for the standard spin-one-half particles.

With our new notation, Eq. (12.45) looks as simple

$$= \langle +|\psi\rangle = \alpha + \beta + \gamma,$$

and this is just the spin-one amplitude $= |\psi\rangle$. Now, it's apparent, for example, that the other two components of “down” are the β , γ “spin-down” amplitudes, β , γ just related with respect to the α axis by the angle θ ; then from Fig. 6-7

$$\alpha = \langle +| +\rangle = \rho^2 \cos \theta$$

So, from (12.4) we know that the symmetric amplitude is

$$\langle +|T_{+-}|S\rangle = \langle +| +\rangle \langle -| -\rangle = (\rho^2)^2 = \rho^4. \quad (12.46)$$

You can see how it goes.

Now we will work through the general case for all the states. If the proton in our frame had “up” in the frame, the Adjoint, the amplitudes that it must be in my frame (the two possibilities) in another frame (the T4 frame) are

$$\begin{aligned} & \langle +| +\rangle + \langle +| -\rangle = (\rho^2 + \rho^2 \sin^2 \theta) \langle +| +\rangle = \rho^2, \\ & \langle +| +\rangle + \langle -| +\rangle = \langle +| -\rangle = \rho^2 \langle -| -\rangle, \quad \text{etc.} \\ & \langle +| +\rangle = \langle +| -\rangle = \langle -| +\rangle = \langle -| -\rangle = \rho^2, \\ & \langle +| +\rangle = \langle -| -\rangle = \langle +| -\rangle + \langle -| +\rangle = \rho^2. \end{aligned} \quad (12.47)$$

We can, then, write the state $= |\psi\rangle$ as the following linear combination:

$$|+\rangle = \rho^2 |+\rangle + \beta |+\rangle - \gamma |+\rangle + \gamma |-\rangle + \beta |-\rangle - \rho^2 |-\rangle. \quad (12.48)$$

Now we notice that $= |\psi\rangle$ is the state $= |\psi\rangle$ that $= |\psi\rangle = |\psi\rangle$ is just $\sqrt{2}$ times the state $= |\psi\rangle$ (see (12.4)) and that $= |\psi\rangle = |\psi\rangle$. In other words, Eq. (12.48) can be rewritten as

$$|+\psi\rangle = \rho^2 |+\rangle + \sqrt{2} \beta |\psi\rangle + \sqrt{2} \gamma |-\rangle. \quad (12.49)$$

In a similar way we can easily show that

$$|\psi\rangle = \rho^2 |+\rangle + \sqrt{2} \alpha |\psi\rangle + \sqrt{2} \beta |-\rangle. \quad (12.50)$$

Table 12.4

Spin-one-half amplitudes

One-chopper	Composite
$\alpha = \langle + +\rangle$	$= T_{+-} S\rangle$
$\beta = \langle + -\rangle$	$= T_{+-} S\rangle$
$\gamma = \langle - +\rangle$	$= T_{+-} S\rangle$
$\delta = \langle - -\rangle$	$= T_{+-} S\rangle$

for $|+\rangle$ it's a little more complicated, but we

$$|+\rangle_S = \frac{1}{\sqrt{2}}(|+\rangle_1 + |-\rangle_1)$$

But we can express each of the states $|+\rangle_1$, $|\pm\rangle_1$, and $|-\rangle_1$ in terms of the "naïve" states and take the sum. That is,

$$|+\rangle_1 = \omega(|+\rangle^+ - i|\mp\rangle) \quad \text{and} \quad |\pm\rangle_1 = \Re(\omega^{-1} + i\Im(\omega^{-1})) \quad (12.30)$$

and

$$|-\rangle_1 = \omega(|-\rangle^+ - i|\mp\rangle) \quad \text{and} \quad |+\rangle_1 = \Re(\omega^{-1} + i\Im(\omega^{-1})) \quad (12.31)$$

Taking (12.28) times the sum, we get

$$\langle 0|S_1 = \frac{2}{\sqrt{2}}\omega(\omega_1 + \omega_2 - \frac{\omega_1 + i\omega_2}{\sqrt{2}})(|+\rangle^+ - i|\mp\rangle) + \frac{2}{\sqrt{2}}\omega(\omega_1 - \omega_2)$$

It follows that

$$\langle 0|S_1 = \sqrt{2}\omega(\omega_1 + \omega_2 - \omega_1^2 + \omega_2^2) + \sqrt{2}\omega(\omega_1 - \omega_2) \quad (12.32)$$

We have now all of the spin-coordinates wanted. The coefficients of Eqs. (12.28), (12.30), and (12.31) are the matrix elements $\langle Q^a | S_b | 0 \rangle$. Let's put them all together.

$$\langle Q^a | S_b | 0 \rangle = \begin{pmatrix} \omega_1^2 & \omega\sqrt{\omega_1\omega_2} & \omega^2 \\ \sqrt{2}\omega_1 & \omega_1^2 + \omega_2^2 & \sqrt{2}\omega\omega_2 \\ \omega_2^2 & \sqrt{2}\omega\omega_1 & \omega^2 \end{pmatrix} \quad (12.33)$$

We have expressed the spin- a transformation in terms of the spin- b coordinates a , b , c , and d .

The last piece of this S matrix is nothing with respect to b by the angle α , since the result is α . Fig. 5.5. The amplitudes in Table 12.1 are just the matrix elements of $S_a(b)$ in Table 6.2.

$$\begin{aligned} a &= 60^\circ \frac{\pi}{3}, & b &= -60^\circ \frac{\pi}{3}, \\ c &= 30^\circ \frac{\pi}{3}, & d &= -60^\circ \frac{\pi}{3}. \end{aligned} \quad (12.34)$$

Using these in (12.33), we get the form of (12.35), which we give there without proof.

What ever happened to the state $|1222\rangle$? Well, it's a spin-zero system, so it has only one character in the name of all coordinate systems. We can check that everything works out by taking the difference of Eq. (12.30) and (12.31); we get that

$$|+\rangle_1 - |-\rangle_1 = |+\rangle^+ - i|\mp\rangle = (\omega_1^2 - \omega_2^2)|+\rangle + (\omega_1 - \omega_2)|-\rangle.$$

But this is $\omega\omega_1$, the determinant of the spin matrix from (12.28), and so is equal to 1. We see then

$$|\langle 0|S_1 | 1222 \rangle| = |\langle 0|S_1 | 0 \rangle|$$

so the relative orientation of the two coordinate frames.

Propagation in a Crystal Lattice

13-1 States for an Electron in a one-dimensional lattice

You will at first sight think that a low-energy electron would have great difficulty passing through a solid crystal. The atoms are packed together with their outer shells a few angstroms apart, and the distance outside of the outer shell electrons is hardly an angstrom or so. That is, the electrons being bound to their spring, so that you would expect the mean free path between collisions to be of the order of a few angstroms which is practically nothing. You could expect the electron to bump into one atom or another almost immediately. Nevertheless, it is a surprising phenomenon of course that if the lattice is perfect, the electrons are able to travel through. The crystal atoms are easily disturbed; they are in a way like this string which lets me demonstrate electricity so easily; it was also important the development of many practical devices. It is also rather nice that it is possible to manufacture to imitate the lattice also. In a radio tube electrons move freely through a vacuum while in the anode they move freely through a coaxed lattice. The machinery behind the lattice is transistor will be described in this chapter. One day we will describe the application of these principles in various electronic devices.

The conduction of electrons in a crystal is one example of a very common phenomenon. Not only do electrons travel through crystals, but other "things" like current carriers can also travel in a solid medium. So the phenomenon which we are studying appears in many ways in the study of the physics of the solid state.

You will remember that we have discussed many and a lot of discrete vibrations. Let's now think of an electron which can be in either one of two positions in each of which it is in the same kind of environment. Let's also suppose that there is a certain amplitude to go from one position to the other, and of course the same amplitude to go back just as we have discussed for the hydrogen molecule in Section 10-1. The laws of quantum mechanics then give the coupling condition. There are two possible states of definite energy for the electron. But it can be described by the amplitude for the electron to be in each of the two basis functions. In either of the definite-energy states, the magnitudes of these two amplitudes are conserved in time, and the forces vary in time with the same frequency. On the other hand, if we start the electron in a given form, it will have been mixed to another, and so on, so having back again to the first position. This amplitude is analogous to the motion of two coupled pendulums.

Now consider a perfect crystal lattice in which we imagine that an electron can be in one of a kind of "sites" in a regular pattern and with some potential energy. Suppose also that the electron has some amplitude to move into a different potential of the nearby atoms. It's something like the two-sites system, but with an additional complication. When he starts on a site in the neighborhood, he can afterward move on to the nearest position as well as return to his starting point. Now we have a situation analogous not to two coupled pendulums, but to an infinite number of pendulums all coupled together. It's something like what you see in one of those machines—made with a long row of tops mounted on a central wire—that is used in first year physics to demonstrate wave propagation.

If you take a common needle which is coupled to another harmonic oscillator and that one has some initial phase, and if you start in regularity in one place, the regularity will propagate as a wave along the line. The same situation occurs if you place random atoms in one of a long chain of atoms.

13-1 States for an electron in a one-dimensional lattice

13-2 States of definite energy

13-3 Time-dependent states

13-4 An electron in a three-dimensional lattice

13-5 Other states in a lattice

13-6 Scattering by imperfections in the lattice

13-7 Trapping by a lattice imperfection

13-8 Scattering amplitude and bound states

Usually the simplest way of analyzing the technical problem is not to think in terms of what happens if a particle is started at a certain place, but rather in terms of steady-state solutions. There exist certain solutions which propagate through the crystal as a wave of a single fixed frequency. Now the same thing happens with the electron—and for the same reason, because it's described in quantum mechanics by similar equations.

You can appreciate one thing, however. The conditions for the electron to leak a place is unambiguous, not a probability. If the electron were simply falling from one place to another, like water going through a hole, the behavior would be completely different. For example, if we had two levels of state connected by a tube, a general state leakage from one to the other, even the levels would approach each other exponentially. But for the electron, what happens is a definite leakage and not just a plain probability leakage. And it's a characteristic of the quantum theory that it is the differential equation that quantizes the theory—which changes the differential equation to an oscilloscope screen. What happens then is quite different from the leakage between interconnected tanks.

We want now to analyze quantitatively the quantum mechanical situation. Imagine a one-dimensional system made of a long line of atoms as shown in Fig. 13-1(a). (A crystal is of course three-dimensional but the picture is very much the same; just you understand the three-dimensional case; you will be able to understand what happens in three dimensions.) Next, we want to see what happens if we put a single electron on this line of atoms. Of course, in a real crystal there are already millions of electrons. But some of them probably fall far off in insulating crystals, take up positions in some part of the crystal which are bound and everything is quite stationary. However, we are going to this account what happens if we put one extra electron in. We will not consider what the other electrons are doing because we're going to change their motion very little by the addition of energy. We are going to add an electron to 2 to produce one energy bound negative ion. In addition, we'll use extension to 3 to make an approximation which corresponds approximately to the inside workings of the atom.

What does the electron want then to do in another atom, instead of the negative ion it's in its place? We will suppose that just as in the case of an electron jumping between two atoms, the electron can jump from one atom to the neighbor on either side with a certain amplitude.

Now how do we do the nuclei system? What will be reasonably reasonable? If you remember what we did when we had only two possible positions, you can guess how it will go. Suppose that we can put of 2, so the vacancies are all equal, and that we number the atoms in sequence, as shown in Fig. 13-1(b). One of the possibilities is that the electron is at atom number 6, another possibility is that the electron is at atom number 7, or at atom number 8, and so on. We can describe the electron state by saying that the electron is at atom number 6. Let's say that this is the wave state $|6\rangle$. Figure 13-1 shows what we mean by the three basic states

$$|0\rangle = |1\rangle, \quad |5\rangle, \quad \text{and} \quad |4 + 6\rangle.$$

Using these base states, any state $|\psi\rangle$ of our one-dimensional crystal can be described by adding all the amplitude $a_i |\psi\rangle$'s. If the state $|\psi\rangle$ is in one of the two states—which means the amplitude a_i is located at the particular atom. Then we can write the state $|\psi\rangle$ as a superposition of the base states

$$|\psi\rangle = \sum_i a_i |\psi_i\rangle. \quad (13-1)$$

Now, we are going to suppose that when the electron is at one atom, there is some amplitude that it can leak to the atom on either side. And we'll take the simplest case for which it can only leak to the nearest neighbors. Imagine the two nearest neighbors, it has to go in two steps. Well, then if the amplitude for the electron just move next door to the next is a_1 open that means

For the moment we would like to write the amplitude ψ (4) to be on the electron as C_n . Then Eq. (13.1) will be written

$$\psi = \sum_n e^{iE_n t} C_n. \quad (13.2)$$

If we know each of the amplitudes C_n at a given moment, we could take their absolute squares and get the probability that you would find the electron if you looked it up at that time.

Now on the successive readouts from time t by analogy with the parallel system we have studied, we would propose (i.e. the Hamiltonian equations) that the system should be made of equations like this:

$$i\hbar \frac{dC_n(t)}{dt} = E_n C_n(t) - AC_{n-1}(t) - AC_{n+1}(t). \quad (13.3)$$

The last coefficient on the right, A , is, physically, the energy the electron would have if it wouldn't leak away from one of the atoms. (It doesn't matter what we call E_F , as we have seen many times, it represents really nothing but a choice of the zero of energy.) The next term represents the amplitude per unit time that the electron is leaking into the next position (i.e. $n+1$) atom; and the final term is the amplitude for leaking from the $(n-1)$ st atom. As usual, we'll ignore this as a constant (independent of t).

For a full description of the behavior of any state $|n\rangle$, we would have one equation like (13.3) for every one of the amplitudes C_n . But if we want to consider a crystal with a very large number of atoms, we'll assume that there are an indefinitely large number of states (that the atom is going forever in both directions). (For the first case we will have to pay special attention to the exponential terms); if the number N of our wavefunctions is indefinitely large, then obviously the Hamiltonian equations are infinite in number. We'll write down just a sample:

$$\begin{aligned} i\hbar \frac{dC_0}{dt} &= E_0 C_0 - AC_{-1} - AC_1, \\ i\hbar \frac{dC_1}{dt} &= E_1 C_1 - AC_{0,-1} - AC_{2,+1}, \\ i\hbar \frac{dC_2}{dt} &= E_2 C_2 - AC_{1,-1} - AC_{3,+1}. \end{aligned} \quad (13.4)$$

12.2 States of definite charge

We can't say many things about an electron in a lattice but first let's try to find the states of definite energy. As we have seen in earlier chapters this means that we have to find a situation in which the amplitudes all of large at the same frequency if they change at all. We look for such kinds of functions:

$$C_n = c_n e^{-E_n t}. \quad (13.5)$$

The complex numbers c_n will be then the non-time-varying part of the amplitude for the electron in the n th atom. If we put this last solution into the equations (13.4) to test them out, we get the result:

$$c_{n+1} - c_n c_n = -Ac_{n+1} - Ac_{n-1}. \quad (13.6)$$

We have 2.1 infinite number of such equations for the infinite number of unknowns c_n , which is rather gratifying.

All we have to do is take the determinant... but wait! Determinants are the ones that are 2, 3, or n equations. But if there are a lot more than n equations and less than n unknowns, the determinant is not so very meaningful. Well, before you try to prove the equations directly, there are 3 ways to do this:

Similarly, we'll see that the variation is ω_{n+1} if the atom ($n+1$) is at x_{n+1} . If our starting energy is $E_0 = \hbar\omega_0$, we will have that $\omega_{n+1} = \omega_0 + \delta$. By choosing our δ big at about 25%, we can expect now that $\omega_n = \hbar\omega$. We can rewrite Eq. (13.3) to

$$C_n = \alpha(x_n)e^{-iE_n t}, \quad (13.7)$$

and Eq. (13.6) would become

$$\dot{E}_0(x_0) = E_0(x_{n-1}) - i\partial(x_{n-1}) \approx i\partial(x_{n-1}). \quad (13.8)$$

Or, using the fact that $x_0 = \dots = x_n = b$, we might also write

$$\dot{E}_0(x_0) = E_0(x_0) - i\partial(x_0 + \delta) \approx i\partial(x_0 + \delta). \quad (13.9)$$

This equation is somewhat similar to a differential equation. It tells us the ω quantity, $i\partial(x)$, at one point, (x_0) , is related to the same physical quantity at next neighbor inputs, $(x_0 + \delta)$. (A differential equation relates the value of a function at a point to the values at infinitesimally nearby points.) Perhaps, the methods we usually use for solving differential equations will also work here. Let's try.

Linear differential equations with constant coefficients can always be solved in terms of exponential functions. We can try the same thing here to get a CR solution:

$$i\partial(x) = e^{i\omega x}. \quad (13.10)$$

Then Eq. (13.9) becomes

$$E_0e^{i\omega x} = E_0e^{i\omega x} - i\partial e^{i\omega x} \approx -i\partial e^{i\omega x}. \quad (13.11)$$

We can now divide out the common factor $e^{i\omega x}$; we get

$$E_0 - E_0 = i\partial e^{i\omega x} = ie^{i\omega x}. \quad (13.12)$$

The last two terms are just equal to (3.1) or (3.9), so

$$E_0 - E_0 = 2ie^{i\omega x}/\hbar. \quad (13.13)$$

We have found that for any value of ω (or the constant \hbar) there is a solution whose energy is given by this equation. There are various possible constant energy inputs, and each corresponds to a different solution. There are an infinite number of solutions. This is not surprising, since we started our orbit in the middle of base states.

Let's see what these solutions look like. For each ω , the ψ 's are given by Eq. (13.10). The amplitudes C_n are then given by

$$C_n = e^{i\omega x_n} e^{-iE_n t}. \quad (13.14)$$

Since you should remember that the energy E also depends on ω as given in Eq. (13.13), the space dependence of the amplitude is $e^{i\omega x}$. The ω dependence is $e^{-iE_n t}$, which is zero when t is even or the rest.

We might note, in space, the amplitude goes as a complex oscillation—the wavelength is the same at every atom, but the phase at a given site is chosen by its ω and t (odd). You can see this in the next. We can visualize what's going on by placing a vertical line to show just the real part at each atom as we have done in Fig. 13.2. The envelope of these vertical line functions is the broken line called

$I_{\text{Re}(\psi)}$

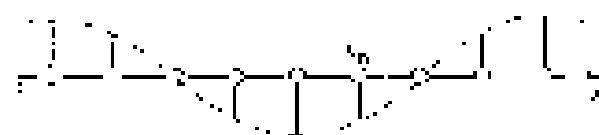


Fig. 13.2. Variation of the real part $\psi(x)$ with x .

is, of course, a cosine curve. The imaginary part of the Naderman oscillating function, however, is shifted 90° in phase so that the absolute square value is the sum of the squares of the real and imaginary parts; this is the state for \pm the C's.

Thus if we pass $\pm k$, we get a stationary state of a particle of energy E . And in any such state, the electron is equally likely to be found at every atom, since there is no preference for location, or the other. Only the phase is different for different atoms. Also, we can pass on the particle wave. From Eq. (13.14) the real and imaginary parts propagate along the crystal as waves—namely, as the real or imaginary parts of

$$\psi(x, t) = e^{iE_0 t} \psi_0(x) \quad (13.25)$$

The wave can travel toward $x < 0$ like a negative k depending on the sign we have picked for k .

Notice that we have been assuming that the number k that we put in our trial solution, Eq. (13.13), was a real number. We can see now why this must be so if we have an infinite line of atoms. Suppose that k were an imaginary number, say iK . Then the amplitude ψ_0 would grow as e^{iKx} , which means that the amplitude would get larger and larger as we go toward large x 's—or toward large negative x 's if K is a negative number. This kind of solution would be O.K. if we were dealing with one or a few atoms, but quickly it would be a physical solution for an infinite chain of atoms. It would give infinite amplitudes—and, therefore, infinite probabilities—which isn't agreement with situation. Later on we will see an example in which the ψ gives it does not do so.

The relation between the energy E and the wave number k is given in Eq. (13.14) & plotted in Fig. 13-3. As you can see from the figure, the energy can go down ($E_1 = -2\pi/\lambda$) & up ($E_2 = 0$) in $(x_0 + \lambda/2)$ at $k = \pi/\lambda$. The graph is plotted for positive k ; if k were negative, the curve would simply be inverted, but the range would be the same. The significant result is that any energy is possible within a certain range or "band" of energies, but no others. According to our assumptions, if an electron in a crystal is in a stationary state, it can have an energy E lower than values in this band.

According to Eq. (13.10), the condition E is not imposed to two-energy states E in (8), (24). As k increases in magnitude (toward either positive or negative k values) the energy at first increases, but then reaches a minimum at $k = -\pi/\lambda$, as shown in Fig. 13-4. For k 's larger than π/λ , the energy would start to increase again. But we do not really need to consider such values of k , because they do not give new states. One just repeat states we already have for smaller k . We can see that in the following way. Consider the lowest energy state for which $k = 0$. The coefficient $a_{k=0}$ is the same for all x_0 . Now we could get the same energy for $k = 2\pi/\lambda$. But then, using Eq. (13.10), we note that

$$a(4,0) = e^{iE_0 x_0},$$

However, taking x_0 to be at the origin, we consider $x_0 = \text{any multiple of } \lambda$, because

$$a(4,0) = e^{iE_0 x_0} = 1$$

The state described by these $a_{k=0}$'s is precisely the same state we got for $k = 0$. It does not represent a different solution.

As another example, suppose that k were real. The real part of $a_{k=0}$ would vary as shown by curve 1 in Fig. 13-4. If k were even times larger ($k = 2\pi/\lambda$), the real part of $a_{k=0}$ would vary as shown by curve 2 in the figure. (The complete

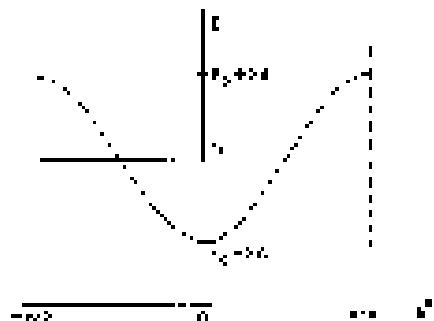


Fig. 13-3. The energy of the stationary wave as a function of the parameter k .

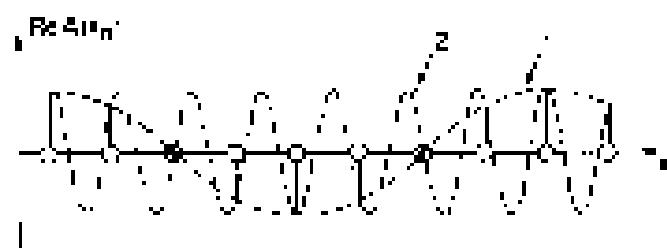


Fig. 13-4. Two values of k which represent the same physical situation, curve 1 is for $k = \pi/\lambda$, curve 2 is for $k = 2\pi/\lambda$.

possible curves don't start anything of course, all that makes ψ the values of the constants. This means we just want to know the energy E_{final} (just). You should note values of E_{final} become symmetric around the E_{initial} 's.

The conclusion that we have an ψ produce what amounts to a problem of calculating E_{final} in a certain limited range. It's not the range between $-E_{\text{initial}}$ and $+E_{\text{initial}}$ —the one shown in Fig. 13-3. In this range, the energy of the stationary states increases linearly with the amplitude L in magnitude of ψ .

One side remark when you're doing your own play with ψ —Suppose that the electron cannot only jump to the nearest neighbor wave amplitude value, but also has the possibility to jump to the next next neighbor with some other amplitudes E_{final} . You will find that the solution of ψ again has written in the form $\psi = \psi_0 + \psi_1 + \psi_2 + \dots$ type of solution, only more complex. You will also find that the stationary states with wave number k have an energy equal to $(E_k - E_{\text{initial}})$ (Prob. 13-1). The solution for the value of the constant E_{final} is no longer unique, it depends upon the particular description of the problem. It is not necessarily the case where E_{final} is not necessarily the same about some individual line. It is true, however, that the new always appears "on" themselves, the interval from $-E_{\text{initial}}$ to $+E_{\text{initial}}$ you never need to worry about other values of ψ .

Let's look a little more closely at what happens for small L , that is when the amplitude of the amplitude from our ψ is the first approximation. Suppose we increase our sum of energies by decreasing E_{initial} . But that is the situation of Fig. 13-2 up to the zero of energy. But just enough, we can write the

$$E_{\text{final}} \approx E_{\text{initial}} + L^2/2,$$

and the energy in Eq. (13-1) becomes

$$E = E_{\text{initial}} + L^2/2. \quad (13-6)$$

We have then the energy of the state is proportional to the square of the wave vector which describes the spatial character of the amplitude L .

13-3 Time-dependent states

In this section we would like to discuss the properties of states in the one-dimensional lattice in more detail. If the amplitude for an electron to be at x is C_x , the probability of finding it there is $|C_x|^2$. For the elementary state described by Eq. (13-1) this probability is very close to 1, and does not change with time. Now let us represent a situation which we would consider as unstable by taking an electron in a binding energy E_0 and then in a certain system ω then it is much likely to be found in one place longer or some other place. We can do that by making a superposition of several solutions like Eq. (13-12) with slightly different values of L and therefore of slightly different energies. Even if $L = 0$ at least the amplitude C_0 will vary with position because of the interference between the various terms, just as the probabilities where there is a mixture of waves of different wave lengths was discussed in Chapter 4E. Very likely some wave mode will have a minimum wave number k_0 , and with various k between this wave number, k_0 is a wave number of stationary states, the amplitudes with different k 's will represent states of slightly different energies, and, the others, of slightly different frequencies. Let's take for convenience of the initial C_0 with the same phase as with $k = k_0$ and $k = k_0 + 1$ (that is, two of them). As we have seen in the writing of column 1, the peaks of the peaks [the peaks where $C(k)$ is largest] will move along in a zig-zag pattern, associated with the quant we have called the "pinning centers." We found that this group of peaks moves with the wave speed c with frequency

$$\omega_{\text{peak}} = \frac{\Delta E}{\hbar}. \quad (13-7)$$

⁴ Provided we do not try to make the period too long.

The same series for would apply equally well here. An electron's energy is a "hump" number one less than for ψ_0 , varying in space just the way ψ_0 does. Fig. 13-5 will show the one-dimensional "transition" with the ψ_0 and ψ_1 in the Ψ , where $\Psi = \psi_0 + \psi_1$. Then if E_0 is for ψ_0 , we get for Ψ

$$\epsilon = \frac{2\pi^2}{\lambda} \psi_0. \quad (13.8)$$

In other words, the ψ_0 terms cancel except at spots where ψ_0 is the right ψ_0 . So since (13.6) tells us that the energy of such an electron is proportional to the square of its velocity, it now looks like a classical particle. So long as we look at things on a scale going enough that we don't see the hump effect, our quantum mechanical picture begins to give results like classical physics. In fact, if we solve Eq. (13.4) by for λ and substitute into (13.6) we can write

$$E = \frac{h^2}{2m} k^2, \quad (13.9)$$

which is equivalent to the extra "kinetic energy" of the electron in a pocket depends on the value k just as for a classical particle. The constant m will be the "effective mass" as given by

$$m^* = \frac{h^2}{2E_0}. \quad (13.10)$$

Also notice that we can write

$$m^* c^2 = E_0. \quad (13.11)$$

If we choose to call m^* the "momentum," it is consistent with our notation, although in the way we have described earlier for a free particle.

Our thought experiment has nothing to do with the real mass of an electron. Energy is quite little time, and the light in real crystals doesn't keep up so long, probably the same order of magnitude, about 1 to 20 times the free space mass of an electron.

We have now established a remarkable property—how an electron in a crystal (or an electron in a vacuum) can move like light through the crystal and flow perfectly freely over the crystal lattice in all the directions. It does so by forming an n -dimensional grid of points in the x -directions x_1, x_2, \dots, x_n , where n is the dimension of the crystal. That is how a solid can conduct electricity.

13-4 An electron in a three-dimensional lattice

Let's look for a moment at how we could apply this same idea to see what happens to an electron in three dimensions. The results turn out to be very similar. Suppose we have a rectangular lattice of atoms with lattice constants a_1, a_2, a_3 in the three directions. (If you've seen a cubic lattice, take the face spacing $a_1 = a_2 = a_3$.) We suppose that the n -component of the wavefunction is ψ_{x_1, x_2, x_3} , where x_1, x_2, x_3 are in the x -directions (x_1, x_2, x_3) and ψ is zero in the x -direction (y_1, y_2, y_3) . Now how shall we describe the basis states? As in the one-dimensional case, the base state is that the electron is at the spot where $x_1 = x_2 = x_3 = 0$, when $\psi(x_1, x_2, x_3)$ is zero at the other points. Choose x_1 with $x_1 = 0$, then there should be a ψ ,

$$\psi = \psi_0, \quad \psi = \psi_1, \quad \psi_0 = 1, \quad \psi_1 = 0,$$

where $\psi_0, \psi_1, \psi_2, \dots$ are any three integers. Using x_1 and x_2 to indicate such points, we will take $x_1 = 0, 1, 2, \dots$ and $x_2 = 0, 1, 2, \dots$ and $x_3 = 0, 1, 2, \dots$ in the lattice points. Thus the base state is represented by the symbol "electron at x_1, x_2, x_3 " and the amplitude for an electron to come into $|1, 1, 1\rangle$ is this function $\psi(x_1, x_2, x_3) = \delta_{(x_1, x_2, x_3), (1, 1, 1)}$.



Fig. 13-5. The real part of $\psi(x)$ as a function of x for a superposition of several states of different energy. The spacing λ is very small on the scale of x shown.

As before, the amplitudes $C(x, y, z)$ may vary with time. With our assumptions, the Helmholtz equation should be like this:

$$\begin{aligned} i\hbar \frac{\partial C(x, y, z)}{\partial t} + E_x C(x, y, z) - A_x C(x + a_x, y, z) - A_x C(x - a_x, y, z) \\ - A_y C(x, y + a_y, z) - A_y C(x, y - a_y, z) \\ = \beta_x C(x, y, z - c) - \beta_x C(x, y, z + c). \quad (13.23) \end{aligned}$$

It looks rather long, but you can see where each term comes from.

Again we can try to find a solution $C(x, y, z)$ in which all the C 's have the same time dependence. Again, the solution is an exponential:

$$C(x, y, z) = e^{-iE_x t/\hbar} e^{ik_x x} e^{ik_y y} e^{ik_z z}. \quad (13.24)$$

If you substitute this into (13.23) you get stuck in worse, provided that the energy E is related to k_x , k_y and k_z in the following way:

$$E = E_0 + \Delta_x k_x a_x + \Delta_y k_y a_y + \Delta_z k_z a_z. \quad (13.25)$$

The total E now depends on the three wave numbers k_x , k_y , k_z , which, recall, are the components of a three-dimensional vector \mathbf{k} . In fact, we can write Eq. (13.25) in vector notation as

$$C(x, y, z) = e^{-iE t/\hbar} e^{i\mathbf{k}\cdot\mathbf{r}}. \quad (13.26)$$

The amplitude varies as a complex plane wave in three dimensions, moving in the direction of \mathbf{k} , and with the wave number $k = (k_x^2 + k_y^2 + k_z^2)^{1/2}$.

The energy associated with this stationary state depends on the three components of \mathbf{k} in the complicated way given in Eq. (13.25). The energy of the vector of E with \mathbf{k} depends on relative signs and magnitudes of A_x , A_y , A_z and β_x . These three numbers are all positive, and, if we are interested in \sin^2 values of k , the dependence is definitely circular.

Expanding the energies we did before to get Eq. (13.16) we can now get this:

$$E = E_{0,0,0} + A_x^2 k_x^2 + A_y^2 k_y^2 + A_z^2 k_z^2. \quad (13.27)$$

For a simple cubic lattice with lattice spacing a we expect that k_x and k_y and k_z would be equal—say $2\pi/a$ for k_x and k_y would be just

$$\begin{aligned} E &= E_{0,0,0} + A_x^2 k_x^2 + k_x^2 + k_x^2 \\ &= E_{0,0,0} + 4A_x^2 k_x^2. \quad (13.28) \end{aligned}$$

This is just like Eq. (13.16). Following the arguments laid down, we could then claim that an electron particle in three dimensions moves by superposing many states with nearly equal energies; also moves like a classical particle with some effective mass.

In a crystal with a lower symmetry, there will be no such symmetry effect so the size of the electron potential term is not symmetric; the three coefficients A_x , A_y and A_z are different. Then the "effective mass" of the electron localized in a small region depends on its direction of motion. It would, for instance, have a different inertia for motion in the x -direction than for motion in the y -direction. (The details of such a situation are consequences described in terms of an "effective mass tensor".)

13.6 Other states in a lattice

According to Eq. (13.24) the electron particles we have been calling now can have energies only in a certain "band" of energies which covers the energy range from the minimum energy

$$E_0 = 2(A_x + A_y + A_z)$$

$$E_1 = \lambda A_1 + A_2 = 4\lambda.$$

Other energies are possible, but they belong to a different class of electron states. For the states we have described, we imagine base states in which an electron is placed on one atom of the crystal in some extended state, say, the lowest energy state.

If you lose a e^- , in empty space, and add an electron to make an ion, the ion can be formed in many ways. The electron can gain enough energy to form the state of lowest energy, or it can go on to form one of another of many possible "excited states" which include with a definite probability all other energy levels. For example, let the energy E_1 be picked above corresponds to base states which are ions of the lowest possible energy. We could also imagine to have set of base states in which the electron is in the other ion in a different way—in one of the excited states of an ion—so that the energy E_1 is now quite a bit higher. As before there is now only one A_1 different from before so that the electron will jump from its excited state in one atom to the same excited state in a neighboring atom. The whole analysis goes as before, we find a band of possible energies centered at a higher energy. There can, in general, be many such bands each corresponding to a different level of excitation.

This is not the only possibility. There may be some amplitude that the electron jumps from one energy condition to another from another neighbor at the next atom. (This is called an interaction between bands.) The mathematical theory is more and more complicated as you take into account more and more neighbors and more and more coefficients for leakage between the possible states. The new ideas are however, however, the equations are set up much as we have done in our simple example.

We should remark that there is a much more elaborate side of the various calculations, such as the amplitude A_1 which appears in the theory. Generally they involve the breakdowns, so in actual cases we hardly know them at all. Only when these are determined, however, can we fully calculate the properties of a real situation.

There are other situations where the physics and mathematics are almost exactly like when we last found for an electron moving in a crystal, but in which the "object" and process is quite different. For instance, suppose that our original crystal is stable, but stable? Not a line of neutral atoms, even with a loosely bound extra electron. Then imagine that we are to remove one electron. Which atom has lost its electron? This is, of course, no important assumption, but the electron is mobile from the atom to atom. There will, in general, be some amplitude with that the electron at a neighboring atom, say the $i - 1$ atom, will jump to be the i leaving the $i - 1$ atom without a contradiction. This is the same as saying that there is an amplitude A for the "wandering electron" to jump from the $i - 1$ atom to the $i - 1$ atom. You can see that the equation ϕ_{i-1} is exactly the same, of course, the value of A need not be the same as we took before. Again we can get the same formulas for the energy levels, the "slopes" of probability which move through the crystal with the ϕ_i 's and ψ_i 's of Eq. (13.1), & the effectiveness, and so on. Only here the waves describe the behavior of the carrier electron. Think of it as called "carrier". That is just what a particle with a 25.00 ± 0.005 m^-1 . You can see that this wave will respond to little local changes. We'll have some more to say about such things in the next chapter.

As another example, we can think of a line of isolated atoms's some one of which has been put into an excited state—that is, with more than its normal ground-state energy. Let C_1 be the amplitude that the atom has the excitation. It can interact with a neighboring atom by leaking out to it and reenergy and returning to the ground state. Call the amplitude for this process λ . You can see that, at the same time, C_1 does something else, too. The C_1 's wave is called an exciton. It behaves like a normal "particle" moving from place to place carrying the excitation energy. Such motion might be involved in certain biological

processes such as vision, or photosynthesis. It has been guessed that the absorption of light in the retina involves an "exciton" which moves through some periodic structure such as the layers in the retina described in Chapter 18. Now, I want Fig. 18-10 to be accumulated at some speed, so that where the energy is used to induce a chemical reaction.

18-6 Straining from imperfections in the lattice

We want now to consider the case of a single electron in a crystal which is not perfect. Our earlier analysis says that perfect crystals have surface energy — that is, energy of a gliding atom through hexagonal close-packed, without friction. One of the most important things that can stop an electron from going on forever is an impurity or irregularity in the crystal. As an example, suppose that somewhere in the crystal there is an atom which is larger than average and has stronger gluons, or one of the gluons here so that things there are different from the other atoms/sites. Say, for example, E_0 is the amplitude it would be different. How would we describe what happens there?

To be specific, we will return to the one-dimensional case and we will assume that our number "zero" is an "impurity" atom and has a different value of E_0 , than any of the other atoms. Take call this energy $(E_0 - E_0)$. What happens? When an electron travels at about "zero" there is some probability that the electron is localized (localized). If a wave packet is moving along and it reaches a place where there is a "hole" or "lump", some of it will cut the wave and some of it will bounce back. It's quite difficult to analyze such a situation using a wave packet, because everything varies in time. It's much easier to work with periodic solutions. So we will work with stationary states, which we will find out. We make up of continuous waves which cover system left and infinite for x . In three dimensions we would call the reflected part the scattered wave, since it would spread out in various directions.

We start our work a set of equations which are just like the ones in Eq. (14-6) except that the constant term = 0 is different, or, in the rest. Use the equations for $a = -\lambda$, 0 , λ , -1 , 0 , 1 , and $-\lambda$ and take the limit:

$$\begin{aligned} E_{a=-\lambda} &= E_{\mu=-\lambda} - Aa_{-\lambda} - Aa_{+\lambda}, \\ E_{a=0} &= E_{\mu=0} = 4a_0 = Aa_{-\lambda}, \\ E_{a=\lambda} &= E_{\mu=\lambda} - Aa_{-\lambda} - Aa_{+\lambda}. \quad (18-8) \\ E_{a=-1} &= E_{\mu=-1} - Aa_{-\lambda} - Aa_{+\lambda}, \\ E_{a=1} &= E_{\mu=1} - Aa_{-\lambda} - Aa_{+\lambda}, \\ &\vdots \qquad \vdots \end{aligned}$$

There are, of course, all the extra equations for a if a is greater than λ . They will look just like Eq. (14-6).

For the general case, we really ought to use a different A for the amplitude that the electron jumps to a from down "zero", but the main features of what goes on will come out of a simplified example which all the A 's are equal.

Equation (18-8) would still work as a solution for all of the equations except the one for atom "zero". Is it right for that one equation? We need a different solution which we can cook up in the following way. Equation (14-6) represents a wave going in the positive x direction. A wave going in the negative x direction would have been an equally good solution. It would be written

$$\phi(x) = e^{-ikx}.$$

The most general solution we could have taken for Eq. (18-8) would be a combination

solution of a forward and a backward wave, namely

$$u_0 = \alpha e^{i k x} + \beta e^{-i k x}, \quad (11.39)$$

This solution represents a complex wave of amplitude α moving in the $+x$ direction and a wave of amplitude β moving in the $-x$ direction.

Now let's look at the set of equations for the new problem. The two in (11.28) together with three for all the other terms. The equations involving u_0 , v_0 , E_0 , ϕ_0 are satisfied by Eq. (11.29) with the condition that β is related to β_0 and the incident spacing h by

$$\beta = E_0 - i k \cos \theta h. \quad (11.40)$$

The physical meaning is an "incident" wave of amplitude α approaching from the left ("extinct" from the left), and a "reflected" wave of amplitude β going back toward the left. We do not care how β behaves if we set the amplitude α of the incident wave equal to 1. Then the amplitude β is, in general, a complex number.

We can see all the same things about the solutions of u_0 for $v \geq 1$. The coefficients α and β are different so we would have for here

$$u_0 = \alpha e^{i k x} - \beta e^{-i k x}, \quad \text{for } v \geq 1. \quad (11.41)$$

From $v > 1$ the amplitude of u_0 wave going to the right and a wave coming from the right, we were to consider the physical situation in which a wave is originally started only from the left, and there is only a "transmitted" wave that continues beyond its source in a pulse form. We will try for a solution in which $\beta = 0$. We can certainly satisfy all of the equations for the u_0 except for the middle three of (11.28) by the following two solutions

$$\begin{aligned} u_0(\text{for } v < 0) &= \alpha e^{i k x} + \beta e^{-i k x}, \\ u_0(\text{for } v > 0) &= -\beta e^{-i k x}. \end{aligned} \quad (11.42)$$

The situation we are setting up is illustrated in Fig. 11-5.

By using the formulae in Eq. (11.22) for $\alpha_{1,0}$ and $\alpha_{0,1}$, the three middle equations of (11.28) will allow us to solve for α and β for the two conditions given. So we have to add a complex solution. Setting $v = 1$ we have found for these solutions

$$\begin{aligned} (\alpha - \beta_0)(e^{i k x} - \beta_0 e^{-i k x}) &= -4(\alpha_1 - \alpha_0 e^{-i k x} - \beta_0 e^{-i k x}), \\ (S - E_0 - i k h) &= -4k^2 e^{i k x} - \beta^2 e^{-i k x} - \beta_0^2 e^{-i k x}, \quad (11.43) \\ (E_0 - S) \alpha e^{i k x} &= -4(\alpha_0 e^{-i k x} - \beta_0). \end{aligned}$$

Remember that S is given in terms of V by Eq. (11.30). If you substitute this value for S into the equations, and remember that $\cos^2 \theta = \frac{1}{2}(e^{i k x} + e^{-i k x})$, you get from the first equation that

$$\alpha e^{i k x} + \beta e^{-i k x} = 0, \quad (11.44)$$

and from the third equation (11.28)

$$\alpha_0 e^{-i k x} = 0. \quad (11.45)$$

These are consistent only if

$$\gamma = 1 - \beta. \quad (11.46)$$

This says that the reflected wave β is just the original incident wave (11.39) with an added wave γ equal to the reflected wave. This is not always true, but happens to be true for a particle impact one atom long. If the ϵ were a lamp of impunity γ and the ϵ were added to the forward wave would not necessarily be the same as the reflected wave.

SCATTERED WAVE

$\theta = 0$	$\theta = \pi/2$	$\theta = \pi$
α_0	$\alpha_{0,1}$	$\alpha_{1,0}$

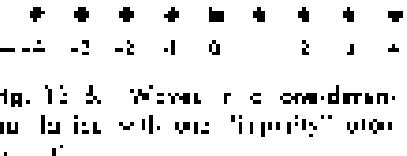


Fig. 11-5. Waves in a one-dimensional barrier with one "impurity" atom ϵ .

We can get the amplitude α of the reflected wave from the middle equation of Eq. (13.33); we find that

$$\alpha = \gamma \frac{-1}{2.0 \sin k_1} \quad (13.34)$$

We have the complete solution for both the wave function and the current.

You may be wondering how the transverse wave can be "brighter" than the incident wave as suggested in Fig. (13.34). Remember though, that β and γ are complex numbers and that the number of electrons (or holes, i.e., probability of finding a particle) in a wave is proportional to the absolute square of the amplitude β ; thus, there will be "transmission of that one" only if

$$|\beta|^2 + |\alpha|^2 = 1. \quad (13.35)$$

You can show that this is true for our solution.

13-7 Trapping by a lattice impurity

There is another interesting situation that can arise if E is a very low number of the energy of the electron is lower at the impurity atom for $k = 0$ than it is anywhere else in the electron gas in the rest of the system. That is, if $E_0 + E_1$ is below the bottom of the band ($E_0 = 0$), then the electron will "trap" itself in a state with $k < k_0$. On such a solution cannot come out of what we have done so far. We cannot, however, if we permit the real solution we look in Eq. (13.15) to have an imaginary number for k , for example, $k = A + iB$, we can have different solutions for $n < 0$ and $n > 0$. A possible solution for $n < 0$ might be

$$\psi_n \text{ for } n < 0 = e^{-kx} \quad (13.36)$$

We have to take a plus sign at the exponent, otherwise the amplitude would get indefinitely large for large negative values of x . Similarly, a possible solution for $n > 0$ might be

$$\psi_n \text{ for } n > 0 = e^{kx} \quad (13.37)$$

If we put these trial solutions into Eq. (13.28) all but the middle three are satisfied provided that

$$E_0 - E_1 = A(e^{k_0} + e^{-k_0}) \quad (13.38)$$

Since the sum of the two exponential terms is always greater than 2, the energy is below the Fermi level and is what we are looking for. Then solving these equations in Eq. (13.38) we find if $e = A$ and $k = B$ is chosen so that

$$A(e^{k_0} - e^{-k_0}) = -E_1 \quad (13.39)$$

Combining this equation with Eq. (13.15) we obtain the energy of the trapped electron, we get

$$E = E_1 - \sqrt{4k^2 + E_1} \quad (13.40)$$

The trapped electron has a binding energy—i.e., it is located below the conduction band.

Note that the amplitudes we have in Eqs. (13.39) and (13.40) do not say that the trapped electron has significant energy above. The probability of finding the electron or nearby atoms is given by the square of these amplitudes. The one particular choice of the parameters it might vary to show: in the k_0 graph of Fig. 13-2, the probability α goes to 0 for finding the electron in the impurity atom. The density denotes the probability that the electron will be far away from the impurity atom. This is another way of "localization" from the point of view of classical physics (the electron doesn't have enough energy to get away from the energy well of the trapping atom). In quantum mechanics it could look out in this way.

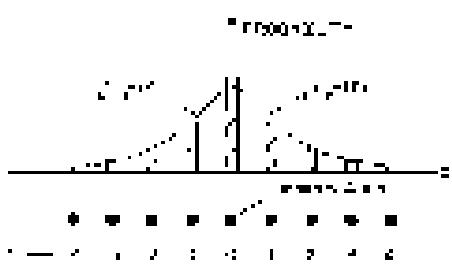


Fig. 13-7. The value of probability of finding a trapped electron at atoms sites near the trapping impurity atom.

13-8 Scattering amplitudes and bound states

Very often examples can be used to illustrate a point which is very useful in the study of the physics of high-energy particles. It has to do with a relationship between scattering amplitudes and bound states. Suppose we have determined strong experimental and theoretical reasons to say that pions can form protons. Then a new particle is discovered and someone wonders whether maybe it is just a combination of a pion and a proton held together in some bound state at $E = 0$? Energy in the way an electron is bound to a proton to make a hydrogen atom. But a bound state we need a combination which has a lower energy than the free propagator.

The π is a "good" theory which says that a bound state will occur if the energy E , where the scattering amplitude becomes infinite, extrapolated algebraically to the mathematical limit is "physically meaningful" to values of E outside of the permitted band.

The physical reason for this is as follows. A bound state is a situation in which there are only waves allowed to be a pion and there's no wave coming in or outgoing, angular momentum and the relative proportion between the so-called "incident" or external wave and the wave being "sent off" is definite. We can test this idea in our example. Let's write an expression Eq. (13.37) for the scattered amplitude directly in terms of the energy P of the particle being scattered from and to a bound state. Since Equation (13.30) can be rewritten as

$$(\Delta \sin \phi) = \sqrt{1/P^2 - (k - k_0)^2}$$

the scattered amplitude is

$$\delta = \frac{-i}{P + \sqrt{1/P^2 - (k_0 - k)^2}}. \quad (13.44)$$

For more satisfaction, this equation should be used only for real states—those with energies in the energy band, i.e., $E_0 < E < E_1$. But suppose we forget this fact and extend the formula into the "unphysical" energy regions where $|E - E_0| > 2.5$. For these unphysical regions we can write

$$\sqrt{1/P^2 - (k_0 - k)^2} = \sqrt{(E - E_0)^2 - 1/P^2}.$$

Then the "scattering" amplitude, whatever it may mean, is

$$\delta = \frac{i}{P - \sqrt{(E - E_0)^2 - 1/P^2}}. \quad (13.45)$$

Now we want to know by energy after which δ becomes infinite (i.e., for which the expression for δ has a "pole"). Yes, so long as E is negative, the denominator of Eq. (13.45) will be zero also

$$(E - E_0)^2 = 1/P^2 - P^2,$$

or when

$$E = E_0 + \sqrt{2 + P^2} = E_1.$$

Term minus sign gives just the energy we found in Eq. (13.43) for the trapped energy.

What about the plus sign? This gives an E larger than the allowed energy band. And, in fact there is another bound state there which we missed when we solved the equations of Eq. (13.25). We leave it as a puzzle for you to find the energy and amplitude δ for this bound state.

The relation between scattering and bound states provides one of the most useful clues in the current search for an understanding of the experimental observations above. The new strange Δ 's are

[†] The sign of the pole to be shown here is a technical point related to the allowed signs of ϵ in Eqs. (13.39) and (13.40). We won't go into it here.

Semiconductors

14.1 Electrons and holes in semiconductors

One of the most valuable contributions developed in recent years has been the application of solid-state electronics to medical diagnosis and electrical devices used in space vehicles. The study of semiconductors led to the discovery of their useful properties and to a large number of important applications. The field is changing so rapidly that what we tell you today may be obsolete next year. It will certainly be incomplete. And it is perfectly clear that with the enormous growth of these electronics many new and more powerful things will be possible as time goes on. You will no doubt continue and go deeper into the subject later in this volume, but you may find it interesting to see that at least something of what you are learning has some relation to the practical world.

There are many numbers of semiconductors. However, however, there are a few that now have the greatest technical importance. These are also the ones that have been studied, and our understanding of them is well advanced. An understanding of most of the others—the semiconductors of interest in this section use today are silicon and germanium. These elements crystallize in the same kind of face-centered cubic lattice as the atoms in the metal, but they bonding with their four nearest neighbors. They are insulators at very low temperatures—near absolute zero—but though they are poor conductors everywhere else, at room temperature. They are *semiconductors*; they are called *nonsolids*.

If we continue our examination of the crystal of silicon or germanium at which is at low temperature, we see no energy regions we expected in the last chapter. The electron will be able to wander around in the crystal, just as our electron did in the last. Actually, we have looked only at the nonconduction electrons in a rectangular lattice, but the situation would be somewhat different in the real lattice, classified as *isotropic*. All of the essential points are, however, shown by the results for the rectangular lattice.

As we saw in Chapter 13, free electrons can have energy only in discrete energy bands called the *conduction bands*. Within this band the energy is allowed to increase in steps of the period of the crystal lattice (see Fig. 13.24); but

$$E = E_0 - 2A_0 \cos k_x - 2A_0 \cos k_y - 2A_0 \cos k_z. \quad (14.1)$$

The A 's are the couplings between the x , y , and z directions, and k 's are the lattice spacing's in the directions.

For energy near the bottom of the band, we can approximate Eq. (14.1) by

$$E = E_{min} - A_0 k^2 + A_0 k^2 = E_0 - \omega^2, \quad (14.2)$$

(see Section 13.4).

Now think of electrons moving, in some particular direction, so that the crystal potentials of A are constant. The state $E_0 - \omega^2$ is the energy of the conduction electrons in the band as we have seen it. The minimum of the electron. We can write

$$E = E_{min} + \omega^2, \quad (14.3)$$

where ω is some constant, and we can take a graph of E versus k as in Fig. 14.1. We call such a graph an "energy diagram." An electron in a particular state of the xy - yz momentum k is indicated by a point such as S in the fig. 14.1.

14.1 Electrons and holes in semiconductors

14.2 Lattice semiconductors

14.3 The Hall effect

14.4 Semiconductor junctions

14.5 Rectification and thermionic junctions

14.6 The transistor

Properties of the solid insulators in solid-state physics. Chapters 13, 14, and 15

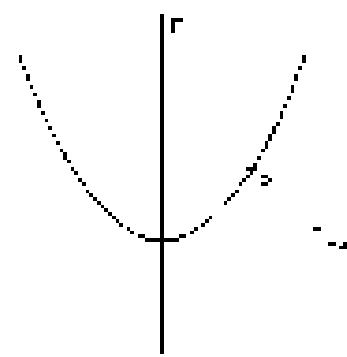


Fig. 14.1. The energy diagram for a lattice system with finite size.

As we also mentioned in Chapter 13, we can have a similar situation if we remove an electron from a neutral atom. The electron can jump over from a nearby atom and fill the hole, thus leaving another hole in the original atom. We can describe this behavior by writing an amplitude to find the hole in any particular atom, and by saying that the hole can jump from one atom to the next. (Clearly, the amplitudes χ that the hole jumps from atom a to the next b [clearly, the amplitudes χ that the hole jumps from atom a to atom b] is just the same as the amplitude that an electron from atom b jumps into the hole in atom a .) The mathematics is just the same [i.e., the form is the same]—electron, and we get again that the energy of the hole is related to its wave number by $E = \hbar c / \lambda$, just like Eq. (14.1) or (14.2), except, of course, with different numerical values for the constants \hbar , c , λ_0 , and A . The hole has an energy relative to the wave numbers of the probability amplitudes. Its energy lies in a discrete band, and near the bottom of the band it increases as its quantification with the wave number—in momentum—just as in Fig. 4-1. Following the arguments of Section 14-3, we would find that the hole also diffuses like a classical particle with a certain effective mass, except that in semiconductors the mass depends on the direction of motion. So the hole behaves like a positive particle moving through the crystal. The charge of the hole particle is positive, because it is caused by the absence of a negative electron; and other "holes" in your mind are temporarily electrons moving in the opposite direction.

If we put several electrons into a neutral crystal, they will move around much like the atoms of a low-pressure gas. If there are no terminals, their interactions will not be very important. If we then put an electric field across the crystal, the electrons will start to move and an electric current will flow. Eventually they would all be driven to one edge of the crystal, and, if there is a metal terminal there, they would be collected, leaving the crystal negative.

Similarly we could put many holes into a crystal. They would roam around at random unless there is an electric field. With a field they would flow toward the negative terminal, and would be "collected"—what actually happens is that they are annihilated by electrons from the metal terminal.

One can also form hole-holes and electrons together. If there are not too many, they will all go their way independently. With an electric field, they will contribute to the current. For convenience, electrons are called the majority carriers and the holes are called the minority carriers.

We have so far considered that electrons are put into the crystal from the outside, or are removed to make a hole. It is also possible to "create" an electron-hole pair by taking a free electron away from one atom's atom and putting it once more away in the same crystal. We then have a free electron and a free hole, and the two can interact, as we have described.

The energy required to put an electron into the S —we say to "create" the state S —is the energy E^+ shown in Fig. 14-2. It is some energy above E_{∞} . The energy required to "create" a hole in the state S is the energy E^- of Fig. 14-3, which is even energy greater than E_{∞}^+ . Now if we create a pair in the states S and S' , the energy required is just $E^+ + E^-$.

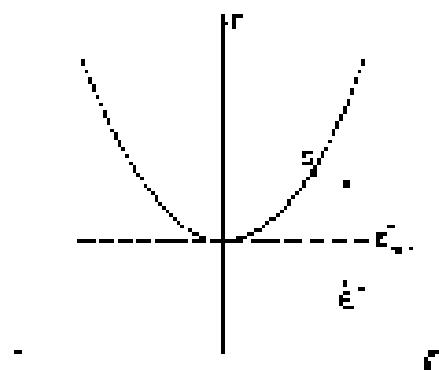


Fig. 14-2. The energy E^+ is required to "create" a free electron.

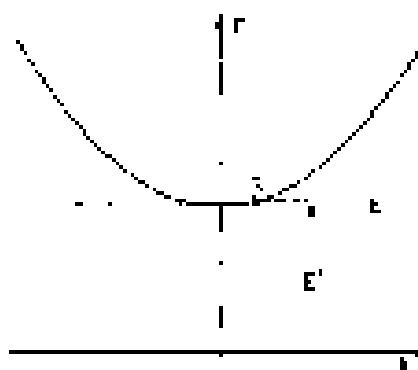


Fig. 14-3. The energy E^- is required to "create" a hole in the state S' .

The creation of pairs in a solid may proceed (as we will see later) so many people like to put Fig. 14-3 and Fig. 14-4 together on the same graph—with the hole energy plotted symmetrically about the zero of carrier energy. We have combined our two graphs in this way in Fig. 14-4. The advantage of this is that a graph of total energy $E_{\text{total}} = E + E'$ is simply a curve of a pair with the electron at S and the hole at S' to set the vertical distance between S and S' as shown in Fig. 14-4. The minimum energy required to break a pair is called the "gap energy" and is equal to $E_{\text{gap}} = E_S + E_{S'}$.

So, when you will see a similar diagram called an energy level diagram which is drawn when people are not interested in backscattering. Such a diagram is shown in Fig. 14-5—just showing the possible energies for the electrons and holes.¹⁴

How can electron-hole pairs be created? There are several ways. For example, photons of light (or x-rays) can be emitted and absorbed. If the photon energy is above the energy of the gap, the rate at which pairs are produced is proportional to the light intensity. If two electrodes are placed on a piece of the crystal and a "bias" voltage is applied, the electrons and holes will be drawn to the electrodes. The current increase is proportional to the intensity of the light. This mechanism is responsible for the phenomenon of photoconduction and the operation of photoconductive cells.

Electron-hole pairs can also be produced by high-energy particles. When a fast-moving charged particle, for instance, a proton or a positron with an energy of tens or hundreds of MeV goes through a crystal, its electric field will knock electrons out of their bound states creating electron-hole pairs. Such events occur hundreds of thousands of times per millimeter of track. After the passage of the particle, the carriers can be collected and measured as a "dissipation electrical pulse." This is the mechanism of "pulsed semiconductor counters" recently popular for experiments in nuclear physics. Such counters do not require semiconductors; they can also be made with crystalline materials. In fact, the best of such counters was made using a diamond crystal, which is at the same time transparent. Very pure crystals (such as diamond) allow holes and electrons to be lost to absorption as the charge carriers without being captured. The semiconductors silicon and germanium are useful because they can be purified with high purity in relatively large sizes (see, i.e., Sec. 14.10).

So far we have been concerned with semiconductors at temperatures near absolute zero. At any finite temperature there is still another mechanism by which electron-hole pairs can be created. The energy can be provided from the thermal energy of the crystal. The thermal energy is released as it transfers its energy to electrons going into "thermionium" emission.

The probability per unit time that the energy of charge is the gas energy E_{gas} will be concentrated around E_{gas} is proportional to $e^{-E_{\text{gas}}/kT}$, where T is the temperature and k is the Boltzmann constant (see Chapter 4), and E_{gas} is the appearance probability, but as the temperature rises there is an increasing probability of producing such pairs. At any finite temperature the generation should continue forever in a random way giving some real mean negative and positive charges. Of course that does not happen because over time the electrons and holes make fully full each state—the electrons drop into the lowest free-carrier energy available to the lattice. When that is the electron and hole "concentration," there is a certain probability per second that a hole meets an electron and the two things annihilate each other.

If the number of electrons per unit volume is N_e and the negative ionized and the density of positive carriers is N_h , the chance per unit time that an electron and a hole will find each other and annihilate is proportional to the product $N_e N_h$. Equilibrium is then found when the rate that pairs are created. You see that

¹⁴ In many books the same energy diagram is also placed in different ways. The carrier μ refers only to electrons. Instead of thinking of the energy of the hole, just think of the energy in case no hole has to be filled to hole. The energy is lower than the backscattered energy—in fact, just the inverse law, just you see in Fig. 14-5. With this interpretation of the energy scale, the gas energy is the minimum energy which must be given to an electron to move it from its bound state to the conduction band.

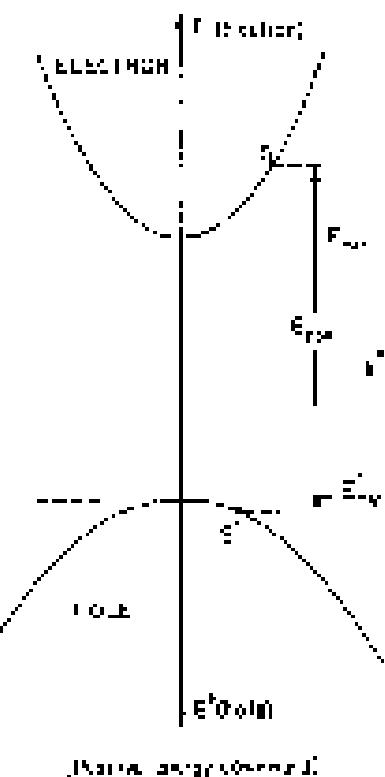


Fig. 14-4. Energy diagram for electron-hole pair creation.

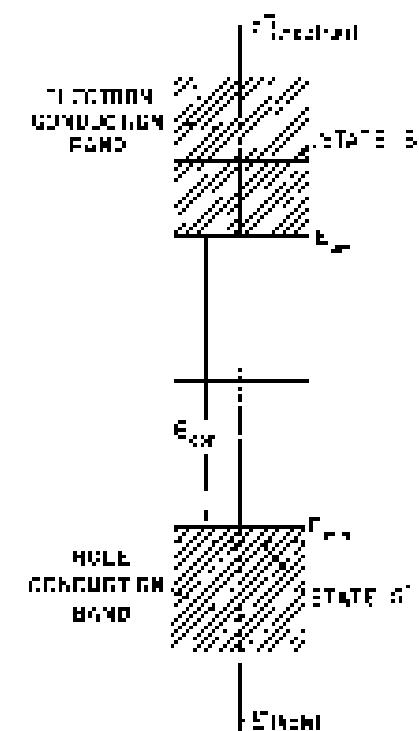


Fig. 14-5. Energy level diagram for electron and hole.

equilibrium the product, $N_e N_h$, should be given by some constant times the "Reitman factor":

$$N_e N_h = \text{constant} \cdot A e^{-E_g/RT}, \quad (14-4)$$

When we say constant, we mean nearly constant. A more complete theory would include more details about how holes and electrons ("fluid" with other) share the band. "Constant" is slightly dependent upon temperature, but the major dependence on temperature is in the exponential.

Let's consider, as an example, a pure material which is originally neutral. At a finite temperature you would expect the number of positive and negative carriers to be equal, $N_e = N_h$. Then each of them should vary with temperature as $e^{-E_g/RT}$. The variation of σ , any of the properties of a superconductor, is "constant" — for example — & mainly determined by the separation E_g between all the other factors very much more slowly with temperature. The gap energy for germanium is about 0.73 eV at the room temperature.

At room temperature (20°C) it is about $1/20$ of an electron volt. At these temperatures there are enough holes and electrons to give a significant conductivity, while at 4°K , 30°K , there is no room temperature. The conductor is then superconducting. The gap energy of germanium is in fact very small — ≈ 0.01 eV at the room temperature.

14-2 Impurity semiconductors

So far we have talked about two ways that electrons can be pulled in a heavily ideally perfect crystal lattice. One way was to let the electrons from an outside source. The other way, was to knock a bound electron out of a donor atom creating simultaneously an electron and a hole. It is possible to produce this in a the conduction band of a crystal in still another way. Suppose we imagine a crystal of germanium in which one of the germanium atoms is replaced by an arsenic atom. The germanium atom is a lone valence of four, the arsenic atom is controlled by the four valence electrons. As it is, in other words, has a valence of 5. It loses part of a single bond atom first at the germanium lattice because it has approximately the largest size, but in doing so it must be to a valence of 4 — using four of its valence electrons to form the crystal bonds and leaving one electron left over. This extra electron is very loosely attached — the binding energy is less than 1/10 of a volt. At room temperature the electron easily picks up sufficient energy from the thermal energy of the crystal, and then dissociates its own — moving away in the lattice as a free electron. An impurity atom such as the arsenic is called a donor site because it can give up a negative charge to the crystal. If it consists of just one atom it is a point source of charge, and instead of the entire face being active, the arsenic atom sites will be distributed throughout the crystal and the crystal will have a certain density of negative charge built in.

You might think that if a high voltage could be applied across the crystal, this will not happen, however, because the arsenic atoms in the body of the crystal will have a positive charge. If the body of the crystal is to remain neutral, the average density of negative ionization must increase in the vicinity of the arsenite. If you put two electrodes on the edges of such a crystal and connect them to a battery, a current will flow, but as the counter-electrons are swept out of one end, new creation of electrons must be introduced from the other end, so that the average density of conduction electrons is left very nearly equal to the density of donor sites.

Since the donor sites are positively charged there will be some tendency to tend to capture some of the conduction electrons as they pass through the crystal. A donor site can therefore act as a trap just as does the discussed in the last section. But if the trapping may be sufficiently small — as is for arsenic — the number of carriers which are trapped in any one site is a small fraction of the total. For a complete understanding of the behavior of semiconductors

one more time introduced to trapping. For the case of no dissipation, however, we see again that the trapping energy is sufficiently low and the temperature is sufficiently high that all of the donor sites have given up their electrons. This is, of course, just an approximation.

It is also possible to work in a germanium crystal with impurity atom whose valence is 2, such as aluminum. The additional atoms are set as a witness & /not/ by adding an extra electron. I can steal an electron from some nearby germanium atom and end up with a negative charge, positive hole or hole carrier. Of course, when it steals the electron from a germanium atom, it leaves a hole there; and this hole can wander around in the crystal as a free hole carrier. An impurity atom which can produce a hole in this way is called an acceptor because it "accepts" an electron. If a portion of a silicon crystal is grown from a melt in which a small amount of aluminum is added, the crystal will have built-in a certain density of holes which can act as positive carriers.

When a donor or an acceptor impurity is added to a semiconductor, we say that the material has been *doped*.

When a portion of a crystal with some built-in impurities is at room temperature, some conduction electrons are compensated by the thermally induced electron hole pairs created as we saw by the donor sites. The electrons from both sources are, usually, equal, $N_{\text{e}} = N_{\text{h}}$, in the total number. N_{e} which comes into play in the statistics processes, lastly, is equilibrium. If the impurity is not too low, the number of negative charges contributed by the donor impurity atoms is roughly equal to the number of its valence atoms added. In equilibrium $N_{\text{e}} \approx N_{\text{h}}$ must still be satisfied at given temperature, T . If $N_{\text{e}} < N_{\text{h}}$, it is determined this means that it is odd to do some trapping which increases N_{h} . The number N_{p} of positive carriers will have to decrease to such an extent that $N_{\text{e}}, N_{\text{p}}$ is unchanged. If the impurity concentration is high enough, the number N_{p} of negative carriers is determined by the number of donors and is mostly independent of temperature. All of the variation in the electrical behavior is supplied by N_{e} , even though it is much less than N_{p} . An alternative picture would be a small concentration of donor impurity with a majority of negative carriers; such a material is called an "n-type" semiconductor.

If the acceptor-type impurity is added to the crystal in the same way, it does not add any net and significant source of free electrons produced by thermal ionization. The process will go on until Eq. (2-4) is satisfied. Under appropriate conditions, the number of positive carriers will be increased and the number of negative carriers will be decreased, leaving the positive a majority. A material with an excess of positive carriers is called a "p-type" semiconductor.

If we put two electrodes on a piece of semiconductor crystal and connect them to a source of potential difference, then we have electric fields inside the crystal. The electric field will cause the positive and the negative carriers to move, and an electric current will flow. Let's consider "n-type" material in an n-type material in which there is a large majority of negative carriers. For such material we can disregard the holes. They can contribute very little to the current because there are so few of them. In addition, if the carriers want to move, without any impurities, in > the crystal at < thin temperature, however, especially in a crystal with some impurities, the electrons do not move completely freely. They are continually making collisions which knock them out of their original trajectories, that is, changing their momentum. These collisions are just exactly the scattering we talked about in the last chapter and occur at ANY frequency in the crystal lattice. In addition, another the main source of scatterings are random thermal sites that are probably present here. Since the conduction electrons have a very slightly different energy at two points here, the probability waves are scattered at all the points. Even in a perfectly pure crystal, however, there are going to be temperature changes, either in the lattice due to thermal vibrations from the classical point of view we can say that the atoms aren't lined up exactly in a regular lattice, but are, at a certain, slightly out of place due to their thermal

currents. The energy loss associated with such current flowing in the resistor as described in Chapter 12 makes it difficult to place so much heat that the losses of resistivity or ohmics are not transmitted perfectly but are scattered in the insulating foreign. At very high temperatures or for very pure materials this scattering may become important, but in most doped substances used in practical devices the impurity atoms contribute more of these losses. We would like now to make an estimate of the electrical conductivity of such a material.

When an electric field is applied to such type semiconductors, that negative carrier will be accelerated in this field, picking up velocity until it is scattered from one of the carrier sites. This means that the carriers which are ordinarily moving occur in a random fashion will have their site, erratic, walk pick up an average drift velocity along the lines of the electric field and give rise to a current through the system. The drift velocity v_d is in general rather slow compared with the typical thermal velocities so that we can estimate the current by assuming that the average time between collisions between scattering is unimportant. Let's say that the negative carrier has an effective positive charge q_e . In an electric field E , the force on the carrier will be $q_e E$. In Recd. 42.5 of Volume 1 we calculated the average drift velocity in units of cm^2/sec and found it to be given by $\frac{q_e E}{\pi \sigma m_e \tau_e}$, where q_e is the charge on one charge, τ_e is the mean free time between collisions, and m_e is the mass. We include the effective mass we mentioned in the last chapter but since we want to do a rough calculation we will suppose that this effective mass is the same as the directions. Then we will call it m_e . With this approximation the average drift velocity will be

$$v_{d,eff} = \frac{q_e E \tau_e}{m_e} \quad (14.1)$$

Knowing the drift velocity we can find the current. Electric current density j is just the number of carriers per unit volume, N_e , multiplied by the average drift velocity, and by the charge on each carrier. The current density is therefore

$$j = N_e v_{d,eff} q_e = \frac{N_e q_e^2 E}{m_e} \tau_e \quad (14.2)$$

We see that the current density is proportional to the electric field, with a semi-conducting constant α says Ohm's law. The coefficient of proportionality between j and E is conductivity σ .

$$\sigma = \frac{N_e q_e^2 \tau_e}{m_e} \quad (14.3)$$

For the n-type material the conductivity is relatively independent of temperature. First, the number of mobile carriers N_e is determined primarily by the density of donors in the crystal (as long as the temperature is not so hot that some of the carriers are suppressed). Second, the mean time between collisions τ_e is mainly controlled by the density of impurity atoms which is, of course, independent of the temperature.

We can apply all the same arguments to p-type materials, changing only the signs of the parameters which appear in Eq. (14.3). If there are comparable numbers of both heavy hole and positive carrier present, then we must add the contributions from each kind of carrier. The total conductivity is then

$$\sigma = \frac{N_h q_e^2 \tau_h + N_p q_e^2 \tau_p}{m_h + m_p} \quad (14.4)$$

For very pure materials, N_h and N_p will be nearly equal. They will be smaller than in a doped material, so the conductivity is "by loss". Also they will vary rapidly with temperature (see $\propto T^{-3/2}$), so the behavior of the conductivity may change extremely fast with temperature.

14.2 The Hall effect

It is certainly a puzzle. Using the circuit resistance where the bulk is relatively free of defects and dislocations, there should be no electrical current carried by holes that behave like positive particles. We would have, therefore, to reverse sign of current. This shows in a rather clear way that the sign of the carrier electric current is again definitely positive. Suppose we have a block made of semiconducting material—again let's make it a metal—and we put an electric field E_x in its x -direction, as shown in Fig. 14.2. Now suppose we put a magnetic field B in the block pointing at a right angle to the current, say along the z -axis of the figure. The moving carriers will feel a magnetic force due to $\mathbf{B} \times \mathbf{v}$, and if v is the carrier velocity, it will be right or left depending on the sign of the charge on the carrier—the average magnetic force on the carriers will be either up or down. Now, that it is right. For the direction we have assumed for the current and the magnetic field the magnetic force on the moving charges will always be up. Positive charges moving in the direction v_x to the right will feel an upward force. If the currents consist of negative charges, they will be moving left (for the same sign of the conductive current) and they will also feel an upward force. Under steady conditions, however, there is an upward motion of the carriers because the current can flow only from left to right. What happens is that a few of the charges, likely 10^6 cm^{-2} per centimeter 2 , have charge density along the upper surface of semiconductor. They are equal and opposite carrier charge density along the bottom surface of the crystal. The charges will do so, the top and bottom surfaces will be electric fields they produce on the moving charges just exactly cancel the magnetic force. For the case $v_x > 0$ the carrier current flows horizontally. The charges on the top and bottom surfaces q_{\pm} produce a potential difference vertically across the crystal, which can be measured with a high-resistance voltmeter, as shown in Fig. 14.3. The sign of the potential difference measured by the voltmeter will depend on the sign of the carrier charges responsible for the current.

Not such experiments were first done. It was expected that the standard potential difference would conjugate signs we could expect for n-type conduction electrons. People were, therefore, quite surprised to find that not some materials the sign of the potential difference was in the opposite direction. It appeared that the carrier current was a positive, not a negative, current. From our discussion of doped semiconductors, it is understandable that an n-type semiconductor should produce the sign of potential difference appropriate to negative carriers, and that a p-type semiconductor should give an opposite positive difference since the current is carried by the positively charged holes.

The original discovery of the anomalous sign of the potential difference in the Hall effect was made in a metal rather than a semiconductor. It has been assumed that in metals the conduction was always by electrons. However, we know our v_x for negative the potential difference had the wrong sign. It is now understood that in metals as well as in semiconductors it is possible, in certain circumstances, that the "holes" responsible for the conductive carriers, i.e., the charge carriers moving in the crystal through the moving, never to least the valence shell of the atom, will have the electric current and the source to external fields, because $v_x > 0$, one would expect for an electric current and the positive carriers.

Let's see if we can make a quantitative estimate of the magnitude of the voltage difference measured from the Hall effect. If the voltage in Fig. 14.3 shows a multiple of e times the charge e , the current must be moving from left to right and the vertical magnetic force must be precisely cancelled by a vertical electric field which we can call E_y , (the "Hall" \rightarrow field, "transverse"). If it is electric field E_y caused the magnetic force we must have

$$E_y = v_x e \times B \quad (14.9)$$

Using the relation between the drift velocity and the electric current density given

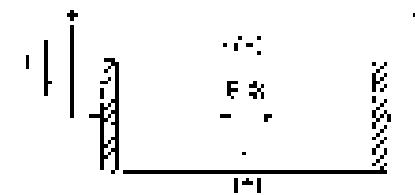


Fig. 14.2. The Hall effect setup from the magnetic forces on the carriers.

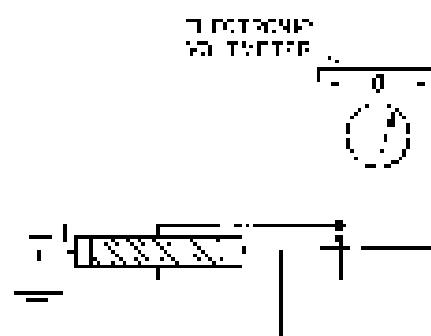


Fig. 14.3. Measuring the Hall effect.

$$V = \frac{q}{\rho A} \cdot \frac{1}{2} B$$

The potential difference between the top and the bottom of the crystal is, of course, the electric field which is supplied by the battery. The electric field strength is proportional to the current density and to the magnetic field strength. The constant of proportionality β_{HJ} is called the Hall coefficient and is usually represented by the symbol R_H . The Hall coefficient depends just on the density of current-carrying carriers of one sign, i.e., on large accuracy. Measurement of the Hall effect is, therefore, one convenient way of determining exactly the density of carriers in a conductor.

14-4 Semiconductor junctions

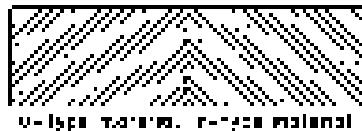


Fig. 14-8. A p-n junction.

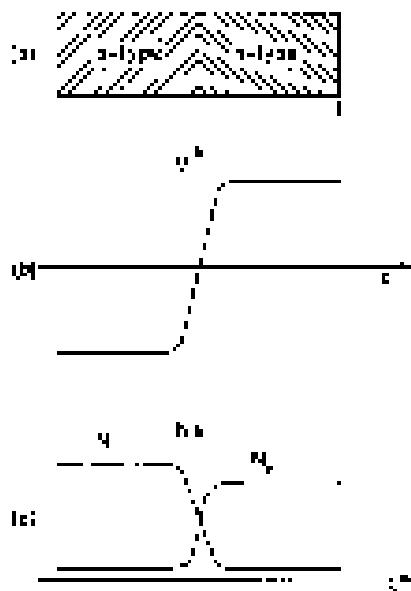


Fig. 14-9. The electric potential and the carrier densities in an idealized semiconductor junction.

We would like to discuss now what happens if we take two pieces of germanium or silicon with different concentrations of dopants, say, one kind of dopants in one piece and another kind in the other. These can be joined together to create a p-n junction, i.e., where we have a type germanium on one side of the boundary, as depicted in Fig. 14-8. Actually, it is not practical to put together two separate pieces of crystal and have them in perfect contact on a microscopic scale. Instead, junctions are made out of a single crystal slice. This was indicated in the two separate regions. One was a small semi-spherical doping impurity in the "n-cell" (the only part of the crystal has grown). Another way is to grow a thin film of a doped element A on the surface of the host. And the crystal contains some impurity atoms to dilute the body of the crystal. In either case, there will always have a sharp boundary through the boundaries can be more or less smooth. A "continuum" of sorts. For example, consider an ideal situation in which each two regions of the crystal with different properties meeting at the boundary.

On the n-type side of the junction there are free electrons which can move about, as well as the fixed donor sites which balance the overall electric charge. On the p-type side there are free holes moving about and an equal number of negative acceptor sites keeping the charge balanced. Actually, last time we saw that in a substitution centre we put the two materials in contact. When they are connected together, the ejection of a charge near the boundary. When the electrons in the n-type material want to be in the p-holes they will not be affected because they would be a free surface, but are able to go right on into the p-type material. Some of the electrons in the n-type material will, therefore, tend to diffuse out into the p-type material where there are fewer electrons. This will cause a build-up because as we lose electrons from the n-side the net positive charge there increases and finally an electric voltage is built up which resists the diffusion of electrons from the p-side. In a similar way, because the carriers of the n-type material can diffuse across the junction into the p-type material, when they do this they leave behind an excess of negative charge. Once equilibrium conditions the net diffusion equals the net recombination. The important about is the electric field which are established in such a way so as to prevent the free electrons from moving the p-type material.

Let us discuss processes we have been describing you in a simple manner and you will notice both act on the diffusion which will charge up the n-type material in positive +ve and the p-type material in a negative sense. Because of the finite conductivity of the semiconductor material, the change in potential from the n-side to the p-side will occur in a relatively narrow region near the boundary; the main bulk of the block of material will have a uniform potential. Let us imagine an electric field direction perpendicular to the boundary surface. Then the electric potential will vary with x , as shown in Fig. 14-9(a). We have a greater concentration of holes in the p-type. The expected variation of the density N_h of holes is $N_h \propto e^{-kx}$ the density N_n of electrons $N_n \propto e^{kx}$ away from the junction. The current densities J_h and J_n should be just the equilibrium density we would expect for individual blocks of the crystal at the same temperature. We have drawn the lines for a

junction at which the p-type material is more heavily doped than the n-type. Because of the potential gradient in the junction, the positive carriers have to climb up a potential hill to get to the n-type side. Thus, there is an equilibrium condition there can be found just below the n-type material that carries no net charge material. Neglecting the way of current, it makes, we expect that the ratio of the carrier densities on the two sides to be given by the "flow equation":

$$\frac{N_{p\text{ in p-side}}}{N_{p\text{ in n-side}}} = e^{-qV/kT} \quad (14.06)$$

The quantity qV is the numerator of the exponential because the energy requires an electric charge qV , through a potential difference V .

We have a parallel similar equation for the density N_n of the negative carriers:

$$\frac{N_{n\text{ in p-side}}}{N_{n\text{ in n-side}}} = e^{-qV/kT} \quad (14.07)$$

With above the equilibrium densities on each of the two extremes, we can use either of the two equations above to determine the potential V across the junction.

Hence that in Eqs. (14.06) and (14.07) you to give the same value for the potential difference V , the product $N_p N_n$ has to be same for the carrier densities on each side. (Remember $N_p = -N_n$.) We have seen earlier, however, that this product depends only on the temperature and the gap energy of the crystal. Provided both sides of the crystal are at the same temperature, the two equations are equivalent with the same values of the potential difference.

Since there is a potential difference from one side of the junction to the other, it induces a drift like a battery. If I hope if we connect a wire from the n-type side to the p-type side we will get an electric current. That would be nice, because then the current would flow forever without using up any material and we would have an infinite source of energy. A violation of the second law of thermodynamics! There is, however, one catch, if you do make a wire from the p-side to the n-side. And the reason is easy to see. Suppose we, in some frequency, turn on a power of induced current. When we connect this wire to the n-type side, we have a junction. There will be a potential difference across this junction. Let's say that there goes one-half the potential difference from the p-type side to the n-type material. When we connect our induced wire to the p-type side of the junction, there is also a potential difference at this junction, so we'll need all the potential difference across the junction. At all other junctions, the potential differences add together themselves so that there is no net current flow in the circuit. Whatever kind of wire you use to connect together the two sides of the crystal, you are producing two new junctions, and in keeping all the junctions at the same temperature, the potential jumps at the junctions to compensate each other and no current will flow in the circuit. It does turn out, however, that if we cut the device so that some of the junctions are at a different temperature than the others, then, perhaps, will flow. Some of the junctions will be heated and others will be cooled by this current and thermal energy will be converted into electrical energy. This effect is known as the Seebeck effect of thermocouples where it is used for measuring temperatures, and of thermoelectric generators. The same effect is also used to cool air by refrigerators.

If we cannot measure the potential difference between the two sides of an *n-p* junction, how can we easily know that the potential gradient shown in Fig. 14-5 really exists? One way is to shine light on the junction. When the light shines on the doped they act as sources or sinks of hole pairs. In the steady state, if a light falls on the junction instead of the edge of the potential barrier of Fig. 14-6 the hole will be driven into the p-type region and the electron will be driven into the n-type region. If the two sides of the junction are now connected to an external circuit, heat and energy will provide a current. The energy of the light is the measured total electrical energy in the junction. The solar cells which generate electrical power for the operation of solar electric satellites operate on this principle.

In our discussion of the operation of a semiconductor junction we have been assuming that the holes and the electrons are moving independently, except that they combine to form a positive ionized equilibrium. When we were describing the current produced by light falling on the junction, we were assuming that all electrons which reached the junction region would get into the metal body of the crystal before being annihilated by a carrier of the opposite polarity. In the intermediate vicinity of the junction, where the density of carriers of both signs is approximately equal, the effect of their mutual annihilation is so small that it is often called "nonradiative recombination." There are, in a crystal and part of a semiconductor junction, must be properly taken into account. We have been assuming that a hole in an electron combined in a junction region has a good chance of getting into the metal body of the crystal before recombining. The typical time for recombination to take place is opposite partner and magnitude is for typical semiconductor junctions about the same between 10^{-1} and 10^{-2} seconds. This time is incidentally much longer than the mean free time τ between collisions with scattering sites in the crystal which we used in the analysis of conduction. In a typical n -type junction, the time for recombination is less formed in the junction region to be negligible over the rest of the crystal is generally much shorter than the recombination time. Most of the pairs will, therefore, contribute to the excess current.

14-5 Rectification at a semiconductor junction

We should like to consider now what happens at a p-n junction under like a rectifier. If we apply a voltage across the junction, a direct current will flow if the polarity is in the direction from the very small current will flow if no voltage is applied in the opposite direction. If an alternating voltage is applied across the junction, a direct current will flow in one direction. The current is "rectified." Let's look again at what is going on in the equilibrium condition represented by the graphs of Fig. 14-4. In the p-type region, there is a large concentration N_p of positive carriers. These carriers are diffusing away and a certain number of them each second approach the junction. The number of positive carriers which approach the junction is proportional to N_p . Most of them, however, are turned back by the high potential hill in the junction and only the fraction $e^{-\frac{qV}{kT}}$ pass through. There is then a current of positive carriers approaching the junction from the p-side. This current is also proportional to the density of positive carriers in the n-type region, but the carrier density here is much smaller than the density in the p-type side. When the positive carriers approach the junction from the n-type side, they find a hill with a negative slope and immediately turn around. In the p-type side of the junction, let's call this current i_0 . Under equilibrium the currents from the two directions are equal. We expect then the following relation:

$$i_0 \sim N_p i_0 e^{-\frac{qV}{kT}} = N_p A_i e^{-\frac{qV}{kT}}, \quad (14-2)$$

You will notice that this equation is really just the same as Eq. (14-10). We have just derived it in a different way.

Suppose, however, that we lower the voltage on one side of the junction by a amount ΔV , which we can do by applying an external potential V to one side of the junction. Now the difference in potential across the junction will be larger by $\Delta V - qV$. The current of positive carriers from the p-side to the n-side will now have this potential difference as its driving factor. Call this current i_1 , we'll get

$$i_1 \sim N_p i_0 e^{-\frac{q(V-\Delta V)}{kT}}.$$

This current is larger than i_0 because the factor $e^{-\frac{qV}{kT}}$ is now less. So we have the following relation between i_1 and i_0 :

$$i_1 = i_0 e^{-\frac{q\Delta V}{kT}}. \quad (14-3)$$

The current from the p-side increases exponentially with the externally applied voltage ΔV . The current of positive carriers from the n-side, however, remains

currents so long as ΔV is not too large. When they approach the barrier, these currents will still be diode currents and will all fall down to the same value if the voltage is less than the natural potential difference V_0 , the x -value would change because other conductors have begun to conduct who are. The maximum of positive current which flows across the junction is then the difference between the currents from the two sides:

$$I_{\text{max}} = I_0 e^{V_0/V} - I_0 \quad (14.14)$$

If the current is a steady flow into the reverse region. Then they have charge the body of the region, where they are eventually annihilated by the majority carriers in the first one. The electrons which are left in the conduction band will be moved by a current of electrons from the source terminal of the negative terminal.

What is done the next instant is Eq. (14.14)'s time. The just as all the other currents rapidly went to negative voltage. But only for A_1 the current increase in size, but the exponential term soon becomes negligible and the negative current never exceeds I_0 . Which makes our assumption is rather valid. That the current I_0 is limited by the rate of recombination current on the side of the junction.

We get through exactly the same analysis for the case of magnetic currents which flows in the junction, first with the potential difference zero, then with a new externally applied voltage, decreased ΔV , yet get again an equation just like (14.14) for the natural conduction. Since the total current is the sum of the current contributed by the two carriers, Eq. (14.14) still applies for the total current provided we identify I_0 as the maximum current which can flow per unit area voltage.

The subexpression of the current of Eq. (14.14) is shown in Fig. 14-10. It shows the typical behavior of small-scale devices such as those used in modern computers. We should remark that Eq. (14.14) is true only for small voltages. For voltage comparable or larger than the natural voltage difference V_0 , there is very little change in the current no longer any exponential equation.

You may remember incidentally that we set exactly the same equation for the law found here in Eq. 14.14 when we discussed the "ideal diode" in Fig. 14-1. In other and probably simpler terms, we might say the same equations in the two situations about the solid-state processes are quite similar.

14-6 The Transistor

Before the discussion can appreciate class characteristics of the transistor, the junction consists of two semiconductor junctions put close together. The junction is usual in part of the same type, but we just described for the semiconductors made the $p-n-p$ junction. Suppose we take a $n-p-n$ type junction. With this diode junctions, a voltage V , a p -type region, and another n -type region, as shown in Fig. 14-11(a). The combination is called a $p-n-p$ transistor. Each of the two junctions in the transistor can never much in the way we have described in the last section. In particular there will be a potential and field at each junction boundary, which prevent the flow from one side region to the other region. If the two junction regions have the same electrical properties, then it will be extended, as we showed in the circuit in (b) shown in the graph of Fig. 14-11(b).

Now let's imagine that we connect each of the two junctions external to the source as shown in part (c) of Fig. 14-12. We will use all voltages to the terminals connected to the left-hand junction as it will be by definition at zero volts. We will call this terminal the emitter. The n -type region is called the base and it is connected to a slightly negative potential. The right-hand p -type region is called the collector, and is connected to a somewhat larger negative potential. Under these circumstances the variation of potential across the crystal will be as shown in the graph of Fig. 14-12(c).

Let's first see what happens to the positive carriers, since it is primarily their behavior which controls the operation of the $p-n-p$ transistor. Since the source is

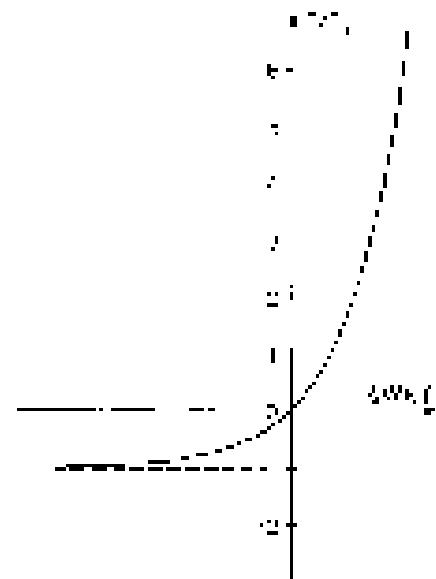


Fig. 14-10. The current density J as a function of the voltage V above V_0 .

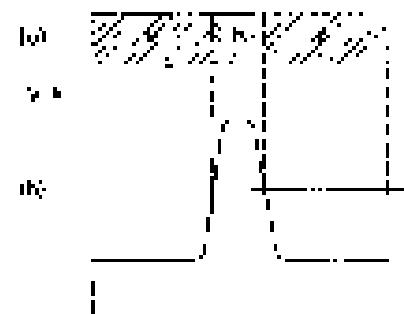


Fig. 14-11. The potential distribution in a diode with no applied voltage.

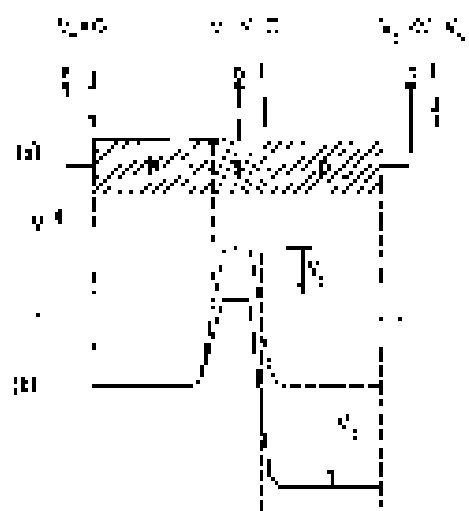


Fig. 14-12. The potential distribution in a $p-n-p$ transistor in operating condition.

at a relatively small positive potential than the base, a current of positive carriers will flow from the emitter region into the base region. A relatively large current flows out between emitter region and collector ("forward voltage") corresponding to the right-hand half of the graph in Fig. 14-11. Now, however, since positive carriers do have to be "brought" from the n -type region into the p -type region, they might think the thickness of the boundary of the n -type region through the base terminal b . Now, however, comes the point of the transistor. The collector region is made very thin, typically 10^{-4} cm or less, much thinner than its base-region dimension. This means that as the holes move toward the region b , there is a very good chance of diffusing across to the other junction because they are annihilated by the electrons in the n -type region. When they get to the uppermost boundary of the n -type region they find a sharp downward potential hill and immediately fall into the right-hand p -type region. This is what the crystal is called the "collector," because it "collects" the holes after they have diffused across the n -type region. In a typical transistor, all but a few tenths of a percent of the hole current which leaves the emitter and enters the base is confined in the collector region, and only the small remainder contributes to the base-emitter current. The sum of the base and collector currents is, of course, equal to the total current.

Now we are ready to see what happens if we apply a negative bias to the base terminal. Since we are at a relatively steady state of the curve of Fig. 14-11, a small variation in the potential V_b will cause a relative percentage change in the collector current I_c . Since the collector voltage V_c is much more negative than the base voltage, these slight variations in potential will not affect appreciably the steep potential hill between the base and the collector. Most of the positive carriers emitted in the negative V_b still reach right by the collector. Thus, as we vary the potential of the base electrode, there will be a corresponding variation in the collector current I_c . The essential point, however, is this: the base current is always zero for a small fraction of the collector current. The resistance is an absolute; a small increase is compensated by the base electrons giving a large current (100 times higher) at the collector electrode.

We return now to the diagram—the way it is when the n -type base has been modulating so far. First, note that we do not expect any significant electron current to flow between the base and the collector. With a large negative voltage on the collector, electrons in the base would have to diffuse over high potential steps, but due to the probability of doing that is very small. There is a very small current of electrons to the collector.

On the other hand, the electrons in the base can drift into the emitter region. In fact, you might expect an electron current in this direction to be proportional to the hole current from the emitter into the base. Since an electron current is useful, and, in fact, working, is bad because it increases the total base current required for a given current of holes in the collector. The transistor is, therefore, designed to minimize the electron current to the emitter. The electron current is proportional to $(N_A/V_b)^{1/2}$, the density of negative carriers in the base region, while the hole current from Amedeo Cesalpino (Noyce et al.), the density of positive carriers in the emitter region. By using relatively thin ribbons—the n -type arsenic N_A (base) can be made much smaller than N_A (emitter). The very thin base region also helps a great deal because the spacing d of the links in this region by the collector increases significantly to average link spacing d from the center into the base, while leaving the electron current unchanged. As a result, if d is the electron current during the intermediate period, it can be made relatively short. In fact, in fact, when the electrons do play a very significant role in operation of the $n-p-n$ transistor, the current is terminated by removal of the holes, and the $n-p-n$ performs as an amplifier as we have described above.

It is also possible to make a transistor by interchanging the p -type and n -type materials in Fig. 14-11. Then we have what is called an $n-p-n$ transistor. In this a p -electrode, the main current is carried by the electrons which flow from the emitter into the base and from there to the collector. Of course, all the arguments we have made for the $p-n-p$ transistor also apply to the $n-p-n$ transistor if the potentials of the electrodes are chosen with the opposite signs.

The Independent Particle Approximation

15-1 Spin waves

In Chapter 10 we worked out the theory for the propagation of an electron or of some other "particle" such as a atomic excitation. In Chap. 11 crystal lattices. In the last chapter we applied the theory to ferromagnets. But when we talked about interactions in which there are many electrons we disregarded any interaction between them. To do this is of course only an approximation. In this chapter we will extend further the idea that you can disregard the interaction between the electrons. We will also see the opportunity to show you some more applications of the theory of the propagation of particles, since we will generally continue to disregard the interactions between particles, this is not unlike what we did in this chapter except for one new application. The first example we'll consider is, however, one in which it is possible to write down quite exactly the exact equations whose case is more than one "particle" present. From them we'll be able to see how the approximation of disregarding the interactions is made. We will, though, analyze the problem carefully.

As our first example we'll consider a "spin wave" in ferromagnetics a solid. We have discussed the theory of paramagnetism in Chapter 9 of Volume II. At zero temperature all the electron spins that contribute to the magnetism in the body are ferromagnetic and are parallel. There is an interaction energy between the spins, which is lower when all the spins are parallel. At any nonzero temperature, however, there is a chance that some of the spins are randomized. We calculate the probability in a approximate manner in Chapter 26. This time we will describe the quantum-mechanical theory—so you will see what you would have to do if you wanted to solve the problem more exactly. We will still make some idealizations by assuming the electron moves fast and far enough that the spin is always anti-parallel to neighboring spins.

The condition in which the electrons at each atom are all paired except one, so that all of the magnetic effects come from one soliton than one atom. Further, we imagine that these electrons are localized in small boxes in the lattice. The model can easily be thought of as a jello.

We also assume that there is an interaction between the two adjacent atoms; electrons which move in the energy of hexagons:

$$E = - \sum_{\langle i,j \rangle} \delta \sigma_i \cdot \sigma_j \quad (15.1)$$

where σ 's represent the spins and the summation is over all adjacent pairs of electrons. We have already discussed this kind of interaction energy when we considered the hyperfine splitting of the magnetism of the nucleus of the hydrogen atom and proton in a hydrogen atom. We represent it then as $\delta \sigma_i \cdot \sigma_j$. Now, for a given pair, say the electrons at atom 1 and atom 3, the Hamiltonian would be $-K \sigma_1 \cdot \sigma_3$. We have σ to indicate pair, and the Hamiltonian is the sum of the energies of each pair of these terms for each interacting pair. The energy is written with the factor $-K$ so that a positive K will correspond to ferromagnetism, that is, the lowest energy results when opposite spins are parallel. In a real crystal, there may be other terms which are the interactions of two nearest neighbors, and so on, but we don't need to consider such complications in this stage.

With the Hamiltonian of Eq. (1), we have a complete description of the system. With our approximation—the properties of the magnetization

15-2 Spin waves

15-3 Independent particle

15-4 The benzene molecule

15-5 More magnetic chemistry

15-6 Other uses of the approximation

should do next. We would also be able to calculate the thermodynamic properties due to the magnetization. If we can find all the energy levels, the probability of the system at temperature T can be found from the principle that the probability that a system will be found in a given state of energy E is proportional to $e^{-E/T}$. This problem has never been completely solved.

We will solve some of the problems by using a simple example in which all three electrons are in a line—a one-dimensional lattice. You can easily extend the idea to three dimensions. At each atomic location there is an electron which has two possible states, either spin up or spin down, and the whole system is described by “flipping over” of the spins as it moves. We take the Hamiltonian of the system to be the operator of the interaction energy. The picture of the spin sectors of Eq. (15.1) is the sign consideration for the signs in the two-electron Heisenberg \hat{H} :

$$\hat{H} = \sum_{i,j} -\frac{A}{2} \hat{\sigma}_i \cdot \hat{\sigma}_j, \quad (15.2)$$

In this equation we have written the constant as $A/2$ for convenience so that some of the later equations will be exactly the same as the ones in Chapter 11.

Now what is the lowest state of this system? The state of lowest energy is the one in which all the spins are parallel, let's say, all up. We can make this state $\psi_0 = |\uparrow\uparrow\uparrow\rangle$, where the “up” symbol is repeated three times. It is easy to figure out the energy for this state. One way is to write out all the cubic terms in terms of $\hat{\sigma}_x$, $\hat{\sigma}_y$, and $\hat{\sigma}_z$, and work through carefully what each term of the Hamiltonian does to the ground state, and then add the results. We can, however, also use a quicker method. We saw in Section 14-2 that $\hat{\sigma}_x$, $\hat{\sigma}_y$, and $\hat{\sigma}_z$ could be written in terms of the Pauli spin exchange operator, like this:

$$\hat{\sigma}_i \cdot \hat{\sigma}_j = i \hat{\sigma}_{ij}^{(1/2)} - 1, \quad (15.3)$$

where the operator $i \hat{\sigma}_{ij}^{(1/2)}$ interchanges the spins of the i th and j th electrons. With this substitution, the Hamiltonian becomes

$$\hat{H} = -A \sum_{i,j} (\hat{\sigma}_{ij}^{(1/2)})^2 - \beta. \quad (15.4)$$

It is now easy to work out what happens in different states. For instance if i and j are both up, then exchanging the spins leaves everything unchanged on $\hat{\sigma}_{ij}^{(1/2)}$, acting on the state just gives the same state back, and is equivalent to multiplying by -1 . The expression $(\hat{\sigma}_{ij}^{(1/2)})^2 = 1$ is just right. In one-half, i there are no we will leave off the recursive superscripts on the $\hat{\sigma}$'s.

For the ground state of spins one and so if you exchange a particular pair of spins you get back the original state. For you get zero in this energy case. If you operate on α with the Hamiltonian you get the state α again multiplied by a sum of terms, $-(A/2)$, for each pair of spins. That is, the energy of the system in the ground state is $-A/2$ per atom.

Next we should like to look at the energies of some of the excited states. It will be convenient to measure the energies with respect to the ground state—the reference ground state has zero energy. We can do this by adding the energy $1/2$ to each term in the Hamiltonian. Each just changes the “ β ” in Eq. (15.4) to “ β_+ .” Our new Hamiltonian is

$$\hat{H}' = -A \sum_{i,j} \hat{\sigma}_{ij}^{(1/2)} + \beta_+. \quad (15.5)$$

With this Hamiltonian the energy of the lowest state is zero. The spin exchange operator is now taken by multiplying by a unity $(1 - \hat{\sigma}_{ij}^{(1/2)})$, and this is canceled by the “ $\hat{\sigma}_{ij}^{(1/2)}$ ” each term.

The ground state here is not “degenerate”; there are no two states with the same energy—for example, all spins down, or all three other directions. The original reference level is the excited state with all up, and the excited state with all down are the only other two energy possibilities.

For decreasing states other than the ground state you will need a subspace $\mathcal{H}_{\text{down}}$. One convenient approach is to group the states according to whether the electron has spin down, or has two down spins, or up, and so on. There are, of course, many states with one spin down. The down spin could be at atom "A," or at atom "B," or at atom "C." We can, in fact, choose just such states for our basis states. We could write them as $|x_1\rangle \otimes |S_1\rangle \otimes |S_2\rangle \dots$ It will, however, be more cumbersome later on to put the total quantum number with the downspinning electron by its coordinate x_i . That is, we'll define the state $|x_1\rangle$ to be one with all the atoms nonspinning up except for the one at A , atom x_1 , which has a downspinning electron (see Fig. 15-1). In general, $|x\rangle$ is the state with one down spin that is located at the coordinate x_1 , x_2 , etc.

Next, let's act on one of the basis function $|S_1\rangle$ on the state $|x\rangle$? One term of the Hamiltonian is say $= A(S_{1,A} - 1)$. The operator $S_{1,A}$ exchanges the two spins of the adjacent atoms 1, 2. But, if the state $|x\rangle$ has $x_1 = 1$, nothing happens; $S_{1,A}$ is equivalent to multiplying by 1:

$$S_{1,A}|x\rangle = |x\rangle$$

It follows that

$$(S_{1,A} - 1)|x\rangle = 0$$

Thus all the terms of the Hamiltonian give zero—except those involving $x_1 = 1$, of course. On the state $|x\rangle$, the operation $S_{1,A}$ exchanges the spin of atom 1 (up) and atom 2 (down). The result is the state with all spins up except the atom at 1, that is

$$S_{1,A}|x\rangle = |x_1\rangle$$

In the same way

$$S_{2,A}|x\rangle = |x_2\rangle$$

Hence, the only terms of the Hamiltonian which survive are $= A(S_{1,A} - 1)$ and $= A(S_{2,A} - 1)$. Acting on $|x\rangle$, they produce $= A|x_1\rangle + A|x_2\rangle$ and $= A|x_1 + 1\rangle + A|x_2\rangle$, respectively. The result is

$$\hat{H}|x\rangle = -A \sum_i (S_{i,A} - 1)|x\rangle = -A(x_1 + x_2 - 2)|x\rangle. \quad (15.6)$$

When the Hamiltonian acts on $|x\rangle$, it gives me the same amplitude to be in states $|x_1\rangle$ and $|x_2\rangle$. That just means that there is a central coupling to have the down spin flip over to the next atom. So because of the interaction between spins, if we begin with one spin down, there must be some probability that an electron whose one will be down instead. Operating on the general state $|x\rangle$, the Hamiltonian gives

$$\hat{H}|x\rangle = -A(|x_1\rangle - |x_2\rangle) - 2|x\rangle. \quad (15.7)$$

Notice particularly that if we take a complete set of atoms with only one spin down, they will only be mixed among themselves. The Hamiltonian will never mix these states with others that have more spins down. So long as you only exchange spins, you never change the total number of down spins.

If we're going to use the matrix equation for the Hamiltonian, we find $(x_1, H|x\rangle, x_2)$ (Eq. 15.7) is equivalent to

$$\begin{aligned} H_{1,1} &= A; \\ H_{1,2} &= H_{2,1} = -A; \\ H_{2,2} &= 0, \quad \text{for } |x_1 - x_2| > 1. \end{aligned} \quad (15.8)$$

Now what are the energy levels for states with one spin down? As usual we let C_x be the amplitude that some state $|p\rangle$ is in the basis $|x\rangle$. If $|p\rangle$ is to be a definite energy state, all the C_x 's must vary with time in the same way, namely,

$$C_x = S_x e^{iE_p t/\hbar}. \quad (15.9)$$

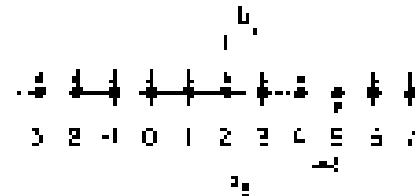


Fig. 15-1. The basis state $|x_1\rangle$ of a chain of three atoms. All the spins are up except the one at x_1 , which is down.

We can put the field ψ on into one- and two-millimeter regions:

$$\psi \left(\frac{2\pi}{3} \right) = \sum_{k=1}^{\infty} (L_k C_k) e^{ikx_1} \quad (15.1)$$

Using Eq. (15.1), we can multiply elements. Of course we get an infinite number of terms, but they cancel by symmetry:

$$C_{k+1} = C_{k+2} = \dots = C_{k+L} = 0 \quad (15.2)$$

We know again exactly what some particular set of coefficients in the wave function ψ tell us because C_k are the wave functions of the k -th mode, which propagates along the lattice with a propagation constant L_k and an energy

$$E_k = L_k C_k = \cos(kL) \quad (15.3)$$

where L is the lattice constant.

The different energy stations correspond to "waves" of even spin ("odd spin waves") and for each such mode there is a constant power. For large scattering by (and thus) inelasticity comes in:

$$E = \omega / 2 \pi \quad (15.4)$$

In this section we can consider a localized wave packet (frontaining, however, only long wavelength) which represents a single mode of motion in one part of the lattice. This domain will behave like a "particle." Because its energy is related to λ via (15.3) this "particle" will have in all directions:

$$m_\lambda = \frac{\lambda^2}{2\pi c} \quad (15.5)$$

15-2 Two-spin waves



Fig. 15-2. A wave with two spins

Note that work must be done prior to parts of these two sections (part A against the previous two parts). We'll discuss states in which there are two spin- $\frac{1}{2}$ fermionic particles, such as the state shown in Fig. 15-2. We can label such a state by the wavefunction in the two spin-wave directions. The two shown can be called $| \psi_1, \psi_2 \rangle$. In general the basis states are $| \psi_1, \psi_2, \psi_3 \rangle$, etc., fully defined with the direction of the spin. If the state is $|\psi_1, \psi_2, \dots, \psi_n\rangle$ it is very likely to be made up of several components of different spin up and down at different sites according to the order. Furthermore, the state $|\psi_1, \psi_2, \dots, \psi_n\rangle$ has no meaning, there isn't such a thing. We can do better by writing $|\psi_1, \psi_2, \dots, \psi_n\rangle$ as a superposition of basis states, $|\psi_1, \psi_2, \dots, \psi_n\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \dots \otimes |\psi_n\rangle$. Now means the amplitude for a given state ψ_i to be in a certain place depends both on the spin and with atoms have a dual spin. The wavefunctions below show two nice examples of ideas. They are much simpler than in mechanics. (Remember the completeness of quantum mechanics is just one weak point. Here there are more degrees of freedom, the condition becomes more and more delicate with lots of other and the equal ones always look very similar, p. but the ideas are not necessarily more complicated than in the simplified case.)

The equations of motion of this spin system are the differential equations for the $C_{\mu k}$. That is,

$$i\hbar \frac{dC_{\mu k}}{dt} = \sum_{\nu} (H_{\mu\nu} - \mu \delta_{\mu\nu}) C_{\nu k} \quad (15.6)$$

Suppose we want to find the stationary states. As usual, we begin here with no motion; time becomes & time the spin angles $\theta_{\mu k}$ and $\phi_{\mu k}$ will be replaced by $\theta_{\mu k}$ and $\phi_{\mu k}$.

coefficients $\alpha_{\pm k}$. And we have to work out carefully the effect of H on a state with spins σ_1 and σ_2 , since H is not linear to figure out. Suppose for a moment that spins σ_1 & σ_2 are strong enough that we don't need to worry about the other constituents. The operator of exchange in the hamiltonian σ_z , will move the down spin either to the $k=+1$ or $k=-1$ room, and so there's an amplitude that the present coefficients come from the state $|\downarrow\downarrow\rangle$, right? And also an amplitude that it has come from the state $|\uparrow\uparrow\rangle$? It may have been the upper spin that moved, so there's a certain amplitude that $C_{\sigma_1\sigma_2}$ is fed from $C_{\downarrow\downarrow}$ or from $C_{\uparrow\uparrow}$. These effects should all be equal. The final result for the Heisenberg equation is Eq. 5.5

$$\dot{C}_{\sigma_1\sigma_2} = -iA(\alpha_{k=+1} + \alpha_{k=-1} + \alpha_{k=+2} + \alpha_{k=-2}) + i\epsilon C_{\sigma_1\sigma_2} \quad (5.16)$$

This equation is *exactly* correct in our situation. If $\sigma_1 = \sigma_2$ there is no spin flip, and if $\sigma_1 = \sigma_2 = 1$, then two of the terms in Eq. (5.16) should be missing. If we are going to diagonalize this equation, we simply ignore the last two terms of these equations & slightly alter A so that C is supposed to be infinite, and we have an infinite number of terms; neglecting a few might not matter much. So for a first rough approximation let's forget about the dotted equation. In other words, we assume that \tilde{T}_1 (Eq. 5.6) is true for σ_1 and σ_2 , even if M and N are next to each other. This is the condition given of our approximations.

Then the equation is not hard to find. We get immediately

$$C_{\sigma_1\sigma_2} = e^{-iE_{\sigma_1\sigma_2}t/\hbar} \quad (5.17)$$

with

$$\epsilon_{\sigma_1\sigma_2} = (\text{polar } \sigma_1)^{\frac{1}{2}} \text{polar } \sigma_2^{\frac{1}{2}} \quad (5.18)$$

where

$$E = 4M + 24 \cos(k_M) - 24 \cos(k_N) \quad (5.19)$$

think for a moment what would happen if we had two independent single spin waves like in the previous section corresponding to $\sigma_1 = \sigma_2$ and $\sigma_1 = -\sigma_2$; they would have merged. From Eq. (5.12), of

$$E_1 = (3M + 24 \cos(k_M))$$

and

$$E_2 = (3M - 24 \cos(k_M))$$

Notice that the energy E in Eq. (5.19) is just $E_1 + E_2$.

$$E = 6M + 24 \cos(k_M) \quad (5.20)$$

In other words we can think of our solution in this way: objects are two particles (that is, two spin waves). One of them has a momentum described by k_M , the other by k_N and the energy of the system is the sum of the energies of the two objects. The two particles are completely independent — that's all there is to it.

Of course we have to assume some approximations, but we do not wish to discuss the precision of our answer at this point. However, you might guess that it is reasonable to expect with billiards or something-and, therefore, by force of mass in the Hamiltonian, keeping out a ΔE to not wouldn't make much of an issue. If we are ϵ -energy downspins, then there was an appropriate constant ϵ , then we would certainly have to worry about the perturbations.

Interestingly enough, an exact solution can be written down if there are just the two down spins. The result is not particularly impressive. But it is interesting that an exact form can be solved exactly for this case. The solution is

$$C_{\sigma_1\sigma_2} = \exp(iE_{\sigma_1\sigma_2}t/\hbar) \delta_{\sigma_1\sigma_2} - \delta_{\sigma_1\sigma_2} \quad (5.21)$$

with the energy

$$E = 4M - 24 \cos(k_M) - 24 \cos(k_N)$$

and with the wave numbers k_1 and k_2 related to k_x and k_y by

$$k_1 = k_x + k_y, \quad k_2 = k_y - k_x = 0. \quad (15.22)$$

This section includes the “interaction” of the two spins. It describes the fact that when the spins come together there is a certain chance of scattering. The spins act very much like particles with an interaction. But the detailed theory of their scattering goes beyond what we want to talk about now.

15.9 Independent particles

In the last section we wrote down a homogeneous eq. (15.9), for a two-particle system. Then, using an approximation which is equivalent to neglecting any “interaction” of the two particles, we found the following state, described by Eqs. (15.17) and (15.18). This state is just the product of two single-particle states—the solution we have given for $\psi_{\alpha_1, \alpha_2}$ in Eq. (15.19) is, however, really not satisfactory. We have very carefully pointed out earlier that the state $\psi_{\alpha_1, \alpha_2}$ is not a different state from $\psi_{\alpha_2, \alpha_1}$; the order of α_1 and α_2 has no significance. In general, the physical interpretation for the amplitude C_{α_1, α_2} must be unchanged if we interchange the values of α_1 and α_2 , and this does not happen with $\psi_{\alpha_1, \alpha_2}$. Either very foolishly represent the amplitude to find a down-spin up-spin and an up-spin down-spin, or else make it symmetric in α_1 and α_2 . Since k_1 and k_2 can in general be different,

The trouble is that we have not solved our equation of Eq. (15.17) to satisfy this additional condition. Fortunately it is easy to fix this up. Because the solution of the Dirac equation just as given in § 15.5 is

$$\psi_{\alpha, k} = R e^{i k_x x} e^{i k_y y}, \quad (15.23)$$

it even has the same energy we got for (15.22). Any linear combination of (15.23) and (15.22) is also a good solution and has no energy still given by Eq. (15.19). The solution we should have chosen—because of our symmetry requirement—is just the sum of (15.17) and (15.22),

$$\psi_{\alpha_1, \alpha_2} = R e^{i k_x x_1} e^{i k_y y_1} + R e^{i k_x x_2} e^{i k_y y_2}. \quad (15.24)$$

Now, given any k_1 and k_2 the amplitude C_{α_1, α_2} is independent of which way we put x_1 and x_2 . If we decide happen to define x_1 and x_2 reversed we get the same amplitude. Our interpretation of (15.24) in terms of “independent” must also be different. We can no longer say that the equation represents one particle with two summed α_1 and a second particle with wave number k_2 . The amplitude (15.24) represents one state with two particles (magnetic). The state is characterized by the two wave numbers k_1 and k_2 . One solution looks like a composite state of one particle with the momentum $p_1 = \hbar k_1$ and another particle with the momentum $p_2 = \hbar k_2$, but... our wave we can't say which particle is which.

By now this discussion should remind you of Chapter 4 and of a story of identical particles. We can just keep crossing out the particles of the spin system the magnetic, because like electrons these particles. A magnetic must be symmetric in the wave numbers of the two particles. What is this same as saying that if we “interchange the two particles” we get back the same amplitude after ± 0 the same sign. But you may be thinking why did we have to add the two terms in making Eq. (15.24). Why not subtract? Well, it might appear that changing x_1 and x_2 would just change the sign of $\psi_{\alpha_1, \alpha_2}$, which doesn't matter. But after changing x_1 and x_2 doesn't change anything—all the elements of the matrix are exactly where they were before, so there is no reason for even the sign of the amplitude to change. The magnetic will be one like these particles!†

† In general, the same particles of one kind won't. Decaying may not like either two fermions or two particles, and as for free particles, the particles will interact with each other while being almost unbound. This “interact” is not the same as the interaction with time. The system is open to it. That means that the entire system is not closed.

The wave pairs of the discussion have been twofold. First we show you something about spin waves, and second we demonstrate a state whose amplitude is a product of two amplitudes, and whose energy is the sum of the energies corresponding to the two amplitudes. For a two-wave packet the amplitude is the product and the energy is the sum. This is exactly what the energy is in the sum. The energy is the sum of the two energy exponentials— $e^{i\omega_1 t}$ plus $e^{i\omega_2 t}$. If two objects are doing something, one of them with the amplitude $e^{i\omega_1 t}$ and the other with the amplitude $e^{i\omega_2 t}$, and if the amplitude for the two things add up together, is the product of the amplitudes? No, even then there is a single frequency in the product which is the sum of the two frequencies. The energy corresponding to the amplitude is proportional to the sum of the two energies.

We have gone through another long-winded argument to tell you a simple thing. When you don't take into account any interaction between particles, you can think of each as being independent. They could potentially exist in two different states; they would then interact, and they will each contribute to energy; they would have had it they were alone. However, suppose we take the ψ . If they are identical particles, they may behave either as bosons or as Fermi particles, depending upon the quantum. Two boson electrons added to a system, for instance, would tend to behave like Fermi particles. When the positions of the electrons are interchanged, the amplitude would reverse sign. In this equivalent situation, a pair of $(1, -1)$ terms would have to be added as sign between the two terms in the total. As a consequence, over there the contribution to energy for one electron will be the same as for the other. The amplitude for this state is zero.

15-4 The benzene molecule

Although quantum mechanics provides the basis from which determine the structures of molecules, there has not yet applied exactly such a detailed description. The chemists have, therefore, worked out various approximate methods for calculating some of the properties of complicated molecules. We would now like to show you how the independent-particle approximation is used to the structures. We begin with the benzene molecule.

We also used the benzene molecule from another point of view in Chapter 12. There we took an s -approaching picture of the molecule at its lowest energy, with the wave functions shown in Fig. 12-3-2. There is a ring of six electrons with a hydrogen attached to the carbon at each head-on. With the conventional picture of valence bonds, it is necessary to use the multiple bonds between half of the carbons, and in the lowest energy conduction bands, we see something like shown in Fig. 12-3. There are also other higher-energy states. When we computed them in Chapter 12, we just took the two states and forgot all the rest. We found that the ground-state energy of the molecule was not the energy of one of the states, the true, but our lower than that by an order proportional to the amplitude of the transition of those states to the rest.

Now we're going to turn to the same molecule from a completely different point of view, and get kind of approximation. The two points of view will give us different answers, but if we include either approximation, it would lead to the true, realistic description of benzene. However, if we don't let in to improve them, which is of course the usual situation, then you can end up as surprised if the two descriptions do not agree exactly. We shall attempt to do this also with our new point-of-view: the lowest energy of the benzene molecule is low, that is, of the three bond structures of Fig. 12-2,

Now we want to use the bonding picture. Suppose we imagine the six electrons of a benzene molecule concerned only by what has done at Fig. 15-1. We have received six electrons— $-\psi_1$ a bond stands for a pair of electrons.

So we have a distance function for this molecule. Now we will consider what happens when we put back the six electrons one at a time, imagining that each one can run freely around the ring. We assume also that all the bonds shown in Fig. 15-4 are satisfied, and don't need to be established further.

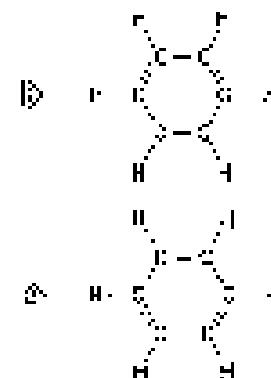


Fig. 15-3. The two base states for the benzene molecule. (1) unpaired

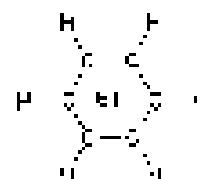


Fig. 15-4. A benzene ring with six electrons removed.



Fig. 15-5. The ethylene molecule.

What happens when we put one electron back into the molecule, and the other, of course, be located in just one of the six positions around the ring—corresponding to the lone states. It would also have a certain amplitude, say ψ_1 , to go from one position to the next. If we analyze the stationary states, it's possible to obtain possible energy levels. That's what we're going to do.

Next put a second electron in. And now we make the most ridiculous approximation that you can think of—that the electrons don't affect each other. Well, we know it's false. Of course they really will interact; they repel each other through the Coulomb force, and furthermore when they are both in the same spot, they must have considerably different energy than twice the energy for one electron. Certainly the expectation value of $\psi_1 \psi_2$ is not zero, because when there are only six sites, particularly when we want to put in an electron. Nevertheless, the approximate have been able to learn a lot by taking this kind of an approximation.

Before we work out the harmonic oscillator, let's consider a simpler example—the ethylene molecule which contains just two carbon atoms with two hydrogen atoms on either side as shown in Fig. 15-5. This molecule has four valence orbitals involving two electrons in each of the two carbon atoms. Now remove one of these electrons. What do we have? We still have just a two-state system, the remaining electron can be at the carbon or the other. We can analyze this as a two-state system. The possible energies for the single electron are either $E_0 = -4$ or $(E_0 + 4)$ as shown in Fig. 15-6.

Now add the second electron. Good, if we have two electrons, we can put the first one in the lower state and the second one in the upper. Not quite we forgot something. Each one of the values is really double. When we say there's a possible state with the energy E_0 , -4 , there are really two. Two electrons can go into the same state if one has to spin up and the other, spin down. This cannot be put in because of the exclusion principle. So there really are two possible states of energy $(E_0 - 4)$. We can draw a diagram as in Fig. 15-7, which indicates both the energy levels and their occupancy. In the condition of lowest energy both electrons will go in the lowest state with their spins opposite. The energy of the extra bond in the ethylene molecule is then is $2E_0$. Why? We neglect the distance between the two electrons.

Let's get back to the harmonic. Now we are two states of Fig. 15-6 has three contributions. Just as there is just one bond in ethylene, and contributes $2(\delta_1 - 4)$ to the energy if E_0 is now the energy is just one electron on one in the one and it is the amplitude to flip it. In this case, the energy due to it is roughly $2(E_0 - 4)$. When we studied benzene before, we got that the energy was lower than the energy of the structure with three extra bonds. That's all the energy the benzene excess is known. But this comes from a new point of view.

We shall now do the six times reduced harmonic, and add one electron. Now we have a benzene system. We haven't drawn one yet, but we know what it is. We can write six electrons in the six napoleons, and so on. Let's save some work by realizing that we've already solved the problem, when we worked out the problem of finding the six ψ 's in the first line of terms. Of course, the second line is an ψ in ψ , it has a nature very similar. But imagine, if I took an ψ and the ψ to a line, and summed the ψ 's along the line from 1 to 6, it is an ψ in the next ψ would be ψ in ψ in ψ etc., which is the same as the original ψ . One reason the situation will be very like the benzene ring. In other words, we can take the solution for ψ , infinite line set in addition to the solution must be periodic with a cycle of six along. From Chapter 13 the idea being, in fact, that ψ is of finite energy when the amplitude at each site is finite. And each ψ the energy is

$$E = E_0 - 2A \cos \frac{\pi k}{6} \quad (15-2)$$

We want to end now with the law of sums which repeat every 6 atoms. Let's find the general case for a ring of N atoms. If the solution is to have a period 12-4

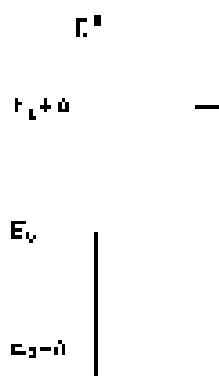


Fig. 15-6. The possible energy levels for the "extra" electrons in the ethylene molecule.

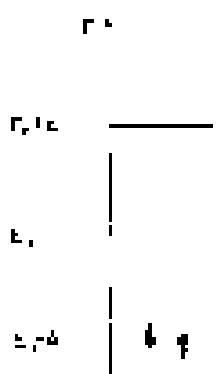


Fig. 15-7. In the extra bond of the ethylene molecule two electrons, one spin up, one spin down, can occupy the lowest energy level.

of N atomic species, each may be unity or it may be a multiple of $\frac{1}{2}$. Taking $\psi = \pm \sqrt{\lambda}$ instead, the conclusion is that

$$\langle \psi | \psi \rangle = N_{\text{sp}} \quad (15.26)$$

We have seen before that there is no meaning to taking ψ 's outside the square root. This means that we get all possible states by taking values of ψ in the range $= \pm \sqrt{\lambda}$.

We find also that N_{sp} in Nalor's case there are six degenerate energy states and they have wave numbers k_x given by

$$k_x = \frac{2\pi}{N_{\text{sp}} a} \quad (15.27)$$

Each state has the energy (15.25). We have a fine spectrum of possible energy levels. The seven numbers $(N_{\text{sp}} = 4)$ is shown in Fig. 15.8(b); the numbers in parentheses indicate the number of different states with the same energy.

There's a nice way to visualize the six energy levels we've been discussing (Fig. 15.8c). Imagine a molecule centered on a level with E_1 , and with a radius of $a/2$. If we start at the bottom and work our way up to a point just below the vertical height of $2\pi/a$ from the center, we'll end up at E_2 (Fig. 15.8c). The six points represent the six possible states. The lowest-energy level is at $(E_1 - 24)$; there are two states with the same energy ($E_1 - 12$), and so on. These are possible energy states for one electron. If we have more than one electron, two with opposite spins can go into each one.

For the benzene molecule we have to put in six electrons. For the ground state they will go into the lowest possible energy states: one at $\epsilon = 0$, two at $\epsilon = -12$, and two at $\epsilon = -24$. According to the independent-particle approximation the energy of the ground state is

$$E_{\text{total}} = N(E_1 - 24) + 2(E_1 - 12) + 2(E_1 - 0) \\ = 6E_1 - 84 \quad (15.28)$$

The energy is increased less than that of three separate double bonds—by the account of 24.

By comparing the energy of benzene to the energy of ethylene it is possible to determine A . It comes out to be 0.4 electron volt, or, in old units the chevalier libri, 14 kilocalories per mole.

We can use this description to calculate or understand other properties of benzene. For example, using Fig. 15.8b we can discuss the excitation of benzene by light. What would happen if we tried to excite one of the electrons? It could move up to one of the empty higher states. The lowest energy of excitation would be a transition from the highest-filled level to the lowest-empty level. That is at the energy 24. Benzene will absorb light of frequency ν when $\hbar\nu = 24$. Light will also be absorbed with the energies 12 and 4. (Indeed, in one of the absorption spectra of benzene has been measured and the pattern of spectral lines is more or less constant except that the lowest transition occurs in the ultraviolet and its frequency would have a definite value of 4 between 12 and 24 electron volts.) That is, the numerical value of A is two or three times larger than is predicted from the theory of bonding energy.

When we calculate these quantities like this it is usually more convenient of a simpler form and you won't remember much. In fact, for example, for calculating bonding energy use such and such a value of A , but for getting the absorption spectrum upon relatively light use another value of A . You may feel

[†] You might think that $2\pi/N$ is even number. Here are $N = 1$ states. That is not so because $\psi = \pm \sqrt{\lambda}$ give the same ψ .

[‡] When there are two states (which will have different amplitude of overlap) with the same energy, we say that the two states are "degenerate." Notice that your electrons can have the same $E_1 - A$.

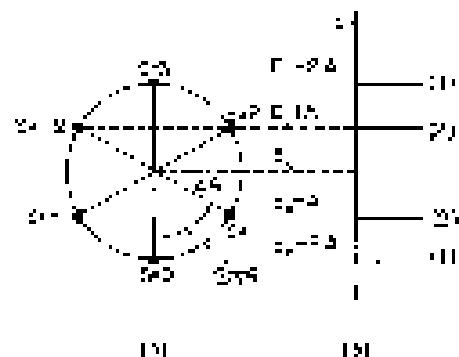


Fig. 15.8. The energy levels belonging to the molecule of benzene. (a) The hexagonal ring with six vertices (levels). (b) The energy levels belonging to the molecule of benzene. (c) The circle with radius $a/2$ centered at E_1 .

But this seems a little absurd. It is not easy to justify from the point of view of a physicist who is trying to understand the world from first principles. But the problem of the chemist is different. He must try to predict what will happen when he is going to heat up a substance so that it may decompose, or in which order various substances will decompose. Now he needs to know of several types of reaction. It doesn't make much difference where they come from. So we use the theory in a site... To satisfy them the physicist. He likes to have theories as abstractions of the facts in them, but then he must always be concrete in them - making experiments.

In the case of human life the "principles" used in the interventionary is not the simpler theory that all elements are independent—the theory we started with is really not legitimate. Nevertheless, it has some shadow of the truth because it results in more or less action in the right direction. We know approximate place where organized sub-environment varies (especially the *organic* climate) takes us way through the rounds of complicated things it cannot be easy. (Dunlop says that the reason a physician can easily calculate from his principles is that he chooses only simple problems. He never comes to a situation with 40 or even 50 variables in it. So far he has been able to calculate reasonably accurately only the 10 or 12 or 15 or 20 minimum items.)

15.5. מושג המוקד: דינמיות

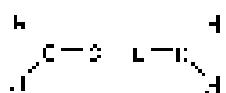


Fig. 14-2 The valence bond approach to the bonding in hydrides. (1, 2)



Fig. 15-18. A tree of *Vitis cordata*.

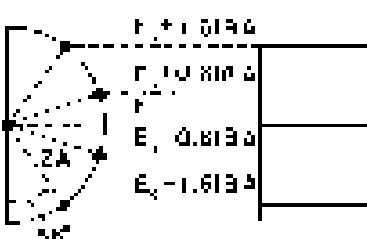


Fig. 15-11. The energy level diagram for the hydrogen atom.

We can play the same game with the wave function except along the two center lines. If we move from left to right, the center is on a line that already corresponds to index 1, Eq. (2), (which I only asked you to solve for infinite λ). But the solution is not unique, we can include the one from Eq. (1). Watch! Now, let me add a few terms on the line and remove them from the basis shown in Fig. 15-10. In writing the equations for the magnitude of position 1 you would not have to be in trouble, but in trouble? Similarly, the equation for position 2 would differ from before, but we need not be in trouble, the rest of time would be nothing new if position 2 is 1, but suppose that we can't get it to zero. In the infinite case we have following procedure, the sum should go to zero even if it is not zero, the sum of the absolute value of the coefficients from the equations from Eqs. 9 or the finite lines 1, 2 to avoided. You might think no such solution exists for a finite line because all the terms all looked like $\lambda^{\pm n}$ which has the same absolute value of the amplitude everywhere. But you are wrong because the energy depends only on the absolute value of λ , so the other solution, which is exactly legitimate in the continuity, would be $\lambda^{\pm n}$. And the same argument applies to a superposition of two such solutions. By superposing them we can set the amplitude equal to zero, which requires the requirement that the amplitude be zero at $\lambda = 0$. It will correspond to the energy $E_0 = -24 \text{ meV}$. Now my question is what the value of λ we can choose to get amplitude zero at $\lambda = 0$? This requires that $\psi_0 = -\psi_1$ as a multiple of ψ_0 or that

$$B = \frac{c}{(q-1)} \cdot S \quad (1^{\text{st}} \text{ eq})$$

where i is an integer from 1 to M . (We take only positive N 's because each solution contains a term $-k$; changing the sign of k gives the same set of all even N 's.) Then, the solution involves $\psi = \phi \times \text{Bessel}(kx) \times \text{cos}(k\theta)$.

$$M_{\rm min} = 1.5 \cdot 10^{-3} M_{\odot}, \quad M_{\rm max} = 10 M_{\odot}, \quad M_{\rm crit} = 5 \cdot 10^{-3} M_{\odot}. \quad (15.1)$$

We can repeat the same procedure as above, however since α is now fixed we do not have to. This time we use a semi-infinite interval for input parameter x , shown in Fig. 12-1. The point at the bottom corresponds to $x = 0$. While you can see $\mu(x)$

case is one of the points at the top, which corresponds to $\delta = 1$, the resonance dipole giving us four allowed curves. These are four valence states, such as what we expect having started with a π -electron with the double bond on the angular interval, and $\pi/3$ to $2\pi/3$ degrees. The lowest energy comes from $\delta_1 = -81.46$ eV, while gradually increasing higher, the golden mean of the Chebnet gives us the lowest energy state of the butadiene molecule according to his theory.¹

Now we can calculating the energy of the excited molecule when θ_2 put in two electrons. With θ_2 next time we will try the lowest π -levels each with two electrons of opposite sign. The total energy is

$$E = 3(E_1) + 1(E_2) + 0(E_3) + 0(E_4) = 0 = 0.407 \text{ eV}$$

This result seems reasonable. The energy is much lower than for two single double bonds, but the π -bond is not so strong as in benzene. Anyway, let's fix now the electron density around this molecule.

The chemist can do not only the energies but the probability amplitudes as well. Knowing the amplitudes in each state, one should either are occupied, he can tell the probability of for any π -electron anywhere in the molecule. These states where the electrons are most likely to be found. In the next we investigate substitutions, which explain the π -electrons are shared with some other group of atoms. The reaction rates increase if there is some in these substitutions effect. It's equivalent to yield an extra electron to the system.

The same here we have one π -ring and give us some unpaired energy of a molecule with two π -systems as in dimethyl-pheophytin shown in Fig. 15-12. Now we can see double and single bonds as well as three methyl groups from a large clockwise ring with twenty electrons. Let's to the electrons of the double bonds remain around this ring. Using the π -dependent basis & method we can see a whole set of energy levels. There is a strong absorption band transition between these levels which is in the visible part of the spectrum and give this molecule its coloring role. Similar compounds molecules such as the methoxygiant, which we've discussed, can be studied.² (see p. 300).

There is one more law which emerges from the application of the kind of theory in organic chemistry. It is probably the most accessible of all laws in a certain sense. The most occurs in the first case with the question: In what substances does a particular strong chemical binding? The answer is very interesting. Take the same π -calculus basis and compare the sequence of curves that occurs now, even with the connection of molecules and adiabatic and photo-electron. We would start by bringing all electrons toward the origin as positive. But since we put the charge of the molecule to zero, takes the number of the number of electrons. First, like $\delta_1 = 0$ since we don't know what it is, we get the curve shown in Fig. 15-13. For the first we calculate the slope of the Coulombic potential V . For each successive group of electrons, increases and decreases exponentially in slope between the groups of electrons. The slope changes when one has just finished filling a set of molecules after these electrons and then goes up to the next higher set of levels for the next electrons.

The second change of the behavior looks very quite different from the curve of Fig. 15-13 because of the interaction of the electrons and because of excited states electrons we have two competing. These corrections will however not with error in the seventh view. For if we scale the values of these corrections, the resulting energy of the molecule still have only a few values in which not happen a Coulomb energy level.

Now consider a new seventh curve due to the points of the average for the one drawn in Fig. 15-14. We clearly see that the points above this curve have "quantum" character, and the points below the curve have "classical character".

¹ The role of resonance of existing interest in molecular mechanics and a similar reference.

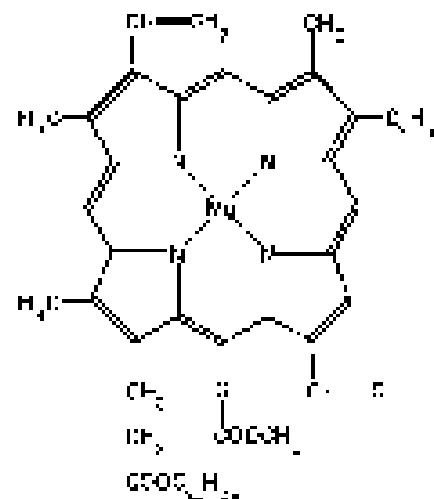


Fig. 15-12. A dimethyl-pheophytin

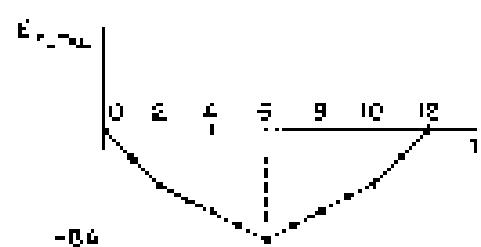


Fig. 15-13. The set of all the electron energies when the lowest value is $E_1 = 0$ and one occupied by a electron. Please note that $\delta_1 = 0$.

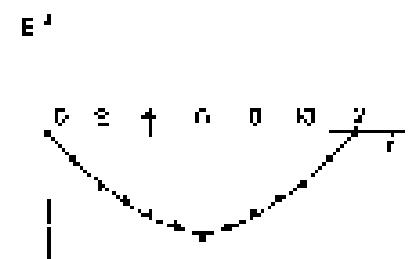


Fig. 15-14. The addition of Fig. 15-13 with a Coulomb curve. Molecular with $n = 2, 5, 10$ are more above than the others.

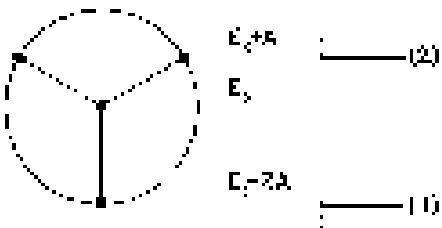


Fig. 15-15. Energy diagram for a ring of three.

example. We would, however, expect that the configurations with a fewer than normal energy would have an above average such hyperactivity-sparking. Notice that the configurations further below the curve always occur at the end of one of the levels. This again is, namely when there are enough electrons left up on "empty shells" so it's willing. This is the very basis of the whole of the theory. Molecules or ions are only really stable (in comparison with other similar configurations) when the available electrons just fit in one empty shell.

This theory has experiment and prediction some very peculiar chemical facts. To take a very simple example, consider a ring of three. It's almost unbelievable that the electrons can make a ring of three and four. You might think that it has been done. The energy diagram for three electrons is shown in Fig. 15-15. Now if you put two electrons in the lower state, you have only two of the three electrons that you require. The third electron must be put in at a much higher level. By our argument this molecule should not be particularly stable, whereas the four-electron state is a scutti-be stable. It does turn out, in fact, that the neutral molecule of triphenyl cyclopropenyl is very hard to make but at the first ionization in Fig. 15-16 is relatively easy to make. Making of them is more really easy because there is always a large stress when the bonds in an organic molecule make an equilateral triangle. To make a stable one would not do, the structure must be stabilized in some way. Anyway if you add three benzene rings on the corners, the positive ion can be made. (The reason for this requirement of odd-benzene rings is not really understood.)

In a similar way the five-membered ring can also be analyzed. If you draw the energy diagram, you can see in a qualitative way that the five-electron system should be the especially stable structure so that each molecule should be most stable as a negative ion. Now the five-ring is well known and does not make and always makes a negative ion. Surprisingly, you can only get a ring of four & is not very interesting, but that a ring of five in IC base along with should be especially stable as a neutral object.

15-6 Other uses of the approximation

There are now other similar situations which we will describe only briefly. In considering the structure of an atom, we can consider that the electrons fill successive shells. The Schrödinger theory of electron motion can be worked out easily only for an electron moving in a "circular" field—one which varies only with the distance from a point. We can then understand what goes on in an atom which has 12 electrons! One way is to use a kind of independent-particle approximation. First you calculate what happens with one electron. You get a number of energy levels. You put an oxygen into its usual energy shell. You can, for a rough model, continue to ignore the electron interactions and go on filling successive shells. In there is a way to get both electrons by taking the account—an approximate way of saying the effect of one extra charge carried by one electron. Each time you add an electron you compute its amplitude to be at various places and then see this amplitude is actually a kind of spherically symmetric charge distribution. You see the field of the distribution—regular with respect to the positive nucleus and all the previous electrons—to calculate the states available to the next electron. In this way you can get reasonably accurate estimates for the energies for the central atom and the outer occupied shells. You find that there are energy wells, just as we saw for one electron in a ring molecule. With a partially filled shell, the atom will seek a preference for having one or more electrons in the lowest state that has to go into the next shell outside of a full shell.

The theory explains the oddities behind the fundamental chemical properties which show up in the periodic table of the elements. The last group of these elements in which a single valence electron is present, and it is especially difficult to make them react. (Some of them do react of course—such as carbon and oxygen, for example, but such compounds are very weakly bound, the associated interaction is nearly zero.) An atom which has one valence electron has less

the α has got off easily losing just one electron to set into the excited state); low-energy excitation which comes from here to a completely filled shell—they are the very noble chemical elements of valence = 0 or $-$.

The other situation is found in nuclear physics. In nuclei, nucleons—protons and neutrons interact with each other quite strongly. From the independent particle model can again be well-defined a shell structure. Twisted discoveries eventually in nuclei were reported shortly after they calculated certain magic numbers by Gurney (1932, 1933, 1934, 1935, 1936). Since there was literally no explanation for these numbers they were called "magic numbers" at nuclear physics. It is well known that nucleons don't interact strongly with each other—people were therefore quite surprised when it was discovered that an independent particle model predicted a shell structure which came out with the first few magic numbers. The model assumed that each nucleon (proton or neutron) remained in a central potential which was created by the exchange effects of all the other nucleons. This model failed, however, to give the correct values for the magic numbers. Then it was discovered by Maria Goeppert Mayer, and independently by Jensen and his collaborator, that by taking the independent particle model and adding only a correction for what is called the "spin-orbit interaction," one could calculate an improved model which gave all of the magic numbers. (The spin-orbit interaction states the energy of a nucleon to be lower if its spin has the same direction as its orbit's angular momentum from motion in the nucleus.) The "spin-orbit correction" is also called "the so-called 'shell structure'" of the nuclei, though it is not related with the necessities of nuclei and their interactions.

"An independent particle approximation has been...a valuable range of subjects: from solid-state physics, to chemistry, to biology, to nuclear physics. It is often only a crude approximation, but it is very useful in understanding why there are especially stable configurations of shells. Since it is all of the complexity of the interactions between the individual particles, we thank you for accepting that a student's opportunity to grow correctly many important details."

The Dependence of Amplitudes on Position

16-1 Amplitudes on a Line

We are now going to discuss how the probability amplitude of a atom is related to ψ_{1s} in space. In some of the earlier chapters you may have had a prior impression that all of the wave function was being cut off. This is only what we are talking about here, in principle, you do not do this in the theory of quantum theory. But one has to say which the situation in which it is necessary, since we "lived" in the plane of the three oxygen atoms, and for the other two ways we took on the condition is set that one oxygen atom was "below" the plane of the three oxygen atoms. Very likely we pick just these two sites.¹ Why is it not possible that the oxygen atom could be at a hydrogen above the plane? C_6H_6 ? In there are many positions that the hydrogen atom is allowed. Again we see we take along the hydrogen molecular bond, in which there is one site surrounded by two hydrogens, we know well, we have shown, and so the electron in the hydrogen bond of C_6H_6 jumps from and the other for the electron in the oxygen, enclosed of proton number two. Clearly we were being very many details. The electron is not necessarily at proton number two, but is only to the neighborhood it could be somewhere there or perhaps even when below the proton, or maybe even to the left of the proton or somewhere to the right of the proton.

We fundamentally assumed wavefunction Little details. We said that we were interested in only certain features of the position so we were imagining that with the electron we in the vicinity of proton number two, it would take up a certain energy definite state. In that condition the probability to find the electron would have come to be definitely restricted around the proton, and we were not interested in the details.

We can also put another way. An electron in a hydrogen molecule has got to move in approximately a region where we described the situation in terms of two base states. Interacting there are four kinds of the states. And then one has to be up essentially around 2.3732 ± 0.00007 , or greater than one, but there are many excited states. For each excited state the distribution of the electron around the proton is different. We ignore all these excited states, saying that we were interested in only the components of free energy. But it is not these other excited states which give the possibility of various distributions of the electron around the proton. If we want to do this, to recall the oxygen negative ion, we have to take into account these other possible base states. We could do this in several ways, one one way is to consider the orbital paths in which the motion of the electron in space is more or less described.

We are now ready to consider a more detailed discussion where we allow us to talk in detail about the position of the electron by giving a probability amplitude of five. In fact in most books you will see everywhere the given as six. This more complete theory provides the model fitting for the calculations we have been making in our earlier discussions. The same number of equations can be derived as a kind of generalization to the more complete theory.

You may be wondering why we did not learn such the more complete theory and make the approximations as we want doing. We have said that it would be much easier for you to gain an understanding of the basic mechanics of quantum mechanics by beginning with the two-state approximation and working gradually up to the more complex theory than to approach the subject in the usual way. It is good. For this reason the approach to the subject requires to be in the reverse order to the one you will find in many books.

16-1 Amplitudes on a line

16-2 The ψ_{1s} function

16-3 States of definite momentum

16-4 Normalization of the states in L^2

16-5 The Schrödinger equation

16-6 Quantized energy levels

As we go into the subject of this chapter you will realize that we are breaking a rule we have always followed in the past. Whenever we have taken up any subject, we have always tried to give a more or less complete description of the theory—showing you as much as we could about it. Since the idea has been “We have tried to describe the general consequences of a theory as well as describing some specific detail—but you can see where the theory would lead.” We are now going to break that rule because we are trying to describe how one can fit the probability amplitudes in space and show you the different equations which they satisfy. We will not have time to go into all the details of the various implications which come out of the theory. Instead we will just have to go far enough to relate this theory to some of the approximations which we have used earlier. For example, to the hydrogen molecule or to the diatomic molecule. Perhaps you might have a hunch—uninformed and unprepared. We are approaching the end of our course, and we must surely confess with trying to give you an induction to the general ideas and start indicating the connections between what we have been describing and some of the other ways of approximating the subject of quantum mechanics. We have to give you enough opportunity that you can go off by yourself and by reading books here about some of the implications of the equations that we are going to describe. We must after all leave something to the future.

Let's review once more what we have found out about how an electron can move along a line of atoms. When an electron has an amplitude to jump from one atom to the next, there are definite energy states in which the probability amplitude to find the electron at a distance along the lattice is the form of a travelling wave. For long wavelength, i.e., small values of the wave number k , the energy of the state is proportional to the square of the wave number. For a typical state with the spacing Δx at which the amplitude per unit size for the electron to jump from one atom to the next is a_0 , the energy of the state is related to k (per atom) by

$$E = \frac{e^2}{4\pi\epsilon_0} k^2 \quad (16.1)$$

(see Section 13.3). We also saw that groups of such waves with similar energies would make up a wave packet which would behave like a classical particle with a mass m_e given by

$$m_e = \frac{\hbar^2}{2\Delta x} \quad (16.2)$$

Since waves of probability amplitude a cannot behave like a particle, one might well expect that the general quantum mechanical description of a particle would show the same kind of wave behavior we observed for the lattice. Suppose we were to think of a lattice on a line and imagine that the lattice spacing Δx were to be very small and smaller. In that limit we would be thinking of a state in which the electron could be anywhere along the line. We would have just one continuous distribution of probability amplitudes. We would have no amplitude to find an electron anywhere along the line. This would be one way to describe the motion of an electron in a vacuum. In a framework like this, space can be labeled by an infinity of points all very close together and we can work out the equations that relate the amplitudes at one point to the amplitudes at neighboring points. we will have the continuum mechanics laws of motion of an electron in space.

Let's begin by reviewing some of the general principles of quantum mechanics. Suppose we have a particle which can exist in various conditions in a system in a mechanical space. Any particular condition or state can be found in, we call it, “space,” which we label with a state vector, say $| \psi \rangle$. Some other condition would be labeled with another state vector, say $| \phi \rangle$. We then introduce the idea of base states. We can then choose a set of states $| 1 \rangle, | 2 \rangle, | 3 \rangle, \dots$, and so on, which have the following properties. First, all of these states are quite distinct—we say they are orthogonal. By this we mean that for any two of the base states, $| \beta \rangle$ and $| \gamma \rangle$, the amplitude $\langle \beta | \gamma \rangle$ that a electron, known to be in the state $| \beta \rangle$ is also in the

state $|j\rangle$ is equal to zero—unless, of course, $j = 0$ and ψ is odd for the wave function. We represent this symbolically by

$$\langle j | \hat{j} | \psi \rangle = \delta_{j0}. \quad (16.1)$$

You will remember that $\delta_{j0} = 0$ if j and j_0 differ, and $\delta_{j0} = 1$ if j and j_0 are the same numbers.

Second, the base state $|\psi\rangle$ must be a complex wave so that any state or field can be described in terms of them. That is, any wave, $\psi(x)$, can be described completely by giving all of the amplitudes of $|\psi\rangle$ that a particle in the state $|\psi\rangle$ will also be found in the state $|j\rangle$. In fact, the wave function $\psi(x)$ is equal to the sum of the wave functions each multiplied by a coefficient which is the amplitude of the state $|j\rangle$ in the wave $|\psi\rangle$.

$$|\psi\rangle = \sum_j c_j |j\rangle, \quad (16.2)$$

Finally, if we consider any wave state $|\phi\rangle$ and $|\psi\rangle$ the amplitudes that we gave $|\phi\rangle$ will also be in the state $|\psi\rangle$ can be found by first preparing the state $|\phi\rangle$ in the base states and then projecting back each one into the state $|\psi\rangle$. We write that in the following way:

$$\langle \phi | \psi \rangle = \sum_j \langle \phi | |j\rangle \langle j | \psi \rangle. \quad (16.3)$$

For summation j , of course, to determine c_j over the whole set of base states $|j\rangle$.

In Chapter 11 when we were working out what goes on in an unoccupied or a linear array of atoms, we chose a set of base states $|j\rangle$ which the electron was localized at one or either of the atoms in the line. The base state $|j\rangle$ represented the atom in which the electron was localized at atom number “ j .” (There is, of course, no significance in the fact that we called our base states $|j\rangle$ instead of $|0\rangle$.) At this time, we found it convenient to label the base states by the coordinate x_j of the atom rather than by the number of atom j in the array. The state $|x_j\rangle$ is either way of writing the state $|j\rangle$. Then, following the general rules, any state in QM, $|\psi\rangle$, is described by giving the amplitudes and that an electron in the state $|\psi\rangle$ is also in one of the states $|x_j\rangle$. Thus, whenever we have to refer to symbols, we want to choose amplitudes,

$$c_j = \langle \psi | x_j \rangle. \quad (16.4)$$

Since each base state associated with a location along the line, we expect each of the amplitudes c_j as a function of the coordinate x and write it as $C(x_j)$. The amplitude $C(x_j)$ will, in general, vary with x and j ; there is also the factor of i . We will generally bother to show explicitly the dependence.

In Chapter 11 we had proposed that the amplitudes $C(x_j)$ should vary with x in a way described by the Hamiltonian equation (Eq. 16.2). In the new notation this equation is

$$i\hbar \frac{dC(x_j)}{dx} = E_j C(x_j) = -4C(x_j + \Delta) + 4C(x_j - \Delta), \quad (16.5)$$

The last two terms on the right-hand side represent the process in which an electron travels from atom j to atom $j + \Delta$ and back again.

We found that Eq. 16.5 has solutions corresponding to relative energy states, which we wrote as

$$C(x_j) = e^{iE_j x_j / \hbar}, \quad (16.6)$$

For the low-energy states the wavelengths are large (k is small), and the energy is related to k by

$$E_j = (E_F - 144) + 2k^2 \hbar^2. \quad (16.7)$$

For electrons near zero of energy we find $(E_F - 144) = 0$ because E_F is given by Eqs. 16.1.

Let's see what might happen if we were to let the lattice spacing Δx go to zero, keeping the wave number k fixed. If ψ_0 is still between x_0 and x_1 , the last term in Eq. (16.9) would just go to zero and there would be no physics. But suppose k and ϵ are varied together so that k goes to zero the product $k\epsilon$ is kept constant. For $k \ll k_0 = \pi/\Delta x$ we will write $k\epsilon^2$ as the constant $M/\Delta x^2$. Then there is no surprise, Eq. (16.8) would not change, but what would happen to the differential equation, Eq. (7.1)?

For ϵ we will use $\epsilon = \epsilon_0 + \delta \epsilon$.

$$i\hbar \frac{\partial \Psi(x)}{\partial t} = (\epsilon_0 - M\chi(x_0) - M\chi(x_1)) + C\epsilon_{x_0} + N + C\epsilon_{x_1} = \psi_0. \quad (16.10)$$

For small values of $\delta \epsilon$, the first term is dominant. Now we can think of a continuous derivative of ϵ that goes smoothly through the point ϵ_0 at x_0 and x_1 . As the amplitude ψ goes to zero, the solution gets close and close together, and (if we keep the variation of ϵ fairly small) the quantity in parentheses is just proportional to the second derivative of $C\chi$. We can write (as you can see by making a Taylor expansion of $\psi(x)$ about the center)

$$\chi(x) = C(x + \delta) = C(x - \delta) \approx -i\hbar \frac{\partial^2 \chi(x)}{\partial x^2}. \quad (16.11)$$

In this limit, then, as δ goes to zero, keeping $\delta^2 \epsilon$ equal to M , Eq. (16.10) goes over into

$$i\hbar \frac{\partial \Psi(x)}{\partial t} = -\frac{C^2}{\Delta x^2} \frac{\partial^2 \Psi(x)}{\partial x^2}. \quad (16.12)$$

We have an equation which says that the time rate of change of Ψ is proportional to Δx and the second derivative of the amplitude. It has the effect of damped oscillations in a way which is proportional to the square derivative of the amplitude with respect to position.

The correct quantum mechanical equation for the motion of an electron in free space was first discovered by Schrödinger. For motion along a line it has exactly the form of Eq. (16.12); if we replace ψ by ψ , the free-space mass of the electron, the motion equation like ours space the Schrödinger equation:

$$i\hbar \frac{\partial \Psi(x)}{\partial t} = -\frac{e^2}{2m} \frac{\partial^2 \Psi(x)}{\partial x^2}. \quad (16.13)$$

We do not intend to have you think we have derived the Schrödinger equation, not only with us, but you and every one thinking about it. When Schrödinger first wrote it down, he gave a long narrative to his friend, the other great Nobel laureate Niels Bohr, in which he explained what he had done. In this narrative he said some rather interesting things. Some of the coefficients he used were even, like, all that does not matter, the only important thing is that the ultimate equation gives a correct description of nature. I am purposefully digressing a bit simply to show you that the correct fundamental quantum mechanical equation is not just the same form you get for the limiting case of an electron or electron-like particle. This makes the wave function of the differential equation to be (6.13) as describing non-antisymmetric probability amplitude from one point to the next along the line. That is, than each point has certain amplitude to go to our pixel. It will be little. One has to have some amplitude to be at neighboring points. In fact the equation looks something like the diffusion equation which we have used in Lecture 1. But there is one main difference: the imaginary coefficient, i, factor of the time derivative makes the solution exponentially different from the ordinary diffusion, such as you would have for a gas spreading out along a thin tube. Oscillatory behavior gives rise to real exponential solutions. Whereas the solutions of Eq. (16.13) are complex waves.

* You can imagine that as the pixels x get closer together, the amplitude ψ is jumping from x_0 up to x_1 will increase.

16-1 The wave function

Show that you have some idea about how things are going to work, as you go back to the beginning and study the problem of describing the motion of an electron along a line without losing its initial state connected with atoms with lattice. We want to go back to the beginning and see what rules we have to use if we want to describe the motion of a free particle in space. Since we are interested in the behavior of a free electron, we come down to the shell up which another chapter of physics is based. You will see the ideas we have developed for dealing with a finite number of states will need some technical modification.

We begin by letting the state vector ψ stand for a state in which a particle is located precisely at the coordinate x . The energy value is along the line $E = E(x)$, Eq. 16.07, or 16.08. This is the corresponding state. We will take these states $|\psi\rangle$ as in these states and, if we include all the points on the line, we will have a complete set of states in one dimension. Now suppose we have a different kind of a state, say $|\psi_x\rangle$, in which the electron is distributed in some way along the line. One way of describing this state is to recall from Chapter 11 that the electron will be also found in each of the base states $|\psi\rangle$. We may give an infinite series of amplitudes, one for each value of x . We will write these amplitudes as $C(x)$. Each of these amplitudes is a complex number and since the C is just such complex number for each value of x , the amplitude $C(x)$ is called just a function of x . We will also write it as $C(x)$,

$$C(x) = \langle x | \psi \rangle \quad (16-1)$$

We have already discussed such amplitudes while they are continuous over with the coordinates when we talked about the variation of amplitude with time in Chapter 9. We observe here, for example, that a particle with a definite momentum should be expected to have a particular relation of x and amplitude in space. If a particle has a definite momentum p and wave number k defined over R , the amplitude is to be found at any position x would be like

$$\langle x | \psi \rangle = C(x) \propto e^{ikx}, \quad (16.05)$$

This equation expresses the important general principle of quantum mechanics which connects the base states corresponding to different wave numbers or different spatial distributions of the state of infinite momentum. The infinite momentum state is often more important than base states in other certain kinds of problems. The set of base states is, of course, usually acceptable for a description of a quantum mechanical situation. We will come back later to the matter of the relation between them. For the moment we want to stick to our connection of a description in terms of the states $|\psi\rangle$.

Before proceeding, we want to make one sort of change in notation which we hope will not be too confusing. The function $C(x)$, defined in Eq. (16.05), is of course here a wave vector, except on the particular point x under consideration. We should indicate that in some way. We should, for example, specify which function $C(x)$; we are talking about by a subscript, say, $C_0(x)$. Although this would be perfectly satisfactory notation, it is a little too cumbersome and is not the one you will find in most books. Most people simply omit the letter C and use the symbol ψ to define the function

$$\psi(x) = C_0(x) = \langle x | \psi \rangle. \quad (16.06)$$

Since this is the notation used by everybody else in the world, you might as well get used to it so that you will not be confused when you come across it somewhere else. Remember though, that we will now be using ψ in two different ways. In Eq. 16.04 ψ refers to a total wave given to a particular physical state of the electron. On the left hand side of Eq. 16.06, on the other hand, the symbol ψ is used to define a mathematical function of x which is equal to the amplitude to be associated with each particular displacement. We hope it will not be too confusing.

Once you get accustomed to the idea, conveniently, the function $\psi(x)$ is usually called "the wave function" — because it encodes all the relevant information about the system.

Since we have defined $\psi(x)$ to be the amplitude that an electron in the state ψ will be found at the position x , we would like to compute the absolute square of $\psi(x)$ to get probability of finding an electron at the position x . Unfortunately, the probability of finding a particle exactly at any one point is zero. The answer will, in general, be averaged out in a certain region of the line, and since, in any small piece of the line, there are an infinite number of points, the probability that it will be at any one of them is then be a finite number. We can only describe the probability of finding an electron in terms of a probability amplitude, which gives the relative probability of finding the electron at certain regions over others. *Again*, along the line. Let's take Δx to be a small interval around x . If we go to a small enough Δx in an applied situation, the probability will be varying smoothly from place to place, and the probability of finding the electron in any small finite line segment Δx will be proportional to Δx . We can modify our definition to take this into account.

We can think of the amplitude $\langle x | \psi \rangle$ as representing a kind of "imprint density" for all the possible states ψ in a small region. Since the probability of finding an electron in a small region Δx is proportional to the "intensity" of the wavefunction definition of $\langle x | \psi \rangle$, so that the following relation holds:

$$|\psi(x)|^2 \Delta x = |\langle x | \psi \rangle|^2 \Delta x$$

The amplitude $\langle x | \psi \rangle$ is therefore proportional to the "intensity" that an electron in the state ψ will be found in the box x , and the constant of proportionality is determined by the absolute square of the amplitude $|\langle x | \psi \rangle|$, giving the probability density of finding an electron in a small region. We can write, equivalently,

$$\text{prob}(x, \Delta x) = |\psi(x)|^2 \Delta x. \quad (16.17)$$

We will now come to modify some of our earlier equations to make them compatible with this new definition of a probability amplitude. Suppose we have an electron in the state $|\psi\rangle$ and we want to know the amplitude to find up in a different state $|\phi\rangle$, which may correspond to a different bound state condition of the electron. When we were talking about bound states in discrete states, we would have used Eq. (16.8). Before modifying our definition of the amplitudes we would have written

$$\langle \phi | \psi \rangle = \sum_{\text{all } n} \langle \phi | n \rangle \langle n | \psi \rangle. \quad (16.18)$$

Now, each of these amplitudes are normalized in the same way as we have been writing them, that is, sum of all the states in a complete region Ω would have to be 1, due to multiplying by Δx , and the sum over n values of $\langle n | \psi \rangle$ is simply because an integral. With our modified definition, the correct form becomes

$$\langle \phi | \psi \rangle = \int_{\text{all } \Omega} \langle \phi | n \rangle \langle n | \psi \rangle / \Delta x \, dx. \quad (16.19)$$

The amplitude $\langle \phi | \psi \rangle$ is what we are now calling $\psi(\phi)$ and, in a similar way, we will choose to let the amplitude $\langle \psi | \phi \rangle$ be represented by $\phi(\psi)$. Furthermore, since $\langle \phi | \psi \rangle$ is the complex conjugate of $\langle \psi | \phi \rangle$, we can write Eq. (16.8) as

$$\langle \phi | \psi \rangle = \int_{\text{all } \Omega} \psi^*(x) \phi(x) \, dx. \quad (16.20)$$

Here are new definitions everything follows with the same formulae as before if you always replace a summation sign by an integral over Ω .

We should mention one qualification to what we have been saying. Any suitable set of base states must be complete if it is to be used for an arbitrary

⁴ For a discussion of probability distribution see Vol. I, Section 6.4.

Consider now what is going on. For an electron's wavefunction it is not really sufficient to specify only the horizontal position because of course there are electrons moving up and down as well as left and right. One way of getting a complete set is to take two sets of states in a wavefunction space and have one for down spin. We will however not worry about such complications for the time being.

16-3 States of definite momentum

Say we have an electron in a state $|\psi\rangle$, which is described by the probability amplitude to find $\psi = \psi(x)$. We know that the momentum in which the electron is spread is along the line in a certain direction, so that the probability of finding the electron in a small interval dx at the location x is just

$$|p(x)|^2 dx = |\psi(x)|^2 dx.$$

What can we say about the momentum of this electron? We might ask what is the probability that the electron has the momentum p ? Let's start out by calculating the amplitude that the state $|\psi\rangle$ is in some momentum $|p\rangle$ (which we define to mean one with the definite momentum p). We can find the amplitude by using our basic equation for the states of amplitudes, Eq. (16.20). In terms of its states, we may

$$\langle \text{mom } p | \psi \rangle = \int_{-\infty}^{\infty} \langle \text{mom } p | x \rangle \psi(x) dx, \quad (16.21)$$

and the probability that the electron with initial wavefunction $|\psi\rangle$ should be given in terms of the absolute square of the amplitude. We have noted, however, a small problem about the normalization. In general we can't care about the probability of finding an electron with a momentum p since if you do at the momentum p , the probability that the momentum is exactly some value p may be zero unless the state $|\psi\rangle$ happens to be a state of definite momentum. Once $\langle \psi | \psi \rangle$ is 1, the probability of finding the momentum in a small range dp at the momentum p will weight a definite probability. This is the reason why the normalization of $|\psi\rangle$ is required. We will choose one of these which we think to be the appropriate, although that may not be apparent to you at the moment.

We take our normalizations such that the probability is related to the amplitude ψ

$$p(x) (p, dx) = \langle \text{mom } p | \psi \rangle^2 \frac{dx}{m}, \quad (16.22)$$

With this definition the normalization of the amplitude (temporarily) is determined. The amplitude (now ψ) is, of course, just the complex scaling of the amplitude (or more ψ) which is just the one we have written down in Eq. (16.18), with no normalization we have chosen, or forced out that the proper constant of proportionality in front of the exponential is just 1. Now this

$$\langle \text{mom } p | \psi \rangle = \langle \psi | \text{mom } p \rangle^* = e^{-ipx/\hbar}, \quad (16.23)$$

Eq. (16.23) then becomes

$$\langle \text{mom } p | \psi \rangle = \int_{-\infty}^{\infty} e^{i(px/\hbar)} \langle p | \psi, dx. \quad (16.24)$$

This combined with Eq. (16.22) allows us to find the momentum distribution for every x in Eq. (2)

Let's look at a particular example. To illustrate this let's take an electron scattered in a vertical region around $x = 0$. Suppose we have a wave function which has the following form:

$$\psi(x) = R_0 e^{-k|x|}. \quad (16.25)$$

The probability distribution in x for this wave function is the case the square in

$$\text{prob}(x, dx) = P(x) dx = R_0^2 e^{-2k|x|} dx. \quad (16.26)$$

The probability density function $P(x)$ is the Gaussian curve shown in Fig. 16-1. Most of the probability is concentrated between $x = -\sigma$ and $x = +\sigma$. We say that the "full-width" of the curve is σ . (Here σ is equal to the root-mean-square of the error.) The x for summing carry out according to the distribution. We would normally choose the constant K so that the probability density $f(x)$ is zero merely \pm equivalent to the probability per unit length of x of finding the electron at x , but here we take with this $f(x)dx$ is equal to the probability $\sim \sigma^{-1}$ of finding the electron in the near x . The constant K which does this can be found by requiring that $\int_{-\infty}^{+\infty} f(x)dx = 1$. Since there must be unit probability that the electron is found somewhere. Here we get just $K = (2\pi\sigma^2)^{-1/2}$. This has been checked that $\int_{-\infty}^{+\infty} e^{-x^2/\sigma^2} dx = \sqrt{\pi}$, see Vol. I, page 40-6.1.

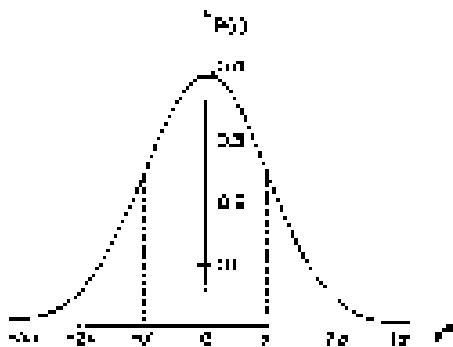


Fig. 16-1. The probability density for the wave function of Eq. (16-24).

Now let's find the distribution in momentum. To do this we $\langle p | \psi(p) \rangle$ and do the amplitude to find the connection with the momentum p .

$$\phi(p) = \langle p | m/m_p | \psi \rangle. \quad (16-27)$$

Substituting Eq. (16-25) into Eq. (16-26) we get

$$\phi(p) = \int_{-\infty}^{+\infty} e^{-x^2/\sigma^2} K e^{-ipx/\hbar} dx. \quad (16-28)$$

The integral can easily be evaluated as

$$K e^{-x^2/\sigma^2} \int_{-\infty}^{+\infty} e^{-ipx/\hbar} dx = K e^{-x^2/\sigma^2} \int_{-\infty}^{+\infty} e^{-i(p/\hbar)x^2} dx. \quad (16-29)$$

We can now make the substitution $u = x + (p/\hbar)x^2/2\sigma^2$, and the integral is

$$\int_{-\infty}^{+\infty} e^{-i(p/\hbar)x^2} dx = 2\pi\sqrt{\frac{\hbar}{p}}. \quad (16-30)$$

(The mathematicians would probably object to this way we go there, but the result is, nevertheless, correct.)

$$\phi(p) = (2\pi\sigma^2)^{-1/2} e^{-p^2/2\sigma^2}. \quad (16-31)$$

We note the interesting result that the amplitude function, in p has precisely the same mathematical form as the amplitude function in x , only the width of the Gaussian is different. We can write this as

$$\phi(p) = (2\pi\sigma^2)^{-1/2} e^{-p^2/\Delta^2}, \quad (16-32)$$

where the half-width Δ of the Gaussian distribution is related to the half-width σ of the distribution by

$$\Delta = \frac{\hbar}{2p}. \quad (16-33)$$

Our last step if we move the width of the distribution in x very small by making Δx small, ζ becomes large and the distance between x 's is very much spread out. On the contrary, if we have a narrow distribution in x , it must correspond to a spread out distribution in y . We can, if we like, consider ζ to be some measure of the uncertainty in the localization of the wavefunction and of the position of the electron in the state we are studying. If we call them Δ_x and Δ_y respectively, Eq. (16.33) becomes

$$\Delta_x \Delta_y = \frac{\hbar}{2}. \quad (16.34)$$

For example enough, it is possible to prove that for any error term Δ_x in the position in x , in y , the product $\Delta_x \Delta_y$ cannot be smaller than the one we have calculated. The Gaussian distribution gives the smallest possible value for the product of the root mean square errors. In general, we can say

$$\Delta_x \Delta_y > \frac{\hbar}{2}. \quad (16.35)$$

This is a quantitative statement of the Heisenberg uncertainty principle, which we have discussed qualitatively in my previous note. We have usually made the assumption above that the minimum value of the product $\Delta_x \Delta_y$ is of the same order as $\hbar/2$.

16-4 Normalization of the states in x

We return now to the discussion of the modifications to our basic principles which we received when we are dealing with a continuum of base states. When we have a finite number of discrete states in the continuum, for what must be satisfied by the set of base states is

$$C|\psi_i\rangle = \delta_{ij}. \quad (16.36)$$

If ψ_i is called i the base state, the amplitude to be at another base state is C . By choosing a suitable normalization, we have defined the amplitude C to be 1. These two conditions are described by Eq. (16.36). We want now to ask how this relation must be modified when we have the two states $|x\rangle$ of a particle on a line. If the particle is known to be in one of the base states $|\psi_i\rangle$, what is the amplitude that it will be in any other base state $|\psi_j\rangle$? If $|\psi_i\rangle$ and $|\psi_j\rangle$ are two different locations along the line, then the amplitude $\langle\psi_i|\psi_j\rangle$ is really 0, as that is consistent with Eq. (16.36). But if x and x' are equal, the amplitude $\langle\psi_i|\psi_j\rangle$ will not be 0, because of the new field normalization problem. To see how we have to proceed, bring up, we go back to Eq. (16.36), and apply this expression to the special case in which the state $|\psi\rangle$ is just the rest state $|\psi_0\rangle$. We write this form

$$C'|\psi\rangle = \int C'|\psi(x)\rangle dx. \quad (16.37)$$

Now the amplitude $C'|\psi\rangle$ is just what we have been calling the "function" $\psi(x)$. Still, the amplitude $C'|\psi\rangle$, since it refers to the same state $|\psi\rangle$, is the state function of the variable x , namely $\psi(x)$. We can, therefore, use its Eq. (5.32) to

$$\psi(x) = \int C'|\psi(x)\rangle dx. \quad (16.38)$$

This operation is to divide an amplitude, resulting from a arbitrary function $\psi(x)$, into components which completely determine the nature of the amplitude $C'|\psi\rangle$, which is, of course, just a function that depends on x and x' .

Our problem now is to find a function $C(x, x')$ which when multiplied into $\psi(x)$ and integrated over all x gives just the quantity $\psi(x')$. It turns out that there is no mathematical function which will do this "at least nothing like what we ordinarily mean by a "function."

Suppose we pick ψ to be the special number 0 and let ϕ be the single-valued $\psi(x)$ to be some function of x , let's say $A(x)$. Then Eq. (16.39) would read as follows:

$$s(t) = \int f(x)\phi(x) dx. \quad (16.40)$$

What kind of function $\phi(x)$ could possibly satisfy this equation? Since the integral must not diverge, it must be true $\phi(x)$ takes for values at x other than 0, $f(x)$ must clearly be 0 for all values of x except 0. But if $f(x)$ is 0 everywhere, the integral will be 0, too, and Eq. (16.40) will not do either. So we have an impossible situation: we wish a function to be 0 everywhere but at a point, and still to give a finite integral. Since we can't find a function that does this, the easiest way out is just to say that the function $\phi(x)$ is defined by Eq. (16.40). Namely, $\phi(x)$ is that function which satisfies (16.40) itself. The function which does this we shall therefore call $\delta(x)$ and call it δ -function. We shall also say that the $\delta(x)$ -function has the charge property $\delta(0)$, if it is integrated for $f(x)$ in the Eq. (16.40), the integral picks out the value $f(0)$ of $f(x)$. This is when x is equal to 0, and, since the integral must be independent of $f(x)$ for all values of x except 0, the function $\delta(x)$ must be 0 everywhere except at $x = 0$. Since obviously we write

$$\delta(0) = \delta(x) \quad (16.41)$$

where $\delta(x)$ is defined by

$$\delta(x) = \int \delta(x)\psi(x)dx. \quad (16.42)$$

Notice what happens if we use the *smooth* function $\psi(x)$ for the function ψ in Eq. (16.42). Then we have the result

$$1 = \int \delta(x)dx. \quad (16.43)$$

This is because $\delta(x)$ has the property that it is 0 everywhere except at $x = 0$ but has a finite integral equal to 1 only. We must realize that the $\delta(x)$ has such a definite value at one point that the total area under it is equal to one.

One way of imagining what this δ -function is like is to think of a sequence of rectangles—or any other stepped function $\psi(x)$ —which gets narrower and narrower and higher and higher, always keeping its width Δx , as depicted in Fig. 16.3. The integral of this function from $-x$ to $+x$ is always 1. If you multiply it by any function $f(x)$ and integrate the product, you get something which is approximately the value of the function at $x = 0$ (an approximation getting better and better as you use the narrower and narrower rectangles). You can if you wish imagine the δ -function in terms of this kind of limit. By process the only important thing, however, is that the δ -function is defined so that Eq. (16.41) is true for any $\psi(x)$; that is, $\delta(x)$ really *represents* the δ -function. Its properties are then as we have described.

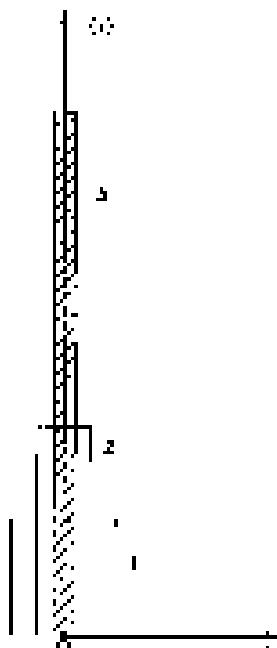
If we change the argument of the δ -function from x to $x - x'$, the corresponding relations are

$$\begin{aligned} \delta(x - x') &= 0, \quad x' \neq 0, \\ \int \delta(x - x')\psi(x)dx &= \delta(x'). \end{aligned} \quad (16.44)$$

If we use $\delta(x - x')$ for the additional $\psi(x)$ in Eq. (16.38), the equation is satisfied. Our reason this is that for non-zero values of x , the condition x corresponding to (16.38) is

$$(x' | x) = \alpha x - x'^2. \quad (16.45)$$

Fig. 16-3. A set of functions, of $\psi(x)$ and $\psi(x')$, which look more and more like $\delta(x)$.



We have now completed the necessary modifications of our basic equations which are necessary for dealing with the continuum of base states corresponding to two points along a line. The situation in three dimensions is fairly obvious; first we replace the coordinate x by the vector \mathbf{x} , then integrate over \mathbf{x} and we are replaced by integrals over x_1 , x_2 and x_3 . In other words, they become volume integrals. Finally, the one-dimensional δ -functions must be replaced by just the product of three δ -functions, one at x_1 , one in y_1 and the other in z_1 : $\delta(x - x_1)\delta(y - y_1)\delta(z - z_1)$. Putting everything together, we get the following set of equations for the amplitudes for particle motion in three dimensions:

$$\langle \psi | \psi \rangle = \int \langle \psi | \psi^*(x, y, z) \rangle d\mathbf{v} d\mathbf{l}, \quad (16.45)$$

$$\langle \psi | \psi \rangle = \langle \psi | \psi \rangle, \quad (16.46)$$

$$\langle \psi | \psi \rangle = \langle \psi | \psi \rangle,$$

$$\langle \psi | \psi \rangle = \int \psi^*(x, y, z) \psi d\mathbf{v} d\mathbf{l}, \quad (16.47)$$

$$\langle \psi | \psi \rangle = \langle \psi | \psi \rangle \delta(x_1 - x') \delta(y_1 - y') \delta(z_1 - z'), \quad (16.48)$$

What happens when there is more than one particle? We will tell you about how to handle two particles and you will easily see what you must do if you want to deal with a larger number. Suppose there are two particles, which we call particle No. 1 and particle No. 2. What shall we get for the base state? First, perfectly good or can be described by us is, that particle 1 is at x_1 and particle 2 is at x_2 , where we can write $x = (x_1, x_2)$. Now, then describing the motion of each separate does not right at base state. That's true since must define the condition of the relative system. You just not think that each particle moves independently as a wave in three dimensions. Any physical state ψ can be defined by a pair of all the coordinates (x_1, x_2, ψ) so that the two particles are located at x_1 and x_2 . This particular combination is therefore a function of the two sets of coordinates x_1 and x_2 . Generally such a function is known as the ψ as an oscillation that moves along in three dimensions. Neither it is really simple position of one individual wave, one for each particle. It is, in general, some kind of wave in the six dimensions defined by x_1 and x_2 . If there are two particles interacting with each other, there is no way of specifying what happens to one of the particles by looking at the wave function for it alone. The former goes does not go consistent in certain situations where the interactions made on one particle were claimed to be able to do what was going to happen to another particle, or were able to destroy an interference—very general people. All sorts of things like measure, they have tried to think of the wave function of one particle about, except that the overall wave function in the six-modes of two particles. The complete description can be given correctly only in terms of functions of the coordinates of both particles.

16.5 The Schrödinger equation

So far we have just been worrying about how we can deal the states which may evolve to elsewhere or anywhere or at the speed. Now we have to work about putting into a description the physics of what can happen in various circumstances. And here we have to worry about how states change with time. If we take a state ψ which goes over the another state ψ' which is also used to describe the situation for $t + \Delta t$ times by taking the wave function—which is just one amplitude $\langle \psi | \psi' \rangle$ a function of time as well as a function of the coordinates. Now take in a given situation ψ that is described by giving a time-varying wave function $\psi(\mathbf{r}, t) = \langle \mathbf{r}, t | \psi \rangle$. This time-varying wave function describes the evolution of successive states that occur as time develops. This is called “quantum superposition,” which gives the projections of the state ψ into the base states, which not always be the true superposition, but not we will consider it later.

In Chapter 8 we described how Eqs. (17.18) and (17.19) in terms of the diagonal A_{ii} . We saw that by some variation of the various amplitudes was given in terms of the matrix element

$$i\hbar \frac{\partial C_i}{\partial t} = \sum_j A_{ij} C_j. \quad (16.49)$$

This equation says that the time variation of each amplitude C_i is proportional to $i\hbar$ of the other amplitudes C_j , with the coefficients A_{ij} .

How would we expect Eq. (16.49) to look when we are using the continuum of states? Well, let's first remember that Eq. (16.49) can also be written as

$$i\hbar \frac{d}{dt} \psi(x) = \sum_j (x, \theta_j) \psi_j / \hbar \psi(x).$$

Now it is clear what we should do. With the wavefunction we would expect

$$i\hbar \frac{d}{dt} \psi(x) = \int (x, \theta) \psi(x) \psi'(x) dx. \quad (16.50)$$

The sum over the basis states ψ_j gets replaced by an integral over x' . Since (x, θ) is a distribution function of x , and we can write it as $\delta(x, x')$ which corresponds to θ_j in Eq. (16.49). Then Eq. (16.50) is the same as

$$i\hbar \frac{d}{dt} \psi(x) = \int \delta(x, x') \psi(x') dx' \quad (16.51)$$

with

$$\delta(x, x') = \langle x | \delta | x' \rangle.$$

According to Eq. (16.51) the rate of change of the $\psi(x)$ would depend on the value of ψ at x' times plus all the factors A_{ij} 's where the amplitude per unit time that the electron will jump from x to x' is δ and ψ is arbitrary. However, this rate amplitude is zero except for points at very close to x . This was shown in the example of the chain of atoms at the beginning of the chapter, Eq. (16.12), that the right-hand side of Eq. (16.13) can be expressed schematically in terms of δ and the derivatives $\delta/\delta x$ with respect to x , all evaluated at the position x .

For a particle moving freely in space with no forces, no disturbance, the solution of physics is

$$\int H(x, x') \psi(x') dx' = - \frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x),$$

where I have gotten from the Schrödinger. It's not possible to write anything more than this since our classical model of Schrödinger, invented as it was, struggle to find an understanding of the experimental observations of the real world. You can perhaps just somehow feel why it should be that way by thinking of our derivative of Eq. (16.51) which came about involving δ — the propagator of an electron in a crystal.

Of course, the particles are not very exciting. What happens if we put forces on the x ? Let's! Will the forces of the x be like forces described in terms of the scalar potential $V(x)$? What about forces due to linking of electric fields and magnetic fields? What about forces due to low energies so that we can ignore spin-orbit effects from relativistic theory, and the Hamiltonian which we did not yet discuss?

$$\int H(x, x') \psi(x') dx' = - \frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) + V(x) \psi(x) \quad (16.52)$$

Again, you can get some idea as to the origin of this equation if you go back to the motion of an electron in a crystal, and see how the equations would have to be modified if the energy of the electron varied slowly from one atomic site to the other, as in the case of time and an electric field outside the crystal. Then § 12

the ω in E_F in Eq. (16.7) would vary slowly with position and would be constant if the term we have added in (16.52).

It may be worth noting why we want a right from $\tilde{\psi}_j$ in (16.51) to Eq. (16.52) instead of just plus some linear combination of the amplitudes $\tilde{\psi}_{j+1/2} = (\psi_j + \psi_{j+1})/2$. We did this because $\tilde{\psi}_{j+1/2}$ can only be written in terms of a single algebraic function, although the whole is equal on the right-hand side of Eq. (16.41); outside outwards of L you are, however, "at" your real coordinates. $\tilde{\psi}_{j+1/2}$ can be written in the following way:

$$\tilde{\psi}(x, \omega) = -\frac{N}{2\pi i} \delta'(\omega - x) - \tilde{\psi}(x) \omega + \omega^2,$$

where δ' denotes the second derivative of the delta function. This is the wavefunction but be replaced by a semirelativistic wavefunction (algebraic differential operator, where ω is complex conjugate)

$$\tilde{\psi}(x, \omega) = \left\{ -\frac{e^2}{m} \frac{d}{dx} + V(x) \right\} \tilde{\psi}(x, \omega).$$

We will not be using these terms, but will work directly with the form in Eq. (16.52).

If we now take expression as above in (16.52) for the integral in (16.51) we get the following differential equation for $\tilde{\psi}(x) = \tilde{\psi}(x, \omega)$:

$$i\hbar \frac{d\tilde{\psi}}{dx} = -\frac{e^2}{m} \frac{d^2}{dx^2} \tilde{\psi} + V(x) \tilde{\psi}. \quad (16.53)$$

It is fairly obvious what we should do instead of Eq. (16.53) to see the wavefunction solution in three dimensions. The only changes are that d/dx gets replaced by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2},$$

and $V(x)$ gets replaced by $V(x, y, z)$. The amplitude $\tilde{\psi}(x, y, z)$ for an electron moving in a potential $V(x, y, z)$ obeys the differential equation

$$i\hbar \frac{d\tilde{\psi}}{dx} = -\frac{e^2}{m} \nabla^2 \tilde{\psi} + V(x) \tilde{\psi}. \quad (16.54)$$

This is called the Schrödinger equation, and was the first non-dimensional equation ever solved. It was written down by Schrödinger before any of the other quantum equations we have described in this book were discovered.

Although we have approached the subject along a completely different route, the year I started research marking the birth of the quantum mechanical description a matter occurred like Schrödinger's. First went down his equation to me. Not many years ago the internal atom of an α particle had been measured. Niels Bohr was able to understand what he had made by going why he was doing what he was doing, and especially how it could be that atoms could be stable. Although Bohr's theory did not give a description of the internal motion of an electron in a hydrogen atom which seemed to explain the observed spectrum of light emitted by this atom, the description did seem to result in this way to a good theory. Niels Bohr's theory of the atomic equations of motion of electrons in atoms was a theory from which atomic phenomena could be calculated one dimensionally, and in detail. In principle Schrödinger's equation is capable of explaining all atomic phenomena, except those involving magnetism and relativity. It explains the energy levels of an atom, and all the laws of chemical binding. This is, however, true only in principle. The mathematics, even however the computer, is more easily, very but not simple problems. Only the hydrogen and helium atoms have been solved to high accuracy. However, with some approximations, come fairly easily, many of the laws of atomic and nuclear atoms, and of the chemical binding, or molecules in general, otherwise. We know that you have a few approximations in some chapters.

The Schrödinger equation as we have written it does not take into account any magnetic effects. It is possible to take such effects into account in a more refined way by adding some terms to the equation. However, as we have seen in Volume I, a magnet is essentially a relativistic effect, and so a correct description of the motion of an electron in an arbitrary electromagnetic field can only be discussed in a more relativistic equation. The correct relativistic equation for the motion of an electron was discovered by Dirac a year after Schrödinger brought forth his equation, and takes on quite a different form. We will not be able to discuss it at all here.

Before we go on to look at some of the consequences of the Schrödinger equation, we would like to show you what it looks like for a system with a large number of particles. We will not be making any use of the variables, but just want to show it to you to emphasize that the wave function ψ is not simply an ordinary wave in space. It is a function of many variables. If there are many particles, the equation becomes

$$-\hbar \frac{\partial^2 \psi(r_1, r_2, \dots, r_n)}{\partial r^2} = \sum_i \frac{\hbar^2}{2m_i} \left[\frac{\partial^2 \psi}{\partial r_{i1}^2} + \frac{\partial^2 \psi}{\partial r_{i2}^2} + \frac{\partial^2 \psi}{\partial r_{i3}^2} \right] + V(r_1, r_2, \dots, r_n) \psi \quad (16.5)$$

The potential $V(r_1, r_2, \dots, r_n)$ corresponds obviously to the total potential energy of all the particles. If there are no external fields acting on the particles, the function V is simply the total kinetic energy of the motion of all the particles. That is, if the i th particle carries the charge q_i/q_0 , then the function V is simply

$$V(r_1, r_2, \dots, r_n) = \sum_{i=1}^n \frac{Z_i Z_i}{r_{ii}} \psi. \quad (16.6)$$

16.4 Quantized energy levels

In this chapter we will look in detail at a solution of Schrödinger's equation for a particular example. We would like now, however, to give you some idea of the most remarkable consequence of Schrödinger's equation—quantum mechanics. The surprising fact that a differential equation involving only continuous functions of continuous variables in space can give rise to quantum effects such as the discrete energy levels in an atom. The even more remarkable is how it can restrict an electron which is confined to a certain region of space by some kind of a potential "wall" must necessarily have only one or another of a series of well-defined discrete energies.

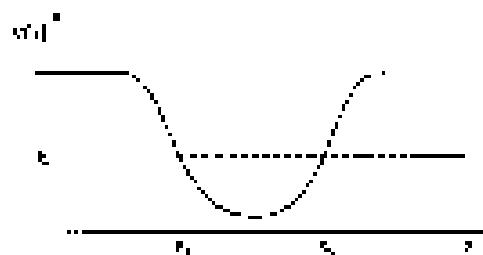


Fig. 16-3 A periodic wave for a particle moving along x .

Suppose we think of an electron in a one-dimensional situation in which its potential energy V is zero with t in a very situation by the graph in Fig. 16-3. We will assume that the potential is stationary—it doesn't vary with time. At a time t_1 later than time t_0 , we would like to find the solutions corresponding to states of definite energy, which means of definite frequency. Let's say a solution of the form

$$\psi = A \sin(\omega t + \phi), \quad (16.7)$$

[†] Requiring the condition of no other volume containing in which $\omega^2 = q_0/m$.
16-1

If we substitute this function into the Schrödinger equation, we find that the function $\psi(x)$ must satisfy the following differential equation:

$$\frac{d^2\psi(x)}{dx^2} + \frac{2m}{\hbar^2} [V(x) - E]\psi(x) = 0 \quad (16.28)$$

This equation says that at each x the second derivative of $\psi(x)$ with respect to x is proportional to itself, the coefficient of proportionality being given by the quantity $V(x) - E$. The second derivative of $\psi(x)$ is the rate of change of the slope. If the potential V is greater than the energy E , the potential, the rate of change of the slope of $\psi(x)$ will have the same sign as $\psi(x)$. This means that the value of $\psi(x)$ will be increasing away from the axis. That is, it will have, more or less, the character of the positive or negative exponential function. This is indicated in the regions to the left of x_1 and to the right of x_2 in Fig. 16.3, where V is greater than the bound energy E , the function $\psi(x)$ would have a break like the one in chapter 10.

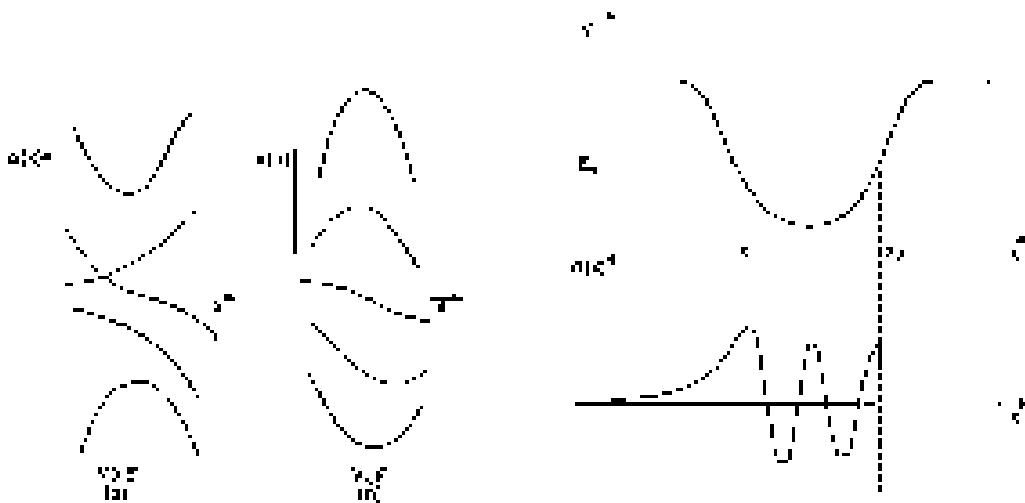


Fig. 16.4. Possible shapes of the wave function $\psi(x)$ for $V > E$ and for $V < E$.

Hg case. A wave function for the energy E which goes to zero for negative x .

shown in part (b) of Fig. 16.4.

On the other hand, if the potential $V(x)$ is less than the energy E , the second derivative of $\psi(x)$ with respect to x has the opposite sign from $\psi(x)$ itself, and the curve of $\psi(x)$ will always be concave toward the axis like one of the curves shown in part (a) of Fig. 16.4. The result is in such a region has, physically, roughly the "form" of a "mound" as in

Now let's see if we can construct graphically a solution for the function $\psi(x)$ which corresponds to a particle of energy E_1 in the potential V shown in Fig. 16.3. Since we are trying to describe a situation in which a particle is found inside the potential well, we want to look for solutions in which the wave amplitude either is very small values of x or is very outside the potential well. We can easily imagine a curve like the one shown in Fig. 16.5 which turns toward zero for large negative values of x and grows smoothly as it approaches x_1 . Since E is equal to V at x_1 , the x -coordinate of the function becomes zero at this point. Between x_1 and x_2 , the quantity $V - E_1$ is always a negative number, so the function $\psi(x)$ is always concave toward the axis and the curvature is a negative constant different than between x_1 and x_2 . If we continue the curve into the region beyond x_2 and x_3 , it should go more or less as shown in Fig. 16.5.

Now let's continue this curve into the region to the right of x_3 . There it comes away from the axis and takes off toward large positive values, as shown in Fig. 16.6. For the energy E_1 we have chosen the condition $|V(x)|$ is large and larger with x increasing. Thus, its curvature is also increasing if the potential continues to stay like V . The amplitude rapidly grows in numerical proportions.

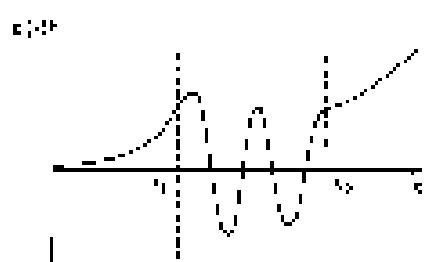


Fig. 16.5. The wave function part of Fig. 16.3 continued beyond x_3 .

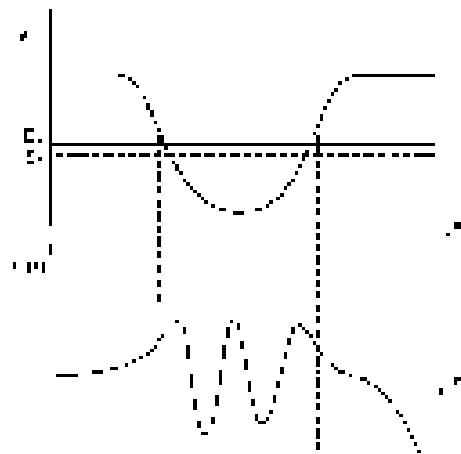
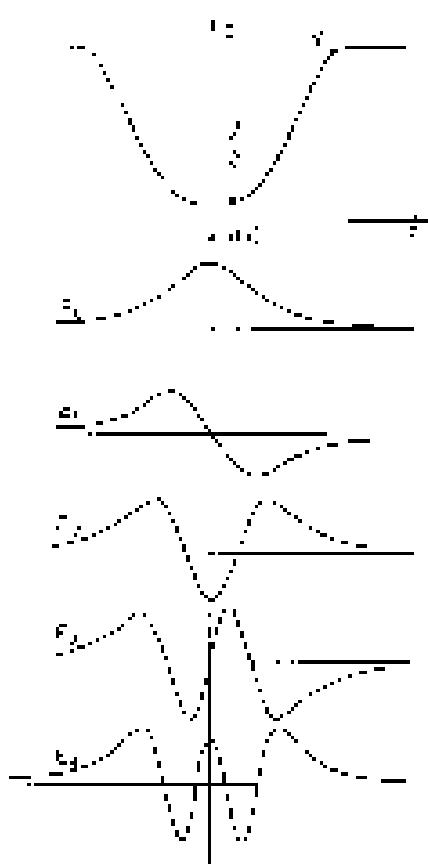


Fig. 16.2. The wave function $\psi(x)$ for an energy E , given by Eq. (16.1).



Fig. 16.3. A new solution for the energy E between E_1 and E_2 .



What does this mean? It implies that the particle is not "trapped" in the potential well. It is **much more likely** to be found *outside* of the well, than inside. For the whence we have manufactured, the electron is most likely to be found at $x = 1$ or, then $x = 2$ etc. We have failed to find a solution for a bound particle!

Let's try another energy, say one a little bit *higher* than E_1 , say the energy E_2 , i.e. Fig. 16.3. If we start with the same conditions as the last, we get the solution shown in Fig. 16.3. It's not *as though* we were going to be better, but it ends up just as bad as the solution for E_1 , except that now $\psi(x)$ is oscillating more and needs *more nodes* to get lower energy values!

Maybe the ψ is the clue. Since changing the energy E from between E_1 to E_2 causes the curve to flip from one side of the axis to the other, perhaps there is some energy lying between E_1 and E_2 for which ψ is a *real* (appreciable) negative value of x ? There is indeed, and we have sketched $\psi(x)$ for the solution in Fig. 16.4.

You should appreciate that the solution we have drawn in the figure is *not* a special case; if we were to vary up or down the energy even so slightly, the function would go over into curves like one of the others of the two broken-line curves shown in Fig. 16.3, and we would not have the proper conditions for a bound particle. We have obtained *one* solution if a particle is to be bound in a potential well, it can do so only if it has a very definite energy.

Does that mean that there is only one energy for a particle bound in a potential well? No. Other energies E are possible, you see energies too close to the barrier that the wave function we have drawn in Fig. 16.4 is no longer bounded in the region between x_1 and x_2 . If we were to pick an energy E which is lower than E_1 , we could choose a solution which crosses the axis only *once*, only twice, only once, or not at all. The possible solutions are sketched in Fig. 16.5. There are also n solutions (n infinite) corresponding to values of E which are higher than the ones shown. Our conclusion is that if a particle is bound in a potential well, its energy can take on only the certain special values in a discrete energy spectrum. You see how a differential equation can describe the basic fact of quantum theory.

We might remark one other thing. If the energy E is above the top of the potential well, then there are no longer any possible solutions, and any possible energy is permitted. Such solutions correspond to the scattering of free particles from potential wells. We have seen in example 6 such solutions when we considered the effects of impurity atoms in a crystal.

Fig. 16.5. The function $\psi(x)$ for the two lowest energy levels etc.

Symmetry and Conservation Laws

17-1 Symmetry

In classical physics there are a number of quantities which are conserved: position, energy, and angular momentum. Corresponding theorems about corresponding quantities also exist in quantum mechanics. The most beautiful thing of quantum mechanics is that the same rules—the laws can, in fact, be derived from very simple ideas; whereas in classical mechanics they are given by the seeming繁雜 of the laws of motion, in quantum mechanics we do an analogous thing to what we will do in quantum mechanics, but it can be done at a very simplified level. In quantum mechanics, however, the conservation laws are very deeply related to the principle of superposition of amplitudes, and to the symmetry of physical systems under various changes. This is the subject of the present chapter. Although we will apply these ideas mainly to the conservation of angular momentum, the central point is that the theorems about the conservation of all kinds of quantities are—in the quantum mechanics—related to the symmetries of the system.

We begin, therefore, by enquiring the question of symmetries of systems. A very simple example would be hydrogen molecule—we could equally well take the helium molecule. In which there are two atoms. For the hydrogen molecule we can look to see how states are in which the electron will be close near proton number 1, and another in which it is far away, but still near proton number 1. Its two states, which we called $|1\rangle$ and $|2\rangle$, are shown again in Fig. 17-1(a). Now, as long as the two nuclei are both outside the atom, that there is no interaction between the two protons. That is to say, if we were to reflect the system in the plane halfway between the two protons—by which we mean that everything on one side of the plane gets moved to the right, its position on the other side we would get the situation in Fig. 17-1(b). Since the protons are identical, the spectrum of reflection things $|1\rangle$ becomes $|2\rangle$ and $|2\rangle$ becomes $|1\rangle$. We'll call this reflection operator F and write

$$F|1\rangle = |2\rangle \quad F|2\rangle = |1\rangle. \quad (17.1)$$

so our F is an operator in the sense that it "does something" to a state to make a new state. One interesting thing is that if you map out the result, produces some sort of state of the system.

Now, if we try any of the wave functions we have described, has no evolution which can be defined by the usual successive action. Namely,

$$\psi_1 = \langle 1 | \hat{p} | 1 \rangle \quad \text{and} \quad \psi_2 = \langle 2 | \hat{p} | 2 \rangle$$

so the ψ_1 and ψ_2 elements we get if we multiply $\hat{p}|1\rangle$ and $\hat{p}|2\rangle$ in the left by $\langle 1 |$ and $\langle 2 |$ respectively. From Eq. (17.1) they are

$$\begin{aligned} \langle 1 | \hat{p} | 1 \rangle &= p_{11} = \langle 2 | \hat{p} | 1 \rangle = 0, \\ \langle 2 | \hat{p} | 2 \rangle &= p_{22} = \langle 1 | \hat{p} | 2 \rangle = 0. \end{aligned} \quad (17.2)$$

In the same way we can get p_{12} and p_{21} . The matrix of \hat{p} —and neglect of the wave function ψ_1 and ψ_2 —is

$$\hat{p} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad (17.3)$$

We see once again that the waves broaden the matrix. In classical mechanics we

17-1 Symmetries

17-2 Symmetry and conservation

17-3 The conservation laws

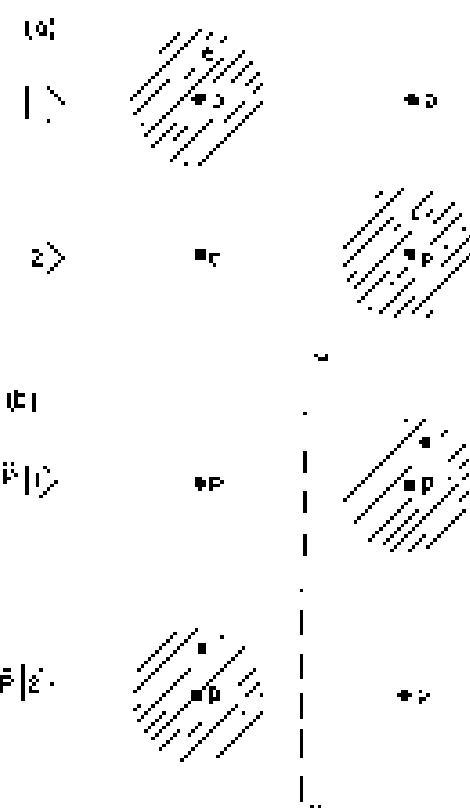
17-4 Polarized light

17-5 The dissociation of the A^+

17-6 Summary of the rotation matrices

Notes: Chapter 13, Sec. 1, Symmetry in Physical Laws

References: Angular Momentum in Quantum Mechanics
A. R. Edmonds, Princeton University Press, 1956



Hyp. 17-1. If the states $|1\rangle$ and $|2\rangle$ are reflected in the plane P-S, they come $|2\rangle$ and $|1\rangle$, respectively.

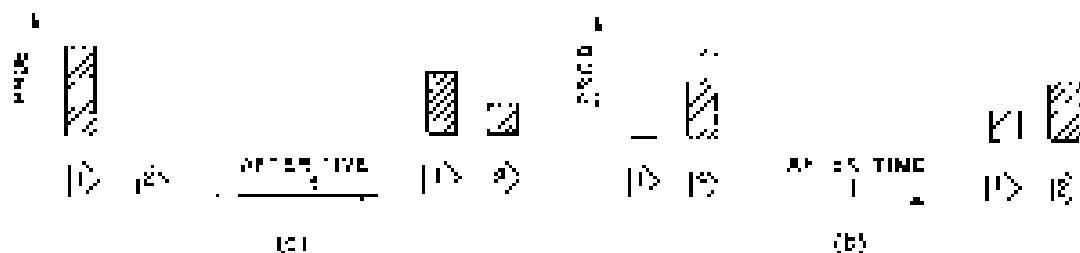


Fig. 17-2. In a symmetric system, if a pure $|1\rangle$ state develops in time it will tell a pure $|2\rangle$ state and develop as in part (c).

specifically interchanges. There are other technical differences— $\langle\psi|\psi\rangle$ is different between a "symmetric" and a "antisymmetric" state—but the distinction is something probably don't need to worry about. So whether Ψ defines an n -electron state or actually used to define a matrix of numbers, we can call it *interchangeable* or *invariant* in a sense.

Now we would like to gain a little insight. We will suppose that the system of the whole hydrogen molecular ion system is *symmetric*. Is there still time?

It depends on a choice, on which one to use. But if the system is symmetric the following just should obviously be true. Suppose we start with the system in the state $|1\rangle$ and find after an interval of time t that the system is now to be in a more correlated situation, in some linear combination of the two basic states. Remember that in Chapter 9 we used an instant "living for a period of time" by multiplying by the operator $e^{-iEt/\hbar}$. That means the "time" after a while, say 10 seconds, to be definite—say in some crazy way. For example, it might be $\sqrt{2}/3$ parts of the state $|1\rangle$ and $\sqrt{1/3}$ parts of the state $|2\rangle$, and we would write

$$|\Psi \text{ at } 10 \text{ sec}\rangle = C|1\rangle + D|2\rangle = \sqrt{2/3}|1\rangle + \sqrt{1/3}|2\rangle \quad (17.0)$$

Now we ask what happens if we start the system in the symmetric state $|1\rangle$ and wait for a second time, the same conditions? It is clear that if the whole is symmetric as we are supposing, we should get the state unchanged to (17.0).

$$|\Psi \text{ at } 15 \text{ sec}\rangle = T^2|1\rangle = C = \sqrt{2/3}|1\rangle + \sqrt{1/3}|2\rangle \quad (17.1)$$

The same ideas are sketched diagrammatically in Fig. 17-2. So if the physics does agree, it's consistent with some theory, and we work out the behavior of a particular atom, we also know the behavior of the rest, we would just by referring the original state to the symmetry plane.

We should like to do the same thing perhaps a bit more generally. We'd need a little more algebra. Let \mathcal{Q} be any sort of a number of operations that you do, \mathcal{Q} because to a system without changing the physics. For instance, let \mathcal{Q} be might be $C(\theta, \phi, \psi)$, the operation of a rotation in the plane between the two atoms in the hydrogen molecule. Or, in a system of n two-electron atoms, we might be thinking of the operation of swapping the two electrons. Another possibility would be in a spherically symmetric system, the operation of a rotation of the whole system through a finite angle around some axis, which wouldn't change the physics. Of course, we would normally want to give each physical operation a special name. For \mathcal{Q} , specifically, we will necessarily define $\mathcal{Q}(\theta, \phi, \psi)$ to be the operation "rotate the system about the z -axis by the angle θ ". By \mathcal{Q} we mean just any sort of the operations we have described in any other way, which leaves the basic physical situation unchanged.

Let's think of some more examples. If we have an atom with no *external* magnetic field or no external electric field, and if we want to turn the system into another antisymmetric, we'll be in some physical system. Again, the antisymmetric molecule is symmetric with respect to a reflection in a plane parallel to that of the two hydrogen. So now is there a no external field. Well, there is an electric field, when we make a reflection we would have to flip the electric field also.

and that changes the physical problem. But if we have no external field, the nucleus is symmetric.

Now let's consider a generic situation. Suppose we start with the state $|\psi_1\rangle$ and after some time or over many given physical conditions it has become the state $|\psi_2\rangle$. We can write

$$|\psi_2\rangle = \hat{U}|\psi_1\rangle \quad (17.6)$$

You can be thinking of Eq. (17.6) as "Now imagine we perform the operation \hat{U} on the whole system." The state $|\psi_1\rangle$ will have transformed into state $|\psi_2\rangle$, which we can also write as $\hat{U}|\psi_1\rangle$. Also the state $|\psi_2\rangle$ is changed into $|\psi_3\rangle = \hat{U}|\psi_2\rangle$. More generally, if you act just on the $\hat{\psi}$ (but not the \hat{U} , it's just a general property of systems), then, working for the same time under the same conditions, we should have

$$|\psi_3\rangle = \hat{Q}|\psi_2\rangle \quad (17.7)$$

[see Eq. (17.6)] Then we can write $\hat{Q} = \hat{U}$; for $|\psi_1\rangle$ and $\hat{Q}|\psi_1\rangle$ for $|\psi_3\rangle$ you can also write:

$$\hat{Q}|\psi_1\rangle = \hat{U}\hat{Q}|\psi_1\rangle. \quad (17.8)$$

Now we can replace \hat{Q} by $\hat{U}^\dagger\hat{Q}$ [Eq. (17.8)] to get

$$\hat{Q}\hat{U}|\psi_1\rangle = \hat{U}^\dagger\hat{Q}|\psi_1\rangle. \quad (17.9)$$

We must think to understand what this means. Thinking of the hydrogen atom it says there: "nothing is reflected and nothing is split." The expression on the right of Eq. (17.9)—with some as "nothing is reflected and nothing is split"—are equivalent on the left of Eq. (17.9). These should be the same so long as \hat{U} doesn't change under the reflection.

Since (17.9) is true for any starting state $|\psi_1\rangle$, it is really an equation about the operators:

$$\hat{Q}\hat{U} = \hat{U}\hat{Q}. \quad (17.10)$$

This is why we wanted to go to a more abstract measure of symmetry. When Eq. (17.10) is true, we say that the operators \hat{Q} and \hat{U} commute. We can then draw "top marks" in the following way. A classical system is symmetric with respect to the operator \hat{Q} when \hat{Q} commutes with it; the operation is by groups of time. [In terms of matrices, the product of two operators is equivalent to the trace product, so Eq. (17.10) also holds for the matrices \hat{Q} and \hat{U} for a system which is symmetric under the transformation \hat{Q} .]

Locally, since for infinitesimal times t we have $\hat{U} = 1 + i\hat{Q}/\hbar$ —where \hbar is the usual He (then in Sec. 5) you can see that if (17.10) is true, it is also true for

$$\hat{Q}i - \hat{Q}\hat{U}. \quad (17.11)$$

So (17.11) is the mathematical statement of the condition for the symmetry of a physical situation under the reflection \hat{U} . It shows a symmetry.

17-3. Symmetry and conservation

Before applying the result we have just found, we should like to discuss the idea of symmetry a little more. Suppose we have a very special state here after we operate only once with \hat{Q} , we get the same state. This is a very special case, but let's suppose it happens to be true for a state $|\psi\rangle$ and that $|\psi'\rangle = \hat{Q}|\psi\rangle$ is physically the same state as $|\psi\rangle$. That means that $|\psi'\rangle$ is equal to $|\psi\rangle$ except for some phase factor f . However, then happens? For instance, suppose that we

[†]In fact, if you can show that \hat{Q} is necessarily a Hermitian operator—which requires that the operator \hat{Q} is a real number times [a] complex number or both the sum of such numbers— \hat{Q} is a small enough time to be close to the identity. Any operator like a reflection in a mirror doesn't have any part of its hermiticity lost if $|\psi'\rangle$ and $|\psi\rangle$ have the same basis; they can only differ by a non-imaginary phase factor.

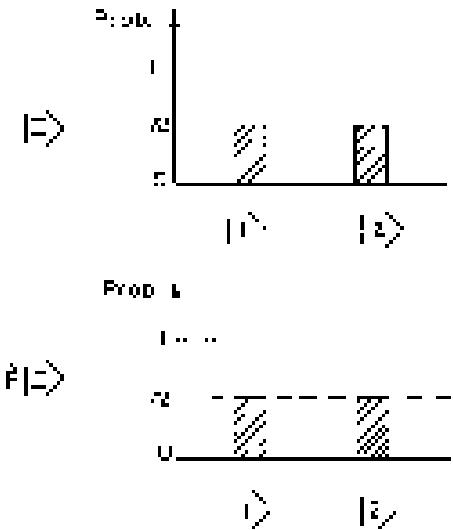


Fig. 17-3. The state $|1\rangle$ and the reflected $|2\rangle$ recorded by reflecting $|1\rangle$ in the central plane.

over in Fig. 17-2 for the case which we considered $|A\rangle$. There it is said that if the input amplitude is ψ_1 at the base states $|1\rangle$ and $|2\rangle$, the probabilities are shown as in Fig. 17-2(a). If we operate on $|1\rangle$ with the reflection operator R , it flips the state over changing $|1\rangle$ to $|2\rangle$ and $|2\rangle$ to $|1\rangle$, so that the probability is shown in Fig. 17-2(b). But notice just the state $|B\rangle$ is left over again. In other words, the state $|B\rangle$ has probabilities before and after reflection look just the same. However, there is a difference if we look at the amplitudes. For the state $|B\rangle$ the amplitudes are the same also, but the state $|B\rangle$ has amplitudes having opposite signs. In other words,

$$P(|B\rangle = P(|1\rangle) = \frac{|\psi_1|}{\sqrt{2}} = \frac{\psi_1 + i\psi_2}{\sqrt{2}} = \psi_1.$$
 (17.12)

$$P(|B\rangle = P(|2\rangle) = \frac{|\psi_1|}{\sqrt{2}} = \frac{\psi_1 - i\psi_2}{\sqrt{2}} = -\psi_2.$$

Now notice $\psi_1 = \psi_2$ unless $\theta = 0$ for the state $|B\rangle$ and $\theta \neq 0$ for the state $|B\rangle$.

Let's look at another example. Suppose we have a Rabi polarized photon propagating in the z-direction. If we do no rotation of the wave function, the z-component, we know that this just rotates the amplitude by $e^{i\theta}$ where θ is the angle of rotation. So for the situation we had in this case, ψ is just equal to the angle of rotation.

Now it is clear that \hat{Q} corresponds to the wave function operator $\hat{\psi}$ just as \hat{Q} is the plane of the state of the wave function $\psi = 0$ is free to rotate. In other words, if we come up with ψ , you can find the state ψ just after a rotation.

$$\hat{Q}(\psi, \psi_1) = |\psi_2\rangle \quad (17.13)$$

and if the symmetry of the equation makes you say that

$$\hat{Q}(\psi) = e^{i\theta} |\psi_2\rangle. \quad (17.14)$$

Then it is also true that

$$\hat{Q}(\psi_2) = e^{i\theta} |\psi_2\rangle. \quad (17.15)$$

This is clear, even

$$\hat{Q}(\hat{Q}\psi) = \hat{Q}(\psi, \psi_2) = \hat{Q}(\psi),$$

and if $\hat{Q}(\psi) = e^{i\theta} |\psi_2\rangle$, then

$$\hat{Q}(\psi_2) = \hat{Q}(\hat{Q}\psi, \psi_2) = \hat{Q}(\psi, \psi_2) = \hat{Q}(\psi).$$

The second of equalities follows from (17.13) and (17.15) for a unitary real operator from (17.14) and then the last two because $e^{i\theta}$ commutes with its conjugate.

So both certain symmetries something which is true initially is true for all times. But isn't that just a conservation law? Yes, it says that if you look at the right-hand side and the working out is complicated on the side because that an operator \hat{Q} is the necessary expression of the special procedure, take a unitary operation by a certain phase, then you know that the symmetry will be true of the final state. The same result can multiply the final state by the same phase factor. This is always true even though we may not know exactly what the internal mechanism of the converse which changes a system from the time to the final state. Even if we would like to look at the details of the machinery by which the system goes from one state to another, we can still see that if a thing is in a state with a certain symmetry character originally and if the time evolution for the many \rightarrow symmetrical under that symmetry operation, then the state ψ will have the same symmetry character for all times. That's the basis of all the conservation laws of quantum mechanics.

Let's look at a special example. Let's go back to the ψ^H system. Now we'll show that it really satisfies our definition of ψ^H . We want to take for P not just a

minus reflection, because that requires defining the plane in which we put the mirror. There is a special kind of reflecting that doesn't require the specification of a plane. Suppose we redefine the operation \hat{P} this way: First you reflect in a mirror in the x -plane so that's done, $\psi_1 \rightarrow \psi_1$, and then rotate you... in the system I'll reflect in a mirror so that's done to go to $-x$ and $y \rightarrow -y$. The whole thing is called an inversion. Every point is reflected through the origin to the diametrically opposite position. All the coordinates of everything are reversed. We will still use the symbol \hat{P} for the operation. It's shown in Fig. 13-1. It is a little more complicated than a simple reflection because it doesn't require that you specify what's called a plane normal for the reflection — you just specify only the point where it is, no other axes symmetry.

Now let's suppose that we have a state $|\psi\rangle$, which under the wave-sink operation goes to $|\psi'\rangle$, $\psi' = \hat{P}\psi$.

$$\hat{P}|\psi\rangle = P\hat{P}_1|\psi\rangle = P\hat{P}_1\psi_1 = |\psi'\rangle \quad (13.16)$$

Now suppose that we invert again. After two inversions we are right back where we started from — nothing is changed at all! We must have that:

$$\begin{aligned} \hat{P}^2|\psi\rangle &= \hat{P}\hat{P}_1|\psi\rangle = |\psi'\rangle \\ \text{Hence } \hat{P}^2 &= \hat{P}\hat{P}_1 = P\hat{P}_1P = (P\hat{P}_1)^2 = \hat{P}\hat{P}_1. \end{aligned}$$

It follows that

$$(\hat{P}\hat{P}_1)^n = \hat{P}^n\hat{P}_1^n.$$

So if the inverse operation is a symmetry operation of a state, there are only two possibilities for \hat{P} :

$$\hat{P}^2 = \pm 1,$$

which means that

$$\hat{P}_1|\psi\rangle = |\psi'\rangle \quad \text{or} \quad \hat{P}_1|\psi\rangle = -|\psi'\rangle. \quad (13.17)$$

Classically, if a state is symmetric under an inversion, the operation gives back the same state. In quantum mechanics, however, there are the two possibilities: we get the same state or we have the same state, $\hat{P}_1|\psi\rangle = |\psi\rangle$, but the state $|\psi\rangle$ has imaginary. Other the sign is inverted so that $\hat{P}|\psi\rangle = -|\psi\rangle$, we say that the state has odd parity. (See inversion rule.) \hat{P} is also known as the parity operator. The state $|\psi\rangle$ of the HJ has even parity, and the state $|\psi'\rangle$ has odd parity — see Eq. (13.17). There are, of course, cases which are not symmetric under the operation \hat{P} ; these are states with no definite parity. For instance, in the HJ system the state $|\psi\rangle$ has even parity, the state $|\psi'\rangle$ has odd parity, and the state $|\psi''\rangle$ has no definite parity.

When we speak of an operation like inversion being performed "on a physical system" we can think about it in two ways. One can think of physically moving a greater body to the inverse point $x' = -x$ in we run the risk of breaking up the same system from a new frame of reference $x' = y$, $y' = z$, $z' = -x$ the body $x = -x$, $y = -y$, and $z = -z$. Similarly, when we think of rotation, we can think of rotating bodily a physical system, or rotating the coordinate frame in the space in which we measure the system. Keeping the system "fixed" in space. Generally, the two points of view are completely equivalent. For rotation they are equivalent since first rotating a system by an angle θ is like rotating the reference frame by the negative of θ . In these lectures we have usually considered what happens when a projection is made in a new set of axes. What you get this way is the same as what you get if you leave the axes fixed and rotate the system back-and-forth by the same amount. What you do, that the signs of the angles are, *etc.*

[†]In older books you may find formulas with different signs; they are probably using different definitions of the angle.

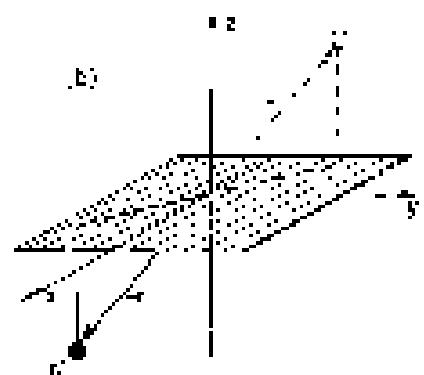
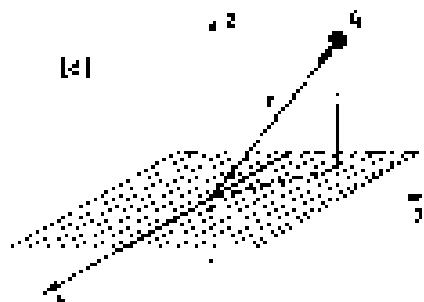


Fig. 13-1. The operation of inversion, \hat{P} . Whatever is at the point A at $t = t_0$ is moved to the point A' at $t = t_0$, $\hat{P} = \pm 1$.

Many of the laws of physics—both classical and quantum mechanics—involve the conservation of the coordinates. They are invariant with respect to an inversion. The laws of fluid dynamics, for instance, are unchanged if we change $x \rightarrow -x$, $y \rightarrow -y$, and $z \rightarrow -z$ in all components. This symmetry is important in gravity, and for the strong interactions of nuclear physics. Only the weak interactions are responsible for beta decay. So, let's take this symmetry. (We discussed this in more detail in Chapter 13, Sec. 13.1.) We will for now leave out any consideration of the β decay. Let's imagine any physical system where β decay is not expected to produce any appreciable effect. An example would be the emission of light by an atom—the same atom at different positions in space-time. Under these circumstances we have the following proposition. If a state originally had even parity, and if you look at the original β state at some later time, it will again have even parity. Furthermore, suppose β is an odd-parity operator in a way that it preserves even parity. If you look at the whole thing—including the photon state—the result since it is β again have even parity (because if you start with odd parity). The principle is called “*conservation of parity*.” (Even though many words (“*conservation of parity*” and “*selection symmetry*”) are closely interwoven in the quantum mechanics. Although until a few years ago it was thought the nature always conserves parity, it is now known that this is not true. In the beta decay of cobalt because the β decay reaction does not care the intrinsic symmetry rules is violated in the rule laws of physics.)

Now we can prove an interesting theorem (which is equivalent to various discrete-space theorems): Any state of definite energy which is not degenerate must have a definite parity. It must have either even parity or odd parity. Remember that we have a situation in which several states have the same energy—in fact such states are degenerate. Our theorem will not apply to them.)

For a state $|\psi\rangle$ in \mathcal{H} , if \mathcal{Q} is the energy, we know that

$$\mathcal{Q}|\psi\rangle = E|\psi\rangle \quad (7.13)$$

where E is just a number—the energy of the state. If we take any operator \mathcal{Q} which is a symmetry operator of the system we can prove that

$$\mathcal{Q}|\psi\rangle = c^*|\psi\rangle \quad (7.14)$$

in which c is a complex $\neq 0$ definite energy. Consider the two states $\langle\psi|\psi\rangle$ and $\langle\psi|\psi'\rangle$ from operating with \mathcal{Q} . If the physics is symmetric, $\langle\psi|\psi\rangle$ and $\langle\psi|\psi'\rangle$ the same energy as $|\psi\rangle$. But we have taken a situation in which there is only one state of that energy, $\langle\psi|\psi\rangle$ and $\langle\psi|\psi'\rangle$ are the same state—it is unity. So ψ is a photon. That's the physics argument.

The same thing comes out of our mathematics. Our definition of symmetry is $\mathcal{Q}|\psi\rangle = E|\psi\rangle$ in Eq. (7.13) (and for any state ψ ,

$$c\mathcal{Q}|\psi\rangle = \mathcal{Q}c|\psi\rangle \quad (7.15)$$

But we are considering only a state $|\psi\rangle$ which is a definite energy state, so that $\mathcal{Q}|\psi\rangle = E|\psi\rangle$. Since E is just a number (in Newtonian terms through \mathcal{Q} if we want), we have

$$\mathcal{Q}\mathcal{Q}|\psi\rangle = \mathcal{Q}E|\psi\rangle = E\mathcal{Q}|\psi\rangle$$

$$E\mathcal{Q}|\psi\rangle = P(\mathcal{Q}|\psi\rangle) \quad (7.16)$$

So $\mathcal{Q}|\psi\rangle = \mathcal{Q}^*|\psi\rangle$ is also a definite energy state of \mathcal{R} and with the same E . But by our logic, there is only one such state, it has to be that $\mathcal{Q}|\psi\rangle = c^*|\psi\rangle$.

What we have just proven is true for any operator \mathcal{Q} that is a symmetry operator of the physical system. Therefore, in a situation in which we consider only electrical forces and strong interactions—and in particular the first interaction symmetry is an accurate approximation, we have that $\mathcal{Q}|\psi\rangle = c^*|\psi\rangle$ (as we have also seen earlier in most chapters of this book). So any state of a definite energy (which is not degenerate) has just either an even parity or an odd parity.

13-3 The conservation law

We turn now to another interesting example of an operator: a rotation. We consider the special case of a dipole, i.e., rotating an atomic system by angle θ around its axis. We will call it operator $\hat{\theta}$ (though this is going to suggest that we have a physical system where we have no influences force up along the z -axis). And that is how a magnetic field is taken to be parallel to the z -axis so that there will be no change in the external conditions if we rotate the whole physical system about the z -axis. For example, if we have an atom in empty space and we rotate the electron around the z -axis by an angle θ , we have the same physical system.

Now then, there are wavefunctions which have the property that after an operation you get a new state which is the original one multiplied by some phase factor. Let me make a quick argument to show you just what I mean. If you trust the plane waves that always be proportional to the $e^{i\vec{k}\cdot\vec{r}}$. Suppose that you would measure only the amplitude. That's because this was the way by definition. If it's proportional to $e^{i\vec{k}\cdot\vec{r}}$ the effect of the rotation is given by a phase shift so that

$$\hat{\theta}(\phi_1, \phi_2) = e^{i\vec{k}\cdot\vec{r}} \phi_1,$$

two such rotations in succession would multiply the state by the factor $e^{i\vec{k}^2} = e^{i\vec{k}_1^2 + \vec{k}_2^2}$, thus

$$\hat{\theta}(\phi_1, \phi_2) = \hat{\theta}(\phi_1) e^{i\vec{k}_1^2} \phi_1 = e^{i\vec{k}_1^2} \hat{\theta}(\phi_1) \phi_1 = e^{i\vec{k}_1^2} \phi_1.$$

The phase change δ must be proportional to \vec{k}^2 . We are considering here those special states ϕ_1, ϕ_2 which

$$\hat{\theta}(\phi_1, \phi_2) = e^{i\vec{k}^2} \phi_1, \quad (13.2)$$

where \vec{k} is some real number.

We also know the remarkable fact that if a system is represented by a wavefunction and \vec{k} is the original one it happens in every case property that $(1+2)$ is true, then it will also have the same property later on. So this number \vec{k} is a very important one. If we know its value initially, we know its value at the end of the process. This is a number which is conserved during a sequence of no inputs. This means that we put out \vec{k} because it hasn't anything to do with the initial and final state because you can't do something in classical mechanics. In quantum mechanics we don't care so little—for each state of a single angular momentum about the z -axis. If we do that we find that in the limit of large systems the state quantity is equal to the x -component of the angular momentum of classical mechanics. So if we have a system for which a conservation law can never just produce a plus or minus, then we have a sense of definite angular momentum about that axis and the angular momentum is conserved. It is not new and forever. Of course, you can't talk about any axis and you just let it be a question of what the parameter for the various axes. You see that the conservation of angular momentum is related to the fact that when you rotate a system you get the same state with only a new phase factor.

One would like to show you how general this idea is. We will apply it to two other conservation laws which have exact correspondence in the physical ideas in the conservation of angular momentum. In classical physics we also have conservation of momentum and conservation of energy, and it is interesting to see that both of these are related in the same way to an operator by another.

¹ Very easily, we could also do it by a rotation of the physical system by $-\phi$: that the result will be the same is making the incident frame by $+\phi$.

² We can always choose a frame in which one of the fields provided there is only an z -field of a time, and its direction doesn't change.

³ For a finite pulse we should make this argument for small angles ϕ . Since one angle is the n^{th} derivative of time, $\phi = m_e K(\phi) = (\hbar/e)^2$ and the total phase change is n^{th} order. One will argue, and it is, therefore, proportional to ϕ .

Suppose that we have a physical system, an atom, some molecules or a atom, or a molecule, or some big atom if you want make any difference. If we take this whole system and move it over to a different place. So we have a Hamiltonian which has the energy that is independent of the state of movement of the atoms, and they may depend on the **displaced position** in space. Under these circumstances there is a special symmetry because we can perform a shift in space. Let's define $\delta \vec{r}$ as the separation of a displacement by the distance along the \vec{r} axis. Then for any state we can make this operation and get a new state. Well again there can be very specific states which lose the property that when you displace them by a along the \vec{r} axis you get the same state except for a phase factor. It's also possible to prove just as we did above that when this happens the phase must be proportional to $e^{i\omega t}$. In we can write ψ this special state $|\psi\rangle$

$$|\psi(\vec{r})\rangle = e^{i\omega t} |\vec{r}\rangle \quad (11.22)$$

The coefficient δ , when multiplied by \hbar , is called ω , an **operator of the momentum**. And the reason it is called that is that the number is numerically equal to the classical momentum p . So we have a large system. The general situation is this. If the Hamiltonian is unchanged when the system is displaced, and if the state starts with a definite momentum in the \vec{r} direction, then the momentum is the same even will remain the same in this system. The total momentum of a system before and after a collision after replacing \vec{r} will be the same.

There is another operation that's quite analogous to the displacement in space. I call it δt . Suppose that we have a physical situation where there is something constant that depends on time, and we start something off at a certain moment in a state, make and let it roll. Now, if we were to start the same thing again (in exactly the same way) our example is everything, except by a time δt and if nothing in the external environment depends on the absolute time, the development would be the same and the final state would be the same as the other. This result, except that it will give them after by the factor $e^{i\omega \delta t}$. Under these circumstances we can also find special states which lose the property that the development in time has the **exact** characteristic that the displaced state is just the old one plus, by a phase factor. That means it's easier than the most general since the phase change must be proportional to i . We can write

$$|\psi(t+\delta t)\rangle = e^{i\omega \delta t} |\psi(t)\rangle \quad (11.23)$$

It is recommended to use the negative sign in defining ω , with this convention about the sense of the system and its movement. So a system of finite energy is one which when displaced in time reproduces itself multiplied by $e^{-i\omega \delta t}$. (This is what we have said before when we did, in a quantum state of definite energy, an ω in a wavefunction with no \vec{r} -dependence.) It means that the system is in a state of definite energy, and the Hamiltonian doesn't depend on t , then no matter what, ω in the system will have these features at all later times.

You see, therefore, one rule is however the conservation law and the symmetry of the world. Symmetry with respect to displacement in time implies the conservation of energy, symmetry with respect to position in \vec{r} implies the conservation of the \vec{p} component of momentum. Symmetry with respect to rotation around the x , y , and z axes implies the conservation of the p_x , p_y , and p_z components of angular momentum. Symmetry with respect to reflection implies the conservation of parity. Symmetry with respect to the exchange of two directions implies the conservation of something we don't have a name for, and so on. Some of these principles have classical analogs and others do not. There are four conservation laws in quantum mechanics, but the one most universal mechanics, in at least, than we usually need most.

In order that you will be able to read other books on quantum mechanics, you must make a small adjustment since to describe the notation that people use. The operation of a derivative with respect to time is, of course, just the expression $i\hbar \partial / \partial t$.

time t , it's worth looking at it:

$$\delta_t |\psi\rangle = \delta(\epsilon - \tau, t) |\psi\rangle \quad (17.25)$$

We'd prefer like to do it in terms of infinitesimal displacements in time, or in terms of infinitesimal displacements in space, or in terms of rotations through infinitesimal angles. Since any finite displacement in a system can be reconstructed by a sequence of infinitesimal displacements, we might as well just do it in the infinitesimal case. The operator that infinitesimally displaces an object in time by ϵ is what we call *displacement*¹ in Chapter 3:

$$D(\epsilon|\psi\rangle) = |1 + \frac{i}{\hbar} \epsilon \hat{p}\rangle |\psi\rangle \quad (17.26)$$

Then D is analogous to the classical quantity we call *energy*, because if $D|\psi\rangle$ converges to be a constant (that is, if) $D|\psi\rangle = |\psi\rangle$, then that constant is the energy of the system.

The same thing is done for the other operators. If we displace a small displacement in x , say by the amount ϵ , a state $|\psi\rangle$ will, in general, pass over into some other state $|\psi'\rangle$. We can write

$$|\psi'\rangle = D_x(\epsilon|\psi\rangle) = \left(1 + \frac{i}{\hbar} \epsilon \hat{p}_x\right) |\psi\rangle \quad (17.27)$$

and as ϵ goes to zero, the $|\psi'\rangle$ should become just $|\psi\rangle$ or $D_x(0) = 1$, and so, could be the change of $D_x(\epsilon|\psi\rangle)$ from $|\psi\rangle$ could be proportional to ϵ . Define \hat{J}_x to be the operator p_x , i.e., the momentum operator—for the measurement of position.

For identical reasons, people usually write the small relation as

$$D_x(\epsilon|\psi\rangle) = \left(1 - \frac{i}{\hbar} \epsilon \hat{J}_x\right) |\psi\rangle \quad (17.28)$$

and call \hat{J}_x the *operator* of the component of angular momentum. For those quantitites for which $D_x(\epsilon|\psi\rangle) = \text{const}$, we can take any small angle—say don't expand the right-hand side to first order in the ϵ and get

$$D_x(\epsilon|\psi\rangle) = e^{i\epsilon \hat{J}_x} |\psi\rangle = (1 - i\epsilon \hat{J}_x) |\psi\rangle$$

Comparing this with the definition of J_x in Eq. (17.28), we get that

$$J_x |\psi\rangle = i\epsilon \hat{J}_x |\psi\rangle \quad (17.29)$$

In other words, if you rotate with J_x by a small finite angular increment ϵ , the state $|\psi\rangle$ is left in the same state; J_x is the amount of component of angular momentum. It is quite analogous to operating on a definite energy since each θ to get $\delta_\theta |\psi\rangle$.

We should now “do the obvious application of” the ideas of this chapter, i.e. of angular momentum—we show you how they work. The point is not easy to be really very simple. You knew before that angular momentum is conserved. The only thing you really have to remember from this chapter is that “a state $|\psi\rangle$ has to be property, but spin is nothing, though it's single, about the axis,” to be sure $i\epsilon \hat{J}_x |\psi\rangle$; it has to be something of angular momentum around θ , not ϵ . That's where θ has to be a number of it, to begin things.

17-4 Polarized light

What it all we would like to check is, for example, in Section 17-3 we showed that when RHC polarized light is viewed in a frame rotated by the angle $\pi/2$ about the \hat{z} -axis, it gets multiplied by $e^{i\pi/2}$. Does that mean then that the photons of light

¹except for the sign with the negative of the one we used in Section 17-3.

Left and right circularly polarized carry an angular momentum of one unit along the wave's direction of travel. If also incoherent, if we have a beam of light containing a large number of photons all circularly polarized the same way, so we would have in a classical view, it will carry angular momentum. If the total energy carried by the beam is E , then the angular momentum is $E/2\pi c \hbar$. Why? Because each photon carries the angular momentum \hbar , so there is a total angular momentum E/\hbar .

$$J = E \frac{\hbar}{c} \quad (7.10)$$

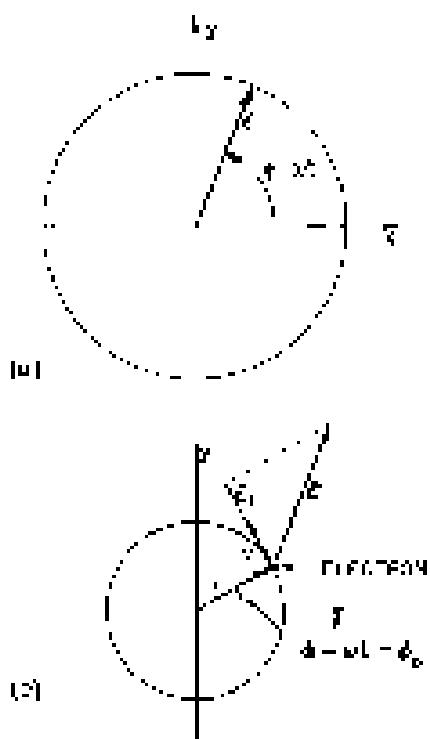


Fig. 17-5. (a) The electric field E in a circularly polarized light wave. (b) The motion of an electron being driven by the circularly polarized light.

Can we prove classically that light which is right circularly polarized carries an energy and angular momentum in proportion $E/\hbar c$? That that is a classical proposition of everything is right. Were we to take a test where we can go along the quantum theory in the classical limit. We should set up the classical physics checks. It will give us an idea whether we have a right to call it the angular momentum. Remember what circularly polarized light is, classically—it's described by an electron oscillating in equilibrium about an oscillating mean position. At a rate of ω , so that the results are periodic, as in Fig. 17-5(a), shown in Fig. 17-5(b). Now in space such light comes out in a wave which is going to absorb it, or at least some of it, and consider an atom in the wave according to the classical physics. We have often described the motion of the electron in the atom as a harmonic oscillator when one drives into resonance by an external electric field. Well suppose that the atom is isotropic, so that it can vibrate equally well in the x -or y -directions. This is the "circular polarized light, the transverse wave" approximation; displacement and velocity here are 90° behind the wave. The net result is that the electron moves in a circle, as shown in Fig. 17-5(b). The electron is displaced at some distance r from its equilibrium point about the origin and goes around with some frequency with respect to the wave ω . The relation between ω and r might be as shown in Fig. 17-5(c). As time goes on, the wave is held constant and the displacement rotates with the same frequency, so that relative motion goes clockwise some. Now, why should the work being done on this electron. The rate and energy is being put into this electron is \propto its velocity. Let's do the component of v_θ parallel to the velocity.

$$\frac{dv_\theta}{dt} = qE_r \quad (7.11)$$

But look, there is angular momentum being passed into the electron, because E_r is always in r -direction about the origin. The torque is qEr , which must be equal to the rate of change of angular momentum dJ/dt :

$$\frac{dJ_r}{dt} = qEr \quad (7.12)$$

Remembering that $r = r_0 \cos \omega t$, we have that:

$$\frac{dJ_r}{dt} = \frac{1}{2} \omega r^2$$

Therefore, if we integrate the total impulse component which is absorbed, it is proportional to the total energy, the situation. A proportionality being $1/\omega$ which agrees with Eq. (7.10). Light does carry angular momentum—but it carries it in light circularly polarized along the axis, and $= 1$ unit along the axis. It is left circularly polarized.

Now let's ask the following question: If light is linearly polarized in the direction, what is its angular momentum? Light polarized in the x -direction, can be represented as the superposition of RHC and LHC polarized light. There fore there is a certain amplitude for the angular momentum J_x and another

¹ It is usually very difficult to measure angular momentum of atomic systems, $\pm 1/2$. They can only have spin one-half, so with no angular momentum $m = 1/2$ with respect to any axis. Of course, if that the z -component of angular momentum is m . You don't need to align the x -axis the time.

amplitude that the angular momentum is $-A$, so it doesn't have a definite angular momentum. It has an amplitude to appear with $+A$, and an equal amplitude to appear with $-A$. The interference of these two amplitudes produces the $\langle A \rangle$ polarization, but it has equal probabilities to appear with plus or minus one unit of angular momentum. The netopic measurements made on a beam of linearly polarized light will show that it is this, a random orientation, because in the $\pi/2$ chamber of photons there are many more numbers of $|\text{D}\pi/2\rangle$ and $|\text{U}\pi/2\rangle$ photons contributing significantly amounts of angular momentum—the average angular momentum is zero. And in the classical theory you don't find the angular momentum unless there is some circular polarization.

We have said that a composite particle can have three values of J_z , namely $-J$, 0 , $+J$ (the transitions we saw in the Stern-Gerlach experiments). But light is screwy; it has only two states. It does not have the transitions. The strings back is related to the fact that "light comes and will. For a particle of spin J which is spinning ω , there are $J(J+1)$ possible states with values of J going in steps of 1 from $-J$ to $+J$. But it turns out that for given J there's going to be only the states with the components $\pm J$ along the direction of rotation ω . The example light does not have three states, but with the ω although a photon is still a object of spin, one thing is not true about it is that it is not in a state "based on what happens in the rotational space—that the said one particles contributes are necessary". That a particle of total angular momentum made about ω without changing the measurement axis. Particles were measured twice (the previous and the next) cannot be about, only rotations about the axis along the direction of motion do not change the orientation of the angular momentum. Rotations around ω does not only see influence to prove that these states are required, give that all of the ω axes are ω under rotations by the angle φ .

One further cold remark. For a very fast ω or like a general, uniform, rotation of the two spin states with respect to the line of motion ($-J$, $+J$) is really necessary. For instance when we spin one half particles—only the parts were the component of angular momentum according to the equation of states $= J/2$ exist at least (and only) magnetic moment ($\pm J/2$ for magnetons). When angular momentum is ω , only (so the energy is conserved, as it is for light) high components ($\pm J$, 0 , $-\omega$) are required.

17-6 The disintegration of the Λ^0

Now we want to give an example of how we use the theorem of conservation of angular momentum in a specifically quantum physical problem. We look at break up of the Lambda particle (Λ^0), which disintegrates into proton and a pion via a "weak" interaction.

$$\Lambda^0 \rightarrow p + \pi^-.$$

Assume we know that the pion has spin zero, that the proton has spin one-half, and that the Λ^0 has spin one-half. We would like to solve the Λ decay problem. Suppose that a Λ^0 were to be produced in a way that cause it to be completely polarized—by which we mean that its spin is, say, "up," with respect to some coordinate system (see Fig. 17-6a). Then what is ψ , in what probability will it disintegrate into the proton gives off an angle θ and π^- goes in the ϕ axis? In Fig. 17-6(b) to decide odds, what is the angular distribution of beta-decay? (Time!) We will look at the easiest case in the coordinate system in which the Λ^0 is at rest; we will measure the angle θ this set the ψ in Fig. 17-6a always be transverse to another frame if we were

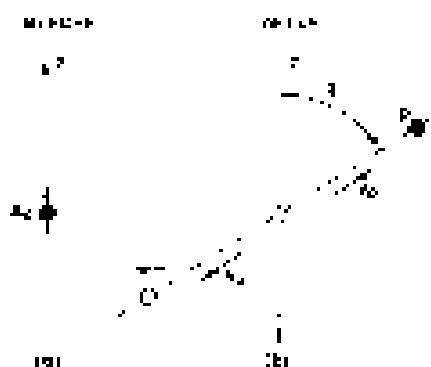


Fig. 17-6. A Λ^0 with spin "up" decays into a proton and a pion [in the CM system]. What is the probability that the proton will go off at the angle θ ?

* We have tried to make it sound as if the component of angular momentum along the direction of motion must form some more particle, but in reality, the type of C-S—and not something like A/λ . Even among all sorts of forms for all the Lorentz invariants, we can take just, we have. There is no law—well have a look about it—but Pauli-Weyl are known about such things.

BC-STATE	EF-STATE
↑↑	↑↑
↑↓	↑↓
↓↑	↓↑
↓↓	↓↓
yes	no
(a)	(b)

Fig. 17.7. Two possibilities for the decay of a spin-up ↑ particle into two going along the \hat{z} -axis. Only (b) conserves angular momentum.

We begin by looking at the second circumstance in which the proton is emitted into a small \hat{z} , the angle subtending the axis (Fig. 17.8). Before the proton is sent on its way it is both spin-up ↑, in part of the system. After a short time—say, 10 microseconds—the hadron is still there except that it is connected with the \hat{z} through the "spin-down" proton and a gluon. Suppose the gluon goes up along the \hat{z} -axis. Then, from the conservation of momentum, the proton must go down. Since the proton is a baryon half as large as a pion, it may be either "up" or "down". This is a principle, but it is possible, as shown in parts (b) and (c) of the figure.

The conservation of angular momentum, however, requires that the proton have spin $+1/2$. This is not really seen from the following reason. A particle moving along \hat{z} has zero spin and can be a very singular number in an absolute sense. If the \hat{z} is a regular baryon, only the spins can contribute to it. They can angular momentum about the \hat{z} with $1/2$ below the degeneracy, so it must also be $-1/2$ (Section 16.2). We can say that since the proton has no spin, the proton spin must be "up".

If you are worried that arguments of this kind may not be valid in quantum mechanics, we can always return to classical theory. The initial state (before the fragmentation), with $\psi = \psi_0(\hat{z})$, spin $+1/2$ has the property that "it is rotated about the \hat{z} " by the angle ϕ . The state after fragmentation by the pion has $\psi' = \psi_0(\hat{z}')$. (In contrast to standard quantum mechanics, ψ' is not a spin-half particle.) Since nature's behavior doesn't depend on our choice of axes, the final state ψ' must also have "it is rotated about the \hat{z}' ". We could write the final state as, for

$$\text{proton going } +\hat{z}, \text{ spin } +1/2; \text{ pion going } -\hat{z}'.$$

It is really no surprise to classify the gluon momenta, since in the frame we are using, the gluon always moves toward the proton, \vec{p}_g and \vec{p}_p are the components of the final momentum

$$\text{proton going } +\hat{z}, \text{ spin } +1/2.$$

Now what happens to this wave vector ψ' when we rotate the coordinates about the \hat{z}' axis by the angle ϕ' ?

Since the gluon and pion are moving along the \hat{z}' -axis of the system, but are separated by the rotation, (that's why we pick the coordinate axes to make the argument otherwise), also, nothing happens to the pion's momentum if it is spinless. The proton, however, has now one-half. If its spin is " $+1/2$ " it will undergo a phase change of $e^{i\phi'}$ in response to the rotation. All its spin were down, the phase change leaves the proton with spin " $-1/2$ ". But its phase change with rotation before and after the fragmentation must be the same if angular momentum is to be conserved. (And it is!) In, since there is no sensible interaction in the fragmentation, the only possibility is that the value spin will be " 0 ," little about going up; its spin must also be " up ".

This is realistic. Let's let the initial value of angular momentum be only the proton's spin, $\psi_0(\hat{z})$ going up (\hat{z}). This spin does not permit the process shown in part (a). Since we know that the fragmentation occurs, there is some amplitude for process (b) (proton going up with spin " up "). Well, we demand to the amplitude for the diagrams we see in this case to be $\psi_0(\hat{z}')$ instead of ψ_0 .

Now let's see what would happen if the "spin orientation" \hat{z}' alone. Again we ask about the decay on \hat{z}' : The proton goes up along the \hat{z}' -axis as shown in Fig. 17.8. You will appreciate that in this case the gluon must have spin "down" if angular momentum is to be conserved. Let's say the amplitude for such a diagram is a .

We can't say anything about the two amplitudes a and b . They depend on the angular momentum of \hat{z}' and on the energy, and we've yet to prove how to

SEEDS	EF-STATE
	↑↑
	↑↓
	↓↑
	↓↓
yes	yes
(b)	(c)

Fig. 17.8. The decay along the \hat{z}' -axis for a ψ^2 with spin " up ".

* Now a more sophisticated, but the important, role the angularity of the proton mechanics is extremely important to you. You can speak about a gluon wave function and take the time to understand it in mathematical detail. In practice, however, you need not do this; you have the general idea, especially in the first few pages of the section.

undetectable from. We'll have to get them from experiment. But with just these two amplitudes we can find out what we want to know about the angular distribution of the disintegration. We only have to be careful always to define completely the states we are talking about.

We want to know the probability that the proton will go off at the angle θ with respect to the x -axis (into a small solid angle $d\Omega$) as drawn in Fig. 17-6. Each point is one α -particle in this direction and each is the x -axis. We know how to specify what happens along this axis. With respect to this new axis, the α no longer has its spin "up," but has a certain amplitude to have its spin "up" and another amplitude to have its spin "down." We have already worked these out in Chapter 6, and again in Chapter 16, Eq. (16.96). The amplitude to be spin "up" is $a \cos \theta/2$ and the amplitude to be spin "down" is $b = -a \sin \theta/2$. When the α spin is "up" along the x -axis it will emit a photon in the y direction with the amplitude a . So the amplitude to find an " $-y$ "-spinning photon coming out along the x' direction is

$$a \cos^2 \frac{\theta}{2}. \quad (17.33)$$

Similarly, the amplitude to find a "down"-spinning photon coming along the post x' is b , which is

$$-b \sin^2 \frac{\theta}{2}. \quad (17.34)$$

The two processes that these amplitudes refer to are shown in Fig. 17-7.

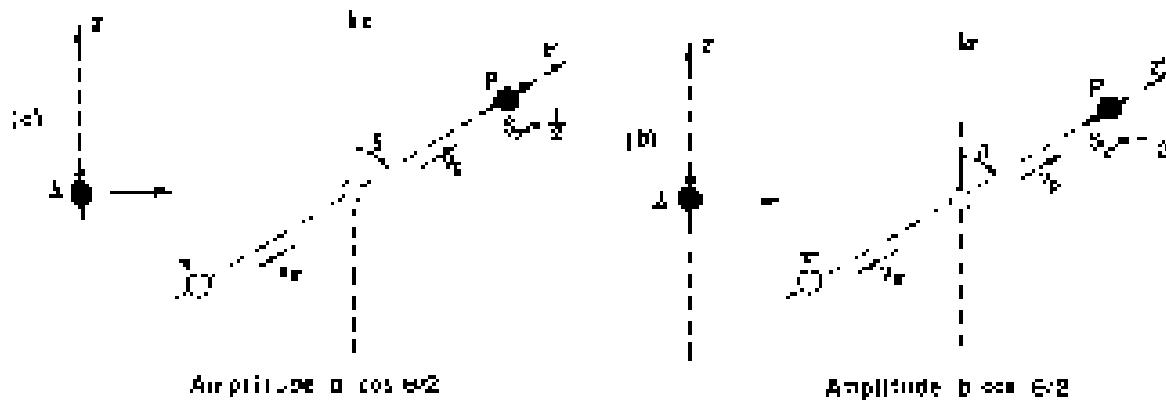


Fig. 17-7. Two possible decay states for the α .

Let's now ask the following easy question. If the α has spin up along the z -axis, what is the probability that the decay process will go off at the angle $\theta/2$? The two spin states ("up" or "down" along z) are distinguishable even though we are not going to look at them. So to get the probability we square the amplitude. The probability $f(\theta)$ of finding a proton in a small solid angle $d\Omega$ is

$$f(\theta) = |a|^2 \cos^2 \frac{\theta}{2} = \hat{p}^2 \sin^2 \frac{\theta}{2}. \quad (17.35)$$

Re-arranging that $\sin^2 \theta/2 = |\hat{p}|^2 = \cos^2 \delta$ and $\cos^2 \theta/2 = \sin^2 \delta$, we can write $f(\theta)$ as

$$f(\theta) = \left(\frac{a_i^2 + b_i^2}{2} \right) - \left(\frac{|p|^2 - \delta|^2}{2} \right) \cos \theta. \quad (17.36)$$

¹ We have chosen to use a or b in the complex and type the matrix elements for $S_{\alpha}(E)$. You would get the same answer for any other choice.

The angular distribution has the form

$$\langle \hat{p}_y \rangle = \langle \psi | \hat{p}_y | \psi \rangle = 0 \quad (17.35)$$

The probability δ has one part that is independent of θ and one part that varies linearly with $\cos \theta$. From measuring the angular distribution we can get δ and α and therefore, β and $|\psi\rangle$.

Now there are many other questions we can answer. Are we interested only in particles with spin "up" along the x -axis? Each of the terms in (17.35) and (17.36) will give us amplitude to find a particle with spin "up" along your spin "Down" and respect to the x -axis ($|+z\rangle$ and $| -z\rangle$). Spin "Up" will respect to the y -axis ($|+y\rangle$) and be orthogonal in terms of the basis states $|+z\rangle$ and $| -z\rangle$. We can then measure the two amplitudes (17.35) and (17.36) and the proper coefficients from 6.2 and (17.21) to get the total amplitude

$$\left(\cos^2 \frac{\theta}{2} + \delta \cos^2 \frac{\theta}{2} \right).$$

This square is the probability that the proton comes out of the angle θ with its spin the same as the ψ^k ("up" along the x -axis).

If parity were conserved, we could say one more thing. The deautlerization of Fig. 17.8 is just the reflection—in say, the x -plane of the deautlerization of Fig. 17.7†. If parity were conserved it would have a coefficient of $+1$ or -1 . Then the coefficient α of (17.37) would be zero, and the deautlerization would be equally likely to occur in all directions.

The experimental result is shown, however, that there is an asymmetry in the deautlerization. The measured angular distribution, over the range $0 < \theta < \pi/2$, is plotted in an additional line in Fig. 17.8. It has the form $|\psi\rangle = |\psi^k\rangle + \alpha \sin \theta |\psi^k\rangle$, and it is not zero, and

You see how much we can get from the conservation of angular momentum. We will give some more examples in the next chapter.

Conversely: By the methods in this section we can find the amplitude and the state $|\psi\rangle$ knowing $\langle \psi | \hat{p}_y | \psi \rangle$ is obtained in an additional line in Fig. 17.8 from the state $|\psi\rangle = |\psi^k\rangle + \alpha \sin \theta |\psi^k\rangle$, and it is not zero, and

$$\text{Spin-orbiting } \langle \psi | \hat{p}_y | \psi \rangle = H |\psi\rangle, \text{ spin } \langle \psi | \hat{p}_y | \psi \rangle = i\hbar\omega. \quad (17.38)$$

where H is the Hamiltonian of the world, or, at least, of whatever is reasonable for the theory. The conservation of angular momentum means that the Hamiltonian must have the property that

$$\text{Cyclic gauge } \langle \psi | \hat{p}_y | \psi \rangle = H |\psi\rangle, \text{ spin } \langle \psi | \hat{p}_y | \psi \rangle = i\hbar\omega. \quad (17.39)$$

By the amplitude β we mean that

$$\text{Spin-orbiting } \langle \psi | \hat{p}_y | \psi \rangle = H |\psi\rangle, \text{ spin } \langle \psi | \hat{p}_y | \psi \rangle = i\hbar\omega. \quad (17.40)$$

Conservation of angular momentum implies that

$$\text{Cyclic gauge } \langle \psi | \hat{p}_y | \psi \rangle = H |\psi\rangle, \text{ spin } \langle \psi | \hat{p}_y | \psi \rangle = 0. \quad (17.41)$$

If the amplitudes given in (17.33) and (17.34) are the ones, we can express them mathematically as follows. By (17.41) we know the amplitude and the β have spin along \hat{p}_y will decompose into a power having along the \hat{p}_y -direction with its sign plus to the \hat{p}_y -direction, or, similarly the amplitude

$$\langle \text{spin-orbiting } -\hat{p}_y, \text{ spin } +\hat{p}_y | H | \psi \rangle, \text{ spin } \langle \psi | \hat{p}_y | \psi \rangle = 0. \quad (17.42)$$

By the general methods of quantum mechanics this amplitude can be written as

$$\sum_i \langle \text{spin-orbiting } -\hat{p}_y, \text{ spin } +\hat{p}_y | H | \psi \rangle, \text{ spin } \langle \psi | \hat{p}_y | \psi \rangle, \quad (17.43)$$

†Remembering that the spin is a vector and the spin is the rotation.

where the sum is to be taken over the base states $\langle A, \beta \rangle$ of the system it creates. Since the hyperfine spin quantum number is zero, we can write this in the form

$$\begin{aligned} \text{operator giving } & \langle A, \beta | \psi_1(\vec{x}', \vec{p}') \rangle \langle \vec{A}, \beta | \psi_2(\vec{x}_1, \vec{p}_1) \rangle \\ & + \langle \psi_1(\vec{x}', \vec{p}') | \psi_2(\vec{x}_1, \vec{p}_1) \rangle \langle A, \beta | \psi_1(\vec{x}', \vec{p}') \rangle \langle \vec{A}, \beta | \psi_2(\vec{x}_1, \vec{p}_1) \rangle \end{aligned} \quad (17.4)$$

The first factor of the first term is a_1 , and the last factor in the second term is a_{22}^* . From the definition of ψ_1 in (17.1), and from (17.4) which is now fully symmetric by momentum conservation, the maximum factor $\langle A, \beta | \psi_1(\vec{x}', \vec{p}) \rangle$ in the first term is the diagonal wave function for half particles with spin β , p' along $\vec{\beta}$ and will have Lorentz spin β , p' along $\vec{\beta}$ and other α at the angle θ , which is one of $2\pi/3$ or $4\pi/3$. So (17.4) is just a scaled Ψ as we wrote in (17.3). The amplitude of (17.4) follows from the same kind of arguments for a spin "doubt" β -particle.

17.6 Summary of the rotation matrices

We would like now to bring together in one place the various things we have learned about the rotations for the fields of spin one-half and spin one. As they will be needed for further calculations, on the next page you will find tables of the two rotation matrices $R_1(\alpha)$ and $R_2(\beta)$ for spin one-half particles, for spin one particles, and for photons (spin one) with zero rest mass. For each spin we will give the terms of the matrix $\langle \psi | \vec{\psi}' \rangle$ corresponding to each component of the particle. They are, of course, exactly equivalent to the amplitudes like $\langle 1, \beta | 0, \beta' \rangle$ we have used in earlier chapters. We can see by $R_1(\alpha)$ that the state is expressed in a new coordinate system which is rotated enough to single out just the non-scattering always the angular part due to defining the primitive sectors of the rotation. By $R_2(\beta)$ we mean that the reference axes are rotated by the angle β about the \hat{n} axis. Knowing these two rotations, you can, of course, work out any arbitrary rotation. As usual, we write our \rightarrow as \rightarrow elements in the ring and our \leftarrow as a free state of the field (rotated!) charge and the state of the field as a base state of the old (unrotated) frame. You can inspect the entries in the tables in many ways. For instance, the entry $+e^2$ in Table 17.1 means that the \rightarrow \rightarrow element $\langle -| \psi | + \rangle = e^2$. If you take β so that $\hat{n} = j = e^{-i\beta^2/2}$, or that $\langle - | \psi | + \rangle = e^{i\beta^2/2}$ it's all the same thing.

Table 17-1

Rotation matrix for spin-1/2

Two states: $|+\rangle$, "up" along the z axis, $m = +1/2$
 $|-\rangle$, "down" along the z axis, $m = -1/2$

$R(\theta)$	$ +\rangle$	$ -\rangle$
$ +\rangle$	$\cos \theta/2$	0
$ -\rangle$	0	$\sin \theta/2$

$R(\theta)$	$ +\rangle$	$ -\rangle$
$ +\rangle$	$\cos \theta/2$	$\sin \theta/2$
$ -\rangle$	$-\sin \theta/2$	$\cos \theta/2$

Table 17-2

Rotation matrices for spin one

Three states: $|+\rangle, m = +1$
 $|0\rangle, m = 0$
 $|-\rangle, m = -1$

$R(\theta)$	$ +\rangle$	$ 0\rangle$	$ -\rangle$
$ +\rangle$	$\cos \theta$	0	0
$ 0\rangle$	0	1	0
$ -\rangle$	0	0	$\sin \theta$

$R(\theta)$	$ +\rangle$	$ 0\rangle$	$ -\rangle$
$ +\rangle$	$(1/2 + i\sin \theta)$	$+i\sqrt{3}/2 \cos \theta$	$(1/2 - i\sin \theta)$
$ 0\rangle$	$-i\sqrt{3}/2 \cos \theta$	$\cos \theta$	$+i\sqrt{3}/2 \cos \theta$
$ -\rangle$	$i(1 - i\sin \theta)$	$-i\sqrt{3}/2 \cos \theta$	$i(1 + i\sin \theta)$

Table 17-3

Theta

Two states: $|X\rangle = \frac{1}{\sqrt{2}}(|+1\rangle - |+2\rangle), m = -1$ (LHC polarized)

$Z = \frac{1}{\sqrt{2}}(|X\rangle - i|Y\rangle), m = -1$ (LHC polarized)

$R(\theta)$	$ X\rangle$	$ Y\rangle$
$ X\rangle$	$e^{-i\theta}$	0
$ Y\rangle$	0	$e^{-i\theta}$

Angular Momentum

18-1 Electric dipole radiation

In the last chapter we developed the idea of the conservation of angular momentum in quantum mechanics and showed how it might be used to account for regular oscillations of the proton from the deionization of the hydrogen atom. We want now to give you a number of other, direct illustrations of the consequences of momentum conservation in atomic systems. Our first example is the radiation of light from an atom. The conservation of angular momentum (among other things) will determine the polarization and angular distribution of the emitted photons.

Suppose we have an atom which is in an excited state of definite angular momentum—say a spin of one—so it makes transitions to states having no magnetic field at a lower energy, emitting a photon. The problem is to figure out the angular distribution and polarization of the photon. If the nucleus is allowed exactly the same as the π^0 deionization, namely that the spin quantum number (spin one-half particles) would be opposite to the atom's spin one, there are three possibilities for the component of angular momentum. The value of m would be $+1, 0$, or -1 . We will take $m = +1$ for our example. Once you see how it goes, you can work out the others as well. We suppose that the atom is rotating with its angular momentum along the $+z$ axis (see Fig. 12-10), and that, with amplitude a along its right-hand polarized \hat{y} axis, it is also rotating. In addition, that the atom ends up with zero angular momentum— \rightarrow shown in part (b) of the figure. (We don't know the answer to that. But we do know that most unpolarized light has one unit of angular momentum out of its direction of propagation.) So after the photon is emitted, the electron would have to be as shown in Fig. 18-1(b)—the atom is again averaging to no magnetism

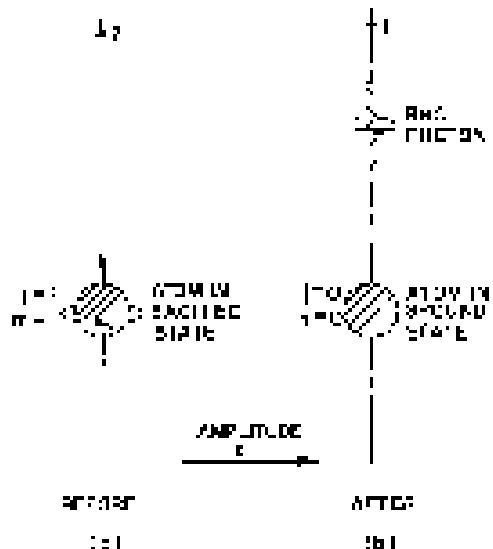


Fig. 18-1. An atom with $m_s = +1$ emits a Rabi photon along the \hat{y} -axis.

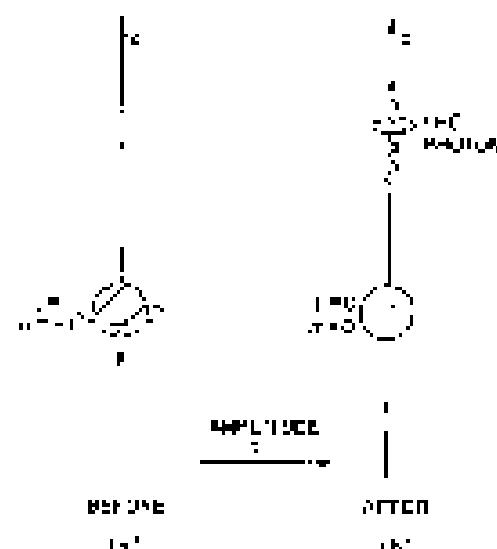


Fig. 18-2. An atom with $m_s = -1$ emits a Rabi photon along the $-\hat{x}$ -axis.

18-1 Electric dipole radiation

18-2 Rabi scattering

18-3 The annihilation of positrons

18-4 Rotation matrix for any angle

18-5 Measuring a nucleon spin

18-6 Conservation of angular momentum

Added Note 1: Derivation of the selection rules

Added Note 2: Conservation of parity in photon annihilation

about the \hat{z} -axis, where we have assumed an atom whose low- $|m|$ state is spinless. We will let a stand for the amplitude for such an event. More precisely, we let a be the amplitude to emit a photon from a certain atomic configuration α , oriented on the \hat{z} -axis, during a time Δt . Notice that the amplitude to emit a LHC photon in the \hat{x} -hat direction is zero. The net angular momentum about the \hat{z} -axis would be -1 . Such a photon would form the system for a total of -1 , which would not conserve angular momentum.

Similarly, if the spin of the atom is initially “Down” ($-\frac{1}{2}$ along the \hat{z} -axis), a downward \rightarrow LHC polarized photon in the direction of $\hat{x}\hat{y} = -\hat{x}\hat{z}$, as shown in Fig. 18-2. We will let b stand for the amplitude for this event, meaning again the amplitude that the photon goes in a certain solid angle $d\Omega$. On the other hand, if the atom is in here $+$ state, a horizontal photon has a different, or all, possible photons can have only one angular momentum $+1$ or -1 along \hat{z} . Check!

Now, we’re going to do something new. Suppose we perform an inversion of the situation in Fig. 18-1, which means that we should imagine what the system would look like if we were to move each part of the system to an equivalent place on the opposite side of the origin. This does not mean that we should reflect the angular momentum vectors, because they are artificial. We should, rather, “turn the usual character” of the motion that would be responsible for such an angular momentum. In Fig. 18-1(a) and (b) we show what the process of Fig. 18-1 looks like before and after an inversion with respect to the center of the atom. Notice that the orientation of the atom is unchanged! In the inverted system of Fig. 18-1(c) we have an atom with $m = +1$ oriented $\hat{x}\hat{y}$ and \hat{z} oriented downward.

If we now rotate the system of Fig. 18-1(b) or 18-1(c) about the \hat{z} -axis, it becomes identical to Fig. 18-2. The combination of the inversion and rotation turns the second process into the first. Using Table 18-1, we see that a rotation of 180° about the \hat{z} -axis just changes $m = -1$ to state inversion $m = +1$, while so the amplitude b is not equal to the amplitude a except for a possibly sign change due to the inversion. The sign change in the transition was dictated by the parity of the initial and final state of the atom.

In other processes, parity is conserved, so the parity of the whole system must be the same before and after the position exchange. What happens will depend on whether the parities of the initial and final states of the atom, or even overall, — the angular distribution of the radiation will be different for different cases. We will take the common case of odd parity for both states and even parity for the transition, as will give what is called “electric dipole radiation.” (If the initial and final states have the same parity we say this is “magnetic dipole radiation,” which has the consequence of the radiation from an atom being a wave in a fixed direction of the \hat{z} -axis.) This is odd, since parity conserves angular momentum, which takes the system from (a) to (c) of Fig. 18-3. The final state of the atom, of course, ± 1 , its amplitude doesn’t change sign. That means it’s going to conserve parity, or equivalently it must be $a \pm b$, to a magnitude less than the amplitude total of the opposite sign.

We conclude that if the amplitude is a that an $m = +1$ atom will emit a photon upward, then for the general parity of the initial and final states the amplitude that an $m = -1$, atom will emit a LHC photon upward is $-a$.

We have all we need to know to find the amplitude for a photon to be emitted at any angle θ with respect to the \hat{z} -axis. Suppose we have an atom originally polarized with $m = -1$. We can consider the state only $+1$, 0 , and -1 states with respect to a fixed frame in the direction of the photon emission. The amplitudes for these three states are just the ones given in the lower half of Table 7-2.

[†] When we change (x, y, z) into $(x, -y, -z)$ you might think that all nature is reversed. That’s not true; just certain like displacements and velocities, but not for all. An inversion preserves momentum—or any vector that is a derived quantity such as a product of two particle vectors. Recall however that the electric component after an inversion.

[‡] You may notice in the argument we have just made on the basis that the final state we have been considering can not have a definite parity. You will find in added Note 2 at the end of this chapter another item of interest, when you may prefer.

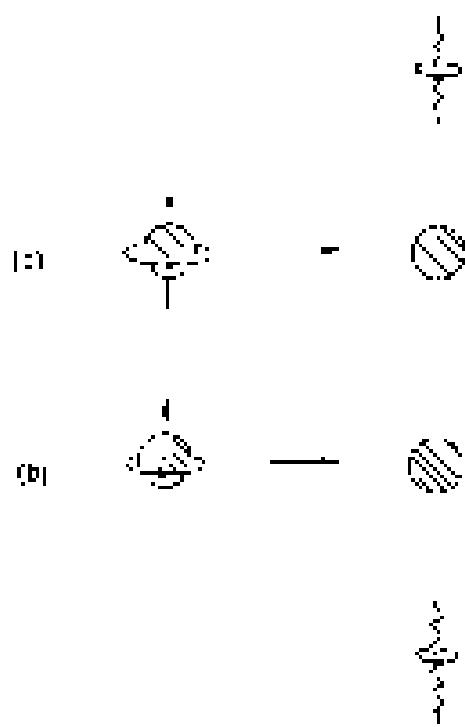


Fig. 18-3. If the process of (b) is transformed by an inversion through the center of the atom, it appears as in (d).

The amplitude for a RHC photon is emitted in the \hat{S} direction is the same as the amplitude for an LHC photon in the \hat{S} direction, except,

$$A_S = A_R(\theta) | \hat{S} \rangle = \frac{\pi}{2} C_1 \cos \theta. \quad (18.1)$$

The amplitude for the LHC photon is emitted in the \hat{S} direction is $-A_S$, unless the proper direction is $\hat{S} = -\hat{S}$ in the new direction. (See Table 17.2, if you like.)

$$-A_S = -A_R(\theta) = -\frac{\pi}{2} C_1 \cos \theta. \quad (18.2)$$

If you are interested in other polarizations you can find an amplitude formula from the superspace box. Use two amplitudes. To get the intensity of any polarization, just sum all of them; you must of course take the absolute square of the amplitudes.

18-2 Light scattering

Let's use these results to solve a somewhat more complicated problem, but also one which is considerably more real. We suppose that the source terms are sitting in the \hat{S} (around \hat{S}) ($\hat{S} = \hat{S}_0$) and scatter an incoming beam of LHC light. (Assume that the light is using only the \hat{S} cell, so no scattering outside \hat{S}). There photons coming up to the beam from the \hat{S} direction, as shown in Fig. 18-4(a). We can calculate the scattering of light in a two-step process. The photon is absorbed, and then is re-emitted. If we start with a RHC photon as in Fig. 18-4(a), and angular momentum is conserved, the beam will come in at $\hat{S} = +\hat{S}_0$ after the absorption, as shown in Fig. 18-4(b). We call the amplitude for this process A_R^* . The photon then emits a RHC photon in the direction \hat{S} , as in Fig. 18-4(c). The total amplitude that a RHC photon is scattered in the direction \hat{S} is just $A_R^* + A_R$. Let's call this scattering amplitude $(R^* + R)$; we have

$$(R^* + R) | \hat{S} \rangle = \frac{\pi}{2} C_1 (1 - \cos \theta). \quad (18.3)$$

There is also an amplitude R^* for a RHC photon \hat{S}' to be absorbed and that a LHC photon with be emitted. The product of these amplitudes has an $\langle R^* | S' | S \rangle = 6$, but a RHC photon is scattered as a LHC photon. Using (18.2), we have

$$\langle R^* | S' | R \rangle = -\frac{\pi^2}{2} (1 - \cos \theta). \quad (18.4)$$

Now let's look about what happens. The LHC photon comes in. With it is scattered, because with $\hat{S}' = -\hat{S}_0$. By these methods of angular \hat{S} we used in the preceding section, we can show that the amplitude must be $\langle R^* | S' | S \rangle = 1$. The amplitude that an electron in the $\hat{S} = -\hat{S}_0$ state will emit a RHC photon of the angle θ is minus the amplitude $(-1) R^*(\theta) = -1$, which is $1 - \cos \theta$. So we have

$$(R^* - S | R \rangle = -\frac{\pi^2}{2} (1 - \cos \theta). \quad (18.5)$$

Finally, the amplitude for a LHC photon to be scattered as a LHC photon is

$$\langle R^* | S | R \rangle = \frac{\pi^2}{2} (1 - \cos \theta). \quad (18.6)$$

(This is the two-photon states which cancel.)

If we make a measure unit of the scattered amplitude for the two-photon decay of an electron plus its bremsstrahlung, the frequency of one of the four amplitudes. For instance, with an incoming beam of LHC light the frequency of the RHC photon in the scattered radiation will be $|1 - \cos \theta|$!

This is π^2 very low, but suppose we scatter with linearly polarized light. What then? If they have \hat{S} -polarized light, it can be represented as a superposition



Fig. 18-4. The scattering of light by electrons via a two-step process.

XKCC and LHC light See note (see Section 1.4)

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|X\rangle + |L\rangle) \quad (18.7)$$

Or, if we have polarized light, we would have

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|R\rangle - |B\rangle) \quad (18.8)$$

Now what do you want to know? Do you want the amplitude that an unpolarized photon will scatter into a RHC photon at the angle? You can get it by the usual rule for summing amplitudes. First, multiply (18.7) by $\langle R' | S | X \rangle$:

$$\langle R' | S | \psi \rangle = \frac{1}{\sqrt{2}}(\langle R' | X | R \rangle + \langle R' | L | R \rangle), \quad (18.9)$$

and then use (18.1) and (18.2) for the two amplitudes. You get

$$\langle R' | S | \psi \rangle = \frac{\alpha}{\sqrt{2}} \cos \theta \quad (18.10)$$

If you wanted the amplitude that an x photon would scatter into a LHC photon, you would get

$$\langle R' | S | \psi \rangle = \frac{\alpha}{\sqrt{2}} \sin \theta. \quad (18.11)$$

Finally, suppose you want to know the amplitude that an unpolarized photon will scatter while keeping its polarization. What you want is $\langle Y | S | \psi \rangle$. This turns out to be

$$\langle R' | S | \psi \rangle = \langle R' | S' \rangle \langle R' | S | \psi \rangle + \langle R' | U | S' | S | \psi \rangle. \quad (18.12)$$

If you take just the real part,

$$|S'\rangle = \frac{1}{\sqrt{2}}(|R'\rangle - |B'\rangle), \quad (18.13)$$

$$|U\rangle = \frac{1}{\sqrt{2}}(|R'\rangle + |B'\rangle), \quad (18.14)$$

it follows that

$$\langle R' | B' \rangle = \frac{1}{\sqrt{2}}, \quad (18.15)$$

$$\langle R' | U \rangle = \frac{1}{\sqrt{2}}. \quad (18.16)$$

So you get that

$$\langle R' | S | \psi \rangle = \alpha \cos \theta. \quad (18.17)$$

This means that a beam of unpolarized light will scatter from an electron (at the x point) with an intensity proportional to $\cos^2 \theta$. If your light is polarized light, you had this:

$$\langle Y | S | \psi \rangle = 0. \quad (18.18)$$

So the unpolarized light is completely polarized in the y -direction.

Now we see something interesting. The result (18.17) and (18.18) correspond exactly to the classical theory of light scattering we gave in Vol. I, Section 3.6. When we imagined that the electron was bound in the atom by a linear restoring force—so that it acted like a classical oscillator, perhaps you are thinking: “It’s no such oscillator in the classical theory, so it gives the right answer only when $\theta = 0$ and $\theta = 90^\circ$.” For one thing, we have considered an isotropic unpolarized—the y -glow—case, which from $\theta = 0$ to $\theta = 90^\circ$ is excited with $E_y = 0$ ground state. If the excited state had split too, you would get a different result. Also, there is no reason why the model of an electron attached to a

spins and orbits by an overlapping electron. It should work for a single proton. But we know that it does not work, and at the present time and interest—incorrect right. And in a certain sense we are bringing the whole theory around to the real truth. Otherwise we have, in fact, a theory of two index of refraction, and of light scattering. By the classical theory, we have now shown the true quantum theory. Now the same result for the more common case. In fact we have now that the polarization of sky light, for instance, by *quantum* the two index approach, which is the only fully legitimate way.

I should like, of course, that there be theories which work on *suppose*, ultimately, by legitimate quantum arguments. Naturally, those ranges which we have used a great deal in calculating, or you see selected from us three parts of classical physics which still have some validity in quantum mechanics. You'll notice that we did not do so in *quantum* my model of the atom which has electrons just positioned in orbits. That's because such a model doesn't give results which agree with quantum mechanics. But the answer is a strong—*why* is not in a sense, is it?—the way in which “looks”—does work, because we used that model for the theory of the index of refraction.

10-3. The annihilation of positronium

We would like now to take an example which is more physical. The question concerning spin, although somewhat hypothetical, we hope not too much so. Our example is the system called “positronium,” which is an “atom” made up of an electron and a positive, a bound state of one and one e^+ . It's like a hydrogen atom, except that a positron replaces the proton. The angular law-like, the hyperfine along many axes. Also like the hydrogen, the ground state is split into a “positive structure” by the interaction of the magnetic moments. The spins of the electron and positron are both enabled, and they can be either parallel or antiparallel to any given axis. (In our ground state there is no other angular momentum in the Kondo model.) So there are five states. These are the only states of a symmetric system, all with the same energy, and one is a sort of twin very near a different energy. The energy splitting is, however, much greater than the 1-20 megacycles of hydrogen, because the reaction magnetic moment is so much larger, 1900 times stronger, than the present moment.

The most important difference, however, is that positronium cannot live forever. The position is the coordinate of the electron; they can annihilate each other. The two particles have opposite momentum, taking energy into each other which appears as two γ -ray photons. In the diagram, this is a good way: I have two massless two or more objects worth have zero rest mass.*

We begin by analyzing the disintegration of the spin-zero state of the positronium. I pointed out that the $\gamma\gamma$ decay with a lifetime of about 10^{-10} second. Initially, we have a positron and an electron moving here and with spins opposite, making the descrete system. After the disintegration there are two photons going off with equal and opposite momenta (Fig. 10-5). The momenta of each equal and opposite because the total momentum after the disintegration must be zero, as it was before, if we are taking the case of annihilation as zero. If the result could be not at 180°, we can't do $\psi\psi$ to solve the problem, and then from it in the ordinary way to the lab system. (So as we can do anything now; we have all the ways.)

First, we note that the angular distribution is not very interesting. Since the initial state has spin zero, it has no “spin” axis. It's something made out of two. The final state must then also be symmetric under π -rotations. This means that all singles the two single photons are equally likely—the amplitude is the same for a photon to go at $\theta = 0$ as at $\theta = \pi$. (Remember, since we have one of the photons in some effect on the other must be opposite.)

* In the chapter immediately after the word “way,” we do not have an easy way to distinguish who has the energy and power in the “masses” from the energy of an electron, because in your mind you think the particle has been “separated.” The true difference is that the present are *isolated* mass.

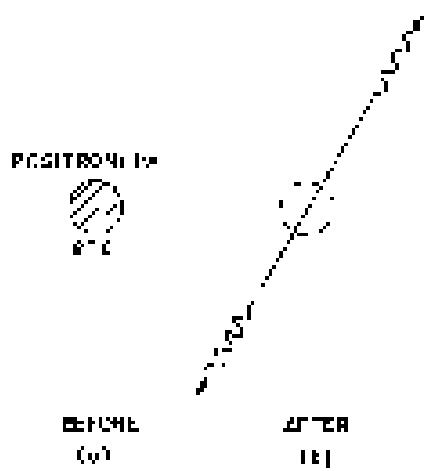


Fig. 10-5. The descrete annihilation of positronium.

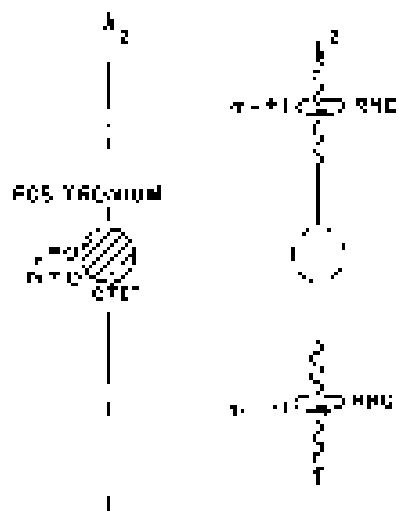


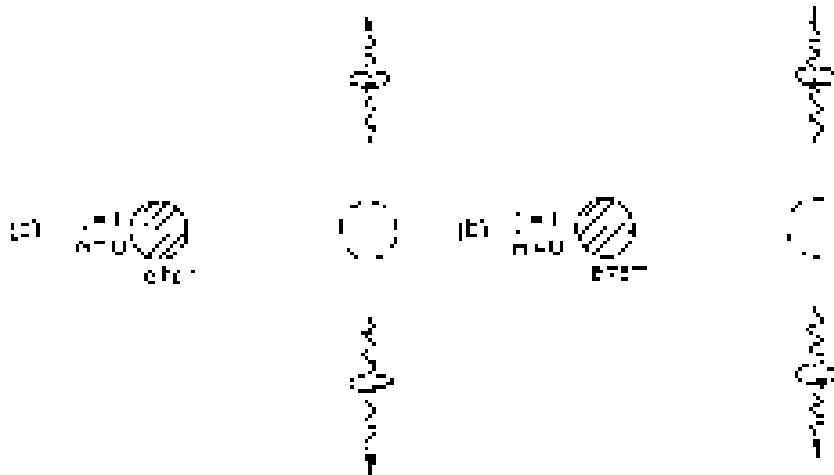
Fig. 1B a. Optimality for continuous mobility along the road.

The only remaining question, which we now want to focus on, is about the orientation of the directors. Let's call the directions of rotation of the two phasers as α (horizontal) and β (vertical). We ignore any rotation of one w.r.t. the other (parallel-axes rule for the rotations); we also choose the first description (right-angle) for a particular phaser rotation, always with respect to the orientation of rotation. Right-angle orientation (but if the phaser is going upward & RPS), then an angular momentum will be generated at the downstream polar position (RUC). It will carry $\pm \frac{1}{2}$ and $\mp \frac{1}{2}$ angular momentum components in its downstream direction, which means plus and minus one unit around the axis. The total will be zero, and the angular momentum after the steering system will be the same as before. See Fig. 1(b).

The static arguments show that if the upward helix placed a RHC, the downward can't be LHC. Then, the first state wouldn't be a unit of angular momentum - this is not guaranteed if the initial one has spin zero. Note that such a state is also not possible for the outer positronium ground state of spin zero, because it has three maxima of one unit of angular momentum in one ring; i.e.

Now we want to see if two-particle annihilation is not possible at all from these constraints. You might think that "well ok, just take a state which has zero angular momentum about the \hat{z} -axis" shows it's like no-gyrotron state, and could deinterfere with the ADL patterns. Certainly, the deinterference sketches in Fig. 18-2(a) conserves angular momentum about the \hat{z} -axis. But now look what happens if we rotate this system around the \hat{x} -axis by 180° : we get the picture shown in Fig. 18-2(c). It's exactly the same as in part (a) of the figure, *but* we have done a full change the two particles. If one particle is a base particle, if we flip it around, its amplitude has the same sign, so the amplitude for the deinterference is per (b) must be the same as in part (a). But we have assumed that the initial photons are spin-one. And when we rotate a spin-one object as a state with $m = 0$ (as 180°) about the \hat{x} -axis, its amplitudes change sign (see Table 17-2 for $e = \pi$). So the amplitudes for (3) and (4) in Fig. 18-2 would have opposite signs, and we can't make conservation of angular momentum work.

When $\lambda = 0$, then you would expect it to end up in the stationary state. All of the time and in the beginning, there exists $\mu = -1/6$, or $+1/3$ of the mass. So half of the time you would expect position and velocity. The other half



Hg. 10-7. For the $\beta \rightarrow \gamma$ state of polarization, the process (a) and (b) (Rb) relation between β 's are exactly the same.

* Note that we always calculate C_2 up to one decimal place above the diagonal or below the π^0 particle. If we try to go to 30% beyond the angle, we run into trouble with any other pions, as we will have to work around the possibility of "self" angles—meaning angles with a π^0 particle. For example, we can't say that the photons came really from the center of the experiment. They could have been coming from us from the far left or something like that. We don't have to worry about such problems unless we are interested in the distribution of medium.

of the two photons beam splitter and detector. That's after multiplying, but it agrees with previous work. The harder part is to do that and the time is 1000 times longer about 10⁻⁷ second. This is what is observed experimentally. We will now go into very much of the details of the spin calculations.

We know that if we only carry about single momenta, the spin-wave state of the descretional doublet has two LHC photons. There is also another possibility, it can split into two LHC photons as shown in Fig. 18-6. The next question is, where is the heavier because the amplitudes for these two possibilities decay inversely. We can find out from the conservation of parity.

To do that, however, we need to know the parity of the positive ion. Now theoretical physicists have come in a way that is not very to explain, but the parity of the electron and the position of its antiparticle must be opposite, so then the spin-wave quantum state of positronium must be odd. We will just assume that it is odd, and since we will get agreement with experiment, we can take that as sufficient proof.

Let's see then what happens if we make an antisym of the process in Fig. 18-6. When we do that, the two photons receive different spin polarizations. The inverted paraxial lenses look like Fig. 18-4. Assuming that the parity of the positronium is odd, the amplitudes for the two processes in Figs. 18-6 and 18-5 must have the opposite sign. Let's let |R, R> stand for the final state of Fig. 18-6 in which both vectors are RHC, and |L, L> stand for the final state of Fig. 18-5, in which both photons are LHC. The final state (labeled in Fig. 18-6) is

$$|\psi\rangle = |\text{R}, \text{R}\rangle - |\text{L}, \text{L}\rangle \quad (18.19)$$

The amplitude changes the RC in Eq. 18-3 and gives the value

$$\langle \psi | \psi \rangle = \langle \text{R}, \text{R} | \psi \rangle = -\epsilon_{123} = -1 \quad (18.20)$$

which is the negative of Eq. 18-3. So the final state |ψ⟩ has negative parity, which is the same as the fine spin wavestate of the positronium. This is the only final state that conserves both angular momentum and parity. There is zero amplitude that the incident ion into the state will occur, which we don't need to worry about here, however, since we are only interested in spin conservation the polarization.

What does the lifetime of Eq. 18-5 bear physically? One thing it does is the following. If we observe the two photons in two detectors which can't be able to detect separately the RHC or LHC photons, we will always see two RHC photons together, or two LHC photons together. That is, if you stand on one side of the positronium and examine the光 on the opposite side, you can measure the polarization and tell me whether it what polarization he got yet. You have a 50-50 chance of catching a RHC photon or a LHC photon; whichever one you get, you can prove that it will get the same.

Since there is a 50-50 chance to RHC or LHC polarization, it sounds as though it might be an linear polarization. Let's ask what happens if we measure the photon in counter that accept only linear polarized light. We change it is not as easy to measure the polarization as it is for light; there is no polarization when waves are, for such short wavelengths. But for something that is, to make the distinction easier. Suppose that you have a counter that only accepts light with polarization, and that there is a gap on the other side that also looks for linear polarization with, say, z-polarization. What is the chance that we pick up the two photons from an antineutron? What we need to ask is the amplitude that

ψ will be a state where I either ends up one, the amplitude

$$\langle \psi | \psi \rangle = P$$

which is, of course, just

$$\langle \psi | \psi \rangle = |\psi\rangle \langle \psi| = P \quad (18.21)$$

Now although we are working with two particle amplitudes for the two photons, we can hardly do just one. So single particle amplitudes, since

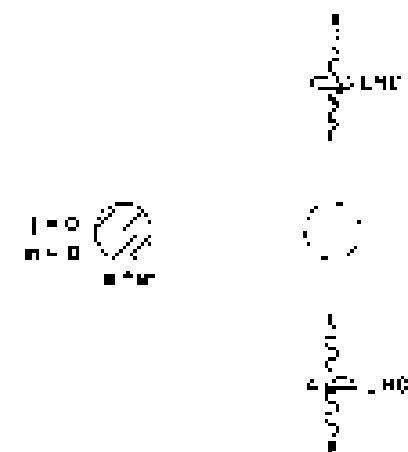


Fig. 18-6. Another doublet process for positron annihilation.

and α_2 does independently of the other. That means that the amplitude $(\alpha_1, \alpha_2) R_1 R_2$ is the product of the two independent amplitudes (α_1, R_1) and (α_2, R_2) . Using Eqs.(18.1), these two amplitudes are $1/\sqrt{2}$ and $i/\sqrt{2}$, so

$$(\alpha_1, \alpha_2) R_1 R_2 = -\frac{i}{2}.$$

Similarly, we find now

$$(\alpha_1, \alpha_2) L_1 L_2 = -\frac{i}{2}.$$

Substituting these two amplitudes into Eq.(18.2) we get that

$$(\alpha_1, \alpha_2) V_{L,R} = 0. \quad (18.2)$$

So there is a 50% probability that if you get a photon in your polarization detector, the other guy will also photon in his polarization detector.

Now suppose that the other guy has a counter for a photon in his. In some way yours. He would never get a count when you present. If you work it through, you will find that

$$(\alpha_1, \alpha_2) S = 0. \quad (18.22)$$

I will, naturally, also work out that if you set your counter for unpolarization, he will get a count about once only if he set his polarization.

Now this all leads to an interesting situation. Suppose you were to set up something like a series of switches which separated the photons into polarized and unpolarized beams, and put a counter in each beam. Let's say that the beam on the left goes to the other side goes the same thing you can always tell him which beam his photon is going to go into. Whether you and he get the photons to do, you can see which of your detectors caught the photon and he will know which of his counters had a photon. Let's say that in a certain chain you can tell that a photon went in to your detector; you can tell him that so must have been a counter in his parameter.

Now, if any people think quantum mechanics is not useful (and I have indeed seen that), they would like to think that since the photons are sent in groups along with a given wave function, they would think that since "they carry photon" has some "hampered" from polarization or unpolarization, there would be some chance of picking it up ... with the counter and that this chance shouldn't depend on whom some other person didn't set about a completely arbitrary photon. They might then "argue" that "since one detector shouldn't be able to change the probability that I will find something". One quantum mechanics says, however, that by looking at a source and on whose detector one you can precisely predict when the particle (or photon in either case) is going to be set, it is detected. This point was never accepted by Einstein, and he wanted to say it is a foolish "argument" because known as the "Einstein-Podolsky-Rosen paradox". But when the situation is described as we have done, it then does doesn't seem to be any paradox at all; it seems quite naturally that what is measured in one place is correlated with what is measured somewhere else. The argument that there is no "hidden" assumption here, like this:

- (1) If you have a counter which tells you whether your photon is RHC or LHC, you can predict exactly what kind of a photon (RHC or LHC) he will find.
- (2) The photons are received just, respectively, say, by partly RHC or partly LHC, some of one kind and some of the other.
- (3) Surely you cannot alter the physical nature of the photons by changing the kind of observation you make on your photons. No matter what measurements you make on your, his must still be either RHC or LHC.

I have here not mentioned our angle, which is multiplied times the amplitude for the transmission into the particular field wave, but we can see that there is no effect because we get the same probability when we look at the other side counter. See Eq.(18.21) that.

- (d) Now suppose he changes his setup so instead of 10 photons he has ten randomly polarized beams with a piece of plastic so that $\frac{1}{2}$ of his photons go either forward or backward, turning it into a 50% left-right beam. Let's say absolutely no way according to quantum mechanics, $\frac{1}{2}$ will go right, $\frac{1}{2}$ will go left. If 1 photon will pass. Then there is 50% probability it will pass to the x-beam and a 50% probability it will pass to the y-beam. And the same goes for a 17% photon.
- (e) Since each photon is RHC or LHC according to (c) and (d) each one must have a 50/50 chance of going into the x sector or the y beam and there is no way in practice which one it will go.
- (f) Yet the theory predicts that if you do your photon path through an experimenter you can set up each existing LHC and RHC will go into his y deflected beam. This is in contradiction to (c) so the theory is a paradox.

No... I apologize. I don't see the "paradox," however, because experimenter choice has the potentiality of biasing the results. We have already discussed the key to this "paradox" in our very first discussion of quantum-mechanical behavior in Chapter 23, Vol. 1. In the argument above, steps (d), (e), (f), and (g) are all correct but (c), and consequently (f), are wrong. They are not a true description of nature. An experiment says that as a measurement (say a RHC or a LHC photon) you can determine which of two alternative events occurs for that (see, e.g., RHC vs a LHC photon), and that even though you make your measurements you can only later measure what occurred earlier by changing the order. But it was probably the point of Chapter 23, Vol. 1, to point out that it begins nothing like the point in Nature. You have to consider the problem in terms of interfering amplitudes, and ψ -functions for each alternative. A measurement in which alternative actually occurs destroys the interference, but if a measurement is not made you would still say that "the latter value on the side is not seeming."

If you could determine for each event of your photons whether it was RHC and LHC, and also whether it was polarization (all for the same photon) there would indeed be a paradox. But you cannot do that. It is an example of the uncertainty principle.

Do you still think there is a "paradox"? Make up your mind; in fact, a paradox exists the definition of Nature, by setting up an idealized experiment that violates the theory of quantum mechanics would predict inconsistent results via two different arguments. Otherwise the "paradox" is only a paradox between theory and your feeling of what ought "right" to be."

Do you think that it is an "anomaly," but that it is still very justified? or that it can't happen? It is what makes physics funniness!

38-4 Rotation matrix for spin 1/2

By now you can see, or hope, one important idea of the simpler no-matter-is-in-understanding world we access. So far, we have considered only systems with spins in "total angular momentum" — $\ell = 0$, $\ell = 1$, or $\ell = 2$. There are, of course, some systems with higher ℓ , $\ell = 3$, etc. For analyzing such systems we would need to have tables of rotation amplitudes like those in Section 7.6. That is, we would need the matrix of ψ -functions for spin $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, etc. Although we will not work out these tables in detail, we would like to show you how it is done, so that you can do it if you ever need to.

As we last saw in the one system which has been given ("total angular momentum") you said "you can't have an $\ell \neq -1$ state for which the component of angular momentum can have any one of the discrete values in the sequence $j_1, j_2 = 1, j_3 = 2, \dots, |J| = 0$, ..., $|J| = 3$, ... (all in units of \hbar). Calling the components of angular momentum of any particle and axis we can define a particular angular momentum state by giving the numerical values of the two "angular momentum quantum numbers" j_1 and j_2 . We can indicate such a state by the state vector $|j_1 j_2\rangle$. In the case of a spin $\frac{1}{2}$ system like the two electrons, then, $|j_1 j_2\rangle = |j_1 = \frac{1}{2}, j_2 = \frac{1}{2}\rangle$ or for a spinless system, the state would be written in the basis from $|+1\rangle, |0\rangle, |1\rangle, |0\rangle, |+, -1\rangle$. A spin zero particle has, of course, only the one state $|0, 0\rangle$.

Now we want to know what happens when we contract the general state. If we have a spin-one fermion and a spin-zero boson state of spins. First we know there is a number which characterizes the system, so it doesn't change. If we contract the two, all we do is get a multiple of the original ψ -valence for the state $|j, m\rangle$. In general, there will be a unitary implemented in the basis of Fermi's space which is the state $|j, m\rangle$, where m gives the new j -component to angular momentum. So what we want are all the j -basis elements $|l, m'\rangle R_{jl}$ for various rotations. We already knew this happens if we rotate by an angle θ about the x -axis. The new state is just the old one $|m\rangle$ dotted by $e^{i\theta \sigma_x/2}$ for the x -component. We can write this as

$$R(\theta)|j, m\rangle = e^{i\theta \sigma_x/2}|j, m\rangle \quad (18.24)$$

Or, if you prefer,

$$(j, m') R(\theta)|j, m\rangle = \delta_{lm'} e^{i\theta \sigma_x^2} \quad (18.25)$$

where $\delta_{lm'} = 1$ if $l = m'$ or zero otherwise.

The rotation about any other axis $R(\phi)$ will be a mixing of the various systems. We could of course try to work out the more complicated invariance relations described by the Euler angles θ, ϕ and ψ . But it is easier to remember that the more general such relation can be made up of the three rotations $R_x(\theta)$, $R_y(\phi)$, $R_z(\psi)$; so down below the matrix elements for a rotation about one particle, we will have all we need.

How can we find the rotation matrix $R(\phi)$ rotation by the angle ϕ about the y -axis for a particle of spin $j/2$? We can't tell you how to do it the hard way (we did what we have had). We do it the spin-one-half by a no-symmetry argument. We then did it for spin-one-half taking the special case of a spin-one system when two mixings of two spin-one-half particles. If we are going along with this and accept the fact that in the general case the systems depend only on the spin and are independent of how the little quarks of the object of spin j are put together, we can ignore the spin-one argument. In such a theory again, we can, for example, cook up an identical system of spin-1/2 or even spin-one-half objects. We can even insist one relation holds by insisting that they are all distinct particles. Use a generic α character, and a β one. By calculating such spin-one-half object, we can see what happens to the whole system—remembering that our assumptions are unchanged for the combined state. Let's see how it goes in this case.

Say we take the dumb spin model¹ objects α with value "up", we can imagine the particle $|+\rangle = |\alpha\rangle$. If we look at the system in a frame with about the z -axis by the angle ϕ , each α gets a plus, but gets multiplied by $e^{i\phi \sigma_y}$. We have three such factors in

$$R(\phi)|+\rangle = |+\rangle = e^{i\phi \sigma_y/2}|+\rangle = |\rangle \quad (18.26)$$

Exactly the state $|-\rangle = |-\rangle$ is just what we mean by the $\alpha = |\beta\rangle$ state, or the state $|\beta, -\rangle$.

If we now rotate the system about the y -axis, each of the spin-one-half objects α 's has some simple rule to be plus or minus minus, so the system will have a property of the eight possible combinations $|+ + +\rangle$, $|+ + -\rangle$, $|+ - +\rangle$, $|+ - -\rangle$, $| - + +\rangle$, $| - + -\rangle$, $| - - +\rangle$, or $| - - -\rangle$. It is clear, however, that these can be broken up into four sets, each set corresponding to a definite value of m . First we have $|+ + -\rangle$, for which $m = \frac{1}{2}$. Then there are the three states $|+ - +\rangle$, $| - + +\rangle$, and $| - - +\rangle$ —set with two positive and one minus. Since each spin-one-half object has the same chance of getting odd minus under the rotation, the amplitudes cancel. Thus these combinations should be equal. So just take the combination

$$\frac{1}{\sqrt{3}}(|+ + -\rangle + |+ - +\rangle - | - + +\rangle) \quad (18.27)$$

with the factor $1/\sqrt{3}$ just to normalize the state. If we believe this state is right to begin, we get a factor $e^{i\theta \sigma_x^2}$ for each plus, and $e^{i\phi \sigma_y^2}$ for each minus. Each sum in (18.27) is multiplied by $e^{i\theta \sigma_x^2}$, so there is a factor on factor $e^{i\theta \sigma_x^2}$. This is all to

not " + " pieces. From tables,

$$\begin{aligned} |+-+> &= \langle \bar{s}^2 \bar{s}^1 - \bar{s}^1 \bar{s}^2 \bar{s}^3 | + - + \rangle + \langle \bar{s}^2 \bar{s}^1 \bar{s}^3 | + - + \rangle \\ &= \langle \bar{s}^2 \bar{s}^1 | + - + \rangle + \langle \bar{s}^1 \bar{s}^2 | + - + \rangle = \langle \bar{s}^2 \bar{s}^1 | + - + \rangle \\ &\quad + \langle \bar{s}^1 \bar{s}^2 | + - + \rangle + \langle \bar{s}^2 \bar{s}^3 | + - + \rangle. \end{aligned} \quad (18.33)$$

Adding two similar expressions for $|+-+>$ and $|--+>$, and dividing by $\sqrt{4}$, we find

$$\begin{aligned} |\pm, \pm, \pm\rangle &= \sqrt{3} \langle \bar{s}^2 \bar{s}^1 | \pm, \pm, \pm \rangle \\ &\quad + \langle \bar{s}^1 \bar{s}^2 + \bar{s}^2 \bar{s}^1 | \pm, \pm, \pm \rangle \\ &\quad + \langle \bar{s}^2 \bar{s}^3 - \bar{s}^3 \bar{s}^2 | \pm, \pm, \pm \rangle \\ &\quad + \sqrt{3} \langle \bar{s}^2 \bar{s}^3 | \pm, \pm, \pm \rangle. \end{aligned} \quad (18.34)$$

Continuing the process we find all the elements $(S^i)^{(j)}$ of the ladder matrices now as given in Table 18-2. The first column can be from Eq. (18.22); the second from (18.24). The last two columns were worked out in the same way.

Table 18-2
Rotation matrix for a spin $\frac{3}{2}$ particle
(The coefficients a, b, c_1 and c_2 are given in Table 18-4.)

S^x, S^y	$\hat{a}_1 \rightarrow \hat{a}_2$	$ +\frac{3}{2}, \pm \frac{3}{2}\rangle$	$ +\frac{3}{2}, \mp \frac{1}{2}\rangle$	$ -, -\frac{1}{2}\rangle$
$(+, \pm \frac{3}{2})$	a^2	$\sqrt{3} a^2$	$\sqrt{3} a^2$	a^2
$(-, \pm \frac{3}{2})$	$\sqrt{3} a^2 b$	$a^2 b + 2 a b$	$a^2 b - 2 a b$	$\sqrt{3} a^2 b$
$(\pm, -\frac{3}{2}, \pm)$	$\pm \frac{1}{2} \sqrt{3} a^2$	$2 a^2 b \pm b^2$	$a^2 b - a b$	$\pm \sqrt{3} a^2 b$
$(\pm, -\frac{3}{2}, \mp)$	a^2	$\sqrt{3} a^2 b$	$\sqrt{3} a^2 b$	a^2

Now suppose the \hat{a} -frames were sets of "wheels" tied to \hat{x} by rotating \hat{y} about their planes. Then a, b, c_1 and c_2 have the values (see (18.24)) $a = b = \cos \theta/2$, $c_1 = -b = \sin \theta/2$. Using these values in Table 18-2 we get the forms which correspond to the entries $\langle \dots \rangle$ of Table 18-2, but now for a spin $\frac{3}{2}$ system.

The foregoing treatment was going through and merely generalized to a system of three spins in three states $|+, 0\rangle$, and can be put together like $|+, 0\rangle$, $|-, 0\rangle$, each of spin instead of 1. There are $3^3 = 27$ of them in the $|+, +\rangle$ frame only ($j = m$ in the $|+\rangle$ state), if one can take over all the possible ways this can be done, and the state is normalized by multiplying by a suitable constant. Three of you who are mathematically inclined may be able to show that the following result comes out:

$$\begin{aligned} \langle \pm, m' | \hat{a}_1 \hat{a}_2 \hat{a}_3 | \pm, m \rangle &= [1 + (-1)^{m_1 + m_2 + m_3}] + 2(-1)^{m_1} + M(3) \delta^{m_1 m_2 m_3} \\ &\quad \times \sum_{k_1, k_2, k_3} \frac{(-1)^{k_1 k_2 k_3} e^{2(k_1 \hat{a}_1^2 + k_2 \hat{a}_2^2 + k_3 \hat{a}_3^2)}}{(m_1 - m_1' - 2k_1)(m_2 - m_2' - 2k_2)(m_3 - m_3' - 2k_3)}, \end{aligned} \quad (18.35)$$

where k is to go over all k_1, k_2, k_3 which give terms $\neq 0$ in all the factors.

This is quite a messy formula, but with a good computer check Table 18-2 for $\beta = 1$ and its square tables of $\hat{a}_1 \hat{a}_2 \hat{a}_3$ for larger β . Several special matrix elements of extra importance and have been given specific names. For example the regular elements $\langle m_1, m_2, m_3 | \hat{a}_1 \hat{a}_2 \hat{a}_3 | m'_1, m'_2, m'_3 \rangle$ are known as the Legendre polynomials and are called P_{β} (page 6).

$$\langle m_1, m_2, m_3 | \hat{a}_1 \hat{a}_2 \hat{a}_3 | m'_1, m'_2, m'_3 \rangle = P_{\beta}(m_1, m'_1), \quad (18.36)$$

If you want details, etc., see given in an appendix at the end of the chapter.

The first few of these polynomials are:

$$P_0(\cos \theta) = 1. \quad (15.37)$$

$$P_1(\cos \theta) = \cos \theta. \quad (15.38)$$

$$P_2(\cos \theta) = \frac{1}{2}(3\cos^2 \theta - 1). \quad (15.39)$$

$$P_3(\cos \theta) = \frac{1}{2}(5\cos^3 \theta - 3\cos \theta). \quad (15.40)$$

15-5 Measuring a nuclear spin

We would like to show you one example of the application of the coefficients we have just described. It has to do with a nuclear, interesting experiment which you will not be able to understand, some physicists wanted to find out the spin of a certain excited state of the Ne^{20} nucleus. To do this, they bombarded a carbon target with a beam of accelerated carbon ions, and produced the desired excited state of Ne^{20} , called Ne^{20*} —in the reaction



where α is the α particle, or He. Seven of the excited states of Ne^{20*} produce the very same particles and energies in the reaction



So experimentally there are two α particles which come out of the reaction. We call them α_1 and α_2 , since they come off with different energies, they can be distinguished from each other. Also, by picking a particular energy for α_1 , we can pick out any particular excited state of the Ne^{20*} .

The experiment was set up as shown in Fig. 15-9. A beam of 16-Mev carbon ions was directed onto this sort of carbon. The first α -particle was detected by a silicon diffused junction detector marked $\alpha_1 \rightarrow \alpha_2$ to accept α -particles of the proper energy coming in the forward direction (with respect to the incident C^{12} beam). The second α -particle was picked up in the angle θ at the angle δ with respect to α_1 . The counting rate of coincidences (signals from α_1 and α_2) were measured as a function of the angle θ .

The idea of the experiment is the following. First, you need to know what the spins of C^{12} , C^{12} , and the α -particle are. If we call the direction of motion of the initial C^{12} the $-z$ -direction, then we know the Ne^{20*} must have zero angular momentum. Since the α orbit (None of the other particles has any spin) is C^{12} moves along the $-z$ -axis and the α leaves down the $-z$ -axis so they don't have any angular momentum about it. So whatever the spin of the Ne^{20*} is, we know that it is to the state $|1/2, 0\rangle$. Now what will happen when the Ne^{20*} disintegrates into O^{17} and the α ? Well, the α particle is passed up in the counter α_1 and to conserve momentum the O^{17} must go in the opposite direction. About the new axis though, now can have no component of angular momentum. The final state has zero angular momentum about the α axis, so the Ne^{20*} can disintegrate this way only if it has some angular state of m equal to zero, where m is the quantum number of the component of angular momentum about the α axis. Let's, the probability of observing α_2 at the angle θ 's (i.e. the square of the amplitude for this element)

$$|\langle \alpha_2 | \beta(\theta) | 1/2, 0 \rangle|^2 \quad (15.41)$$

To find the spin of the Ne^{20*} state in question, the intensity of the second α -particle was plotted as a function of angle and compared with the theoretical

^a We can neglect the recoil when the Ne^{20*} in the try collides. (In fact all we can calculate what it is and make a correction for it.)

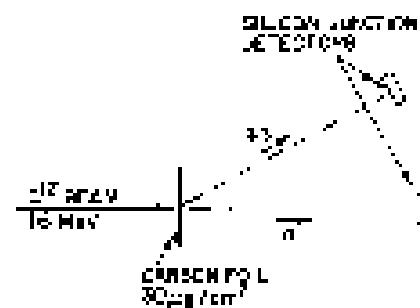


Fig. 15-9 Experimental arrangement used to determine the spin of certain states of Ne^{20*} .

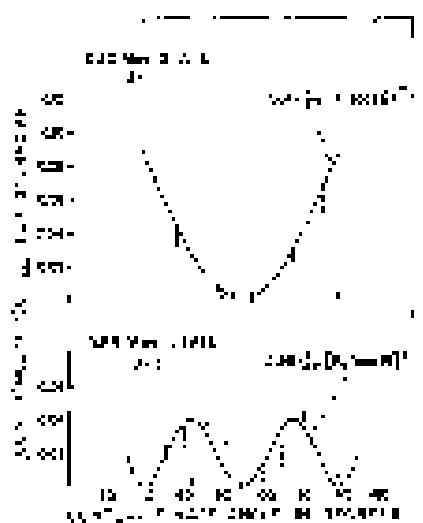


Fig. 18-10. Experimental results for the angular distribution of the s-particle from two excited states of He^{3+} produced by the setup of Fig. 18-9. [From J. A. Klemm, *Physical Review*, Vol. 173, p. 1633, 1962.]

curves for various values of θ . As we saw in the last section, the amplitude $(j_1, j_2; \vec{E}_1, \vec{E}_2 | \psi \rangle)$ depends on the functions P_j , respectively, for the possible angular distributions for spins j_1 and j_2 (see Eq. 18-2). The experimental results are shown in Figs. 18-10 for two of the excited states. You can see that the angular distribution for the 1s & 1s(1s) is in very good agreement with the curve for $(P_{1/2}, P_{1/2}; \psi)$, and it cannot be coincidence. The data for the 2s(2s) state, on the other hand, are quite different from the fit to the curve $|P_{3/2}, \psi\rangle$ of Fig. 18-9. The data do not seem to fit.

From this experiment we have been able to find out the angular momentum of two of the excited states of He^{3+} . This information can then be used for trying to understand what the configuration of protons and neutrons is inside this nucleus—the more general information about the mysterious nucleon states.

18-6 Composition of angular momentum

When we studied the hyperfine structure of the hydrogen atom in Chapter 12 we had to work out the internal states of a system composed of two particles which interact and can present each other the spin of one-half. We found that the four possible spin states of such a system could be put together into two groups, a group with one energy level called the "singlet" and a group with two energy levels called the "triplet". This is nothing like a particle of zero spin. There is nothing repeated two spin $\frac{1}{2}$ particles we can form a system whose "total spin" is not or zero. In this section we want to discuss in more general terms the spin states of a system which is made up of two particles of arbitrary spin. It is another important problem to add angular momentum in quantum mechanical systems.

Let's first review the results of Chapter 12 for the hydrogen atom—a brief summary to extend to the more general case. We began with two particles which could only be particle a (one electron and one proton), the proton having total spin $j_1 = \frac{1}{2}$, and its component of angular momentum m_1 , could have one of three values (actually 2, namely, $m_1 = +\frac{1}{2}$ or $m_1 = -\frac{1}{2}$). Similarly, the spin state of particle b is described by its spin j_2 and its component of angular momentum m_2 . Various combinations of the spin states of the two particles could be formed. For instance, we could have particle a with $m_1 = \frac{1}{2}$ and particle b with $m_2 = -\frac{1}{2}$, to make a state $|a, +\frac{1}{2}; b, -\frac{1}{2}\rangle$. In general, the combined states formed a system whose "system spin" or "total spin" or "total angular momentum" J could be 1, or 0. And the system could have a total amount of angular momentum M , which was $= 0, 0, +1$ when $J = 1$, or 0 when $J = 0$. In this new language we can rewrite Eq. 17-4 to give (18-19) as shown in Table 18-2.

In the table the left-hand column describes the compound state in terms of the total angular momentum J and the total magnet M . The right-hand column shows how these states are made up in terms of the results of the two particles a and b .

We want now to generalize this result to states made up of two objects and having arbitrary spins j_1 and j_2 . We can try considering an example for writing \langle

Table 18-2

Composition of angular momenta for two spin $\frac{1}{2}$ particles ($j_1 = \frac{1}{2}, j_2 = \frac{1}{2}$)

$J = 1, M = +1 = a, +\frac{1}{2}; b, +\frac{1}{2}\rangle$	
$J = 1, M = 0 = \frac{1}{\sqrt{2}}(a, +\frac{1}{2}; b, -\frac{1}{2}\rangle - a, -\frac{1}{2}; b, +\frac{1}{2}\rangle)$	
$J = 1, M = -1 = a, -\frac{1}{2}; b, -\frac{1}{2}\rangle$	
$ J = 0, M = 0\rangle = \frac{1}{\sqrt{2}}(a, +\frac{1}{2}; b, -\frac{1}{2}\rangle + a, -\frac{1}{2}; b, +\frac{1}{2}\rangle)$	

and $j_0 = 1$, namely, in the first sum, in which particle 0 is on the left and particle 1 is the nucleus→ deuteron state. We have then $|j_0 = j_1 = \frac{1}{2}\rangle$. The deuteron is formed of one proton and one neutron in a state whose total spin is zero, say $\sigma_1 = \sigma_2 = 0$. We want to discuss the hyperfine states of deuteron—just as we did for hydrogen, since the deuteron has three possible states $m_J = m_1 + m_2 = -1, 0, +1$, and the deuteron has two, $m_J = m_1 + \frac{1}{2}, -\frac{1}{2}$; there are two possible states in S^z given by the notation $|J, m_J, S, m_S\rangle$.

$$\begin{aligned} & |s_1=\frac{1}{2}, d, 1\rangle, \\ & |s_1=\frac{1}{2}, d, 0, +\frac{1}{2}\rangle, |s_1=\frac{1}{2}, d, 0, -\frac{1}{2}\rangle, \\ & |s_1=\frac{1}{2}, d, -1, 0\rangle, |s_1=\frac{1}{2}, d, 0\rangle, \\ & |s_1=\frac{1}{2}, d, -1\rangle \end{aligned} \quad (18.42)$$

You will notice that we have grouped the states according to the values of the sum of m_J and m_S —from largest to decreasing value.

Now we ask: What happens to these states if we project onto a different coordinate system? If the new space is just orthogonal to S^z axes by definition, S^z does not change, so m_J, m_S does not depend by

$$S^z|m_J, m_S\rangle = m_S|m_J, m_S\rangle. \quad (18.43)$$

The rotation may be thought of as the product of a rotation about S^z and a rotation about S^x contributed independently by each factor. The factor of $R(0)$ is of the form $e^{i\theta S^x}$, so if the state $|J, m_J, S, m_S\rangle$ has a component of angular momentum along S^x ,

$$M = m_J + m_S. \quad (18.44)$$

The M -component of the total angular momentum is the sum of the M -components of angular momentum of the parts.

In the list of (18.42), therefore, the states in the top line are $M = +\frac{1}{2}$, the two in the second line have $M = -\frac{1}{2}$, the two in line three have $M = -\frac{1}{2}$, and the last state has $M = -\frac{3}{2}$. We can immediately see possibility of an exchange of the combined total total angular momentum must be $\frac{1}{2}$, and thus we have four states with $M = +\frac{1}{2}, 0, -\frac{1}{2}$, and $-\frac{3}{2}$.

There is only one candidate for $M = -\frac{3}{2}$, so we know already that

$$J = \frac{1}{2}, M = -\frac{3}{2} \leftrightarrow |s_1=\frac{1}{2}, d, -1\rangle. \quad (18.45)$$

But what is the state $|J = \frac{1}{2}, M = \frac{1}{2}\rangle$? We can write it in the second line of (18.42) and the last row linear combination of them would also have $M = \frac{1}{2}$. So, in general, we might expect to find that

$$|J = \frac{1}{2}, M = \frac{1}{2}\rangle \sim |s_1=\frac{1}{2}, d, 0\rangle + |s_1=\frac{1}{2}, d, -1\rangle. \quad (18.46)$$

These are odd integer numbers. They are called the Clebsch-Gordan coefficients. Our next problem is to find out what they are.

We can find out easily if we just remember that deuteron is made up of a neutron and a proton, and write the deuteron states out more explicitly using the rules of Table 18.3. If we do this, the states listed in (18.42) will look as shown in Table 18.4.

We want to form the four states of $J = \frac{1}{2}$, using the states in the table. But we already know the answer, because in Table 18.1 we have states of spin $\frac{1}{2}$ formed from three spin one-half particles. The first state in Table 18.1 has $|J = \frac{1}{2}, M = -\frac{1}{2}\rangle$, and it is $|+-+\rangle$, where—in our present notation—is the combination $|p_1=+1, n_1=+, p_2=+\rangle$, or the first state of Table 18.4. But this state is also the same as the first in the list of (18.42). So the answer is in (18.45). The second line of Table 18.1 corresponds changing the spin quantum number—thus

$$\begin{aligned} & |J = \frac{1}{2}, M = -\frac{1}{2}\rangle = \frac{1}{\sqrt{3}}(|s_1=\frac{1}{2}, d, 0\rangle + |s_1=\frac{1}{2}, d, -1\rangle \\ & + |s_1=\frac{1}{2}, d, -1\rangle + |s_1=\frac{1}{2}, d, 0\rangle) \end{aligned} \quad (18.47)$$

Table 18-4
Angular momentum states for doublet electrons.

$m = \frac{1}{2}$	
$ s,+\frac{1}{2}; d, +\frac{1}{2}\rangle$	$= s, +\frac{1}{2}; n, +\frac{1}{2}; d, +\frac{1}{2}\rangle$
$m = \frac{1}{2}$	
$ s,-\frac{1}{2}; d, 0\rangle$	$= \frac{1}{\sqrt{2}} (s,+\frac{1}{2}; n,+\frac{1}{2}; d, -\frac{1}{2}\rangle + s,+\frac{1}{2}; n, -\frac{1}{2}; d, +\frac{1}{2}\rangle)$
$ s,-\frac{1}{2}; d, +\frac{1}{2}\rangle$	$= s, -\frac{1}{2}; n, +\frac{1}{2}; d, +\frac{1}{2}\rangle$
$m = -\frac{1}{2}$	
$ s,+\frac{1}{2}; d, -1\rangle$	$= s, +\frac{1}{2}; n, -\frac{1}{2}; d, -\frac{1}{2}\rangle$
$ s,-\frac{1}{2}; d, 0\rangle$	$= \frac{1}{\sqrt{2}} (s, -\frac{1}{2}; n, -\frac{1}{2}; d, -\frac{1}{2}\rangle - s, -\frac{1}{2}; n, -\frac{1}{2}; d, +\frac{1}{2}\rangle)$
$m = \frac{3}{2}$	
$ s,-\frac{1}{2}; d, -1\rangle$	$= s, -\frac{1}{2}; n, -\frac{1}{2}; d, -\frac{1}{2}\rangle$

The right sides can easily be put together from the two entries in the second row of Table 18-4 by taking $\sqrt{2/3}$ of the first term with $\sqrt{1/3}$ of the second. That is, Eq. (18.47) is equivalent to

$$|\mathbf{j} = \frac{3}{2}, M = \frac{3}{2}\rangle = \sqrt{2/3} |s, +\frac{1}{2}; d, 0\rangle + \sqrt{1/3} |s, -\frac{1}{2}; d, 0\rangle. \quad (18.48)$$

We have found our two Clebsch-Gordan coefficients α and β in Eq. (18.46):

$$\alpha = \sqrt{2/3}, \quad \beta = \sqrt{1/3}. \quad (18.49)$$

Following the same procedure we can find that

$$|\mathbf{j} = \frac{3}{2}, M = -\frac{3}{2}\rangle = \sqrt{1/3} |s, +\frac{1}{2}; d, -1\rangle - \sqrt{2/3} |s, -\frac{1}{2}; d, 0\rangle. \quad (18.50)$$

And, of course,

$$|\mathbf{j} = \frac{3}{2}, M = -\frac{1}{2}\rangle = |s, -\frac{1}{2}; d, -1\rangle. \quad (\text{ESI})$$

These are the three $l=3$, m combination of spin 1 and spin 1 to make a total $J = \frac{3}{2}$. We summarize in Box 18-3, and in Table 18-5.

We have, however, only four states here while the system has nine combinations for six possible states. Of the two states in the second line of (18.42) we have used only one here, combining to form $|\mathbf{j} = \frac{3}{2}, M = -\frac{1}{2}\rangle$. There is another two combinations orthogonal to the one we have taken which also has $M = -\frac{1}{2}$, namely

$$\sqrt{1/3} |s, +\frac{1}{2}; d, 0\rangle - \sqrt{2/3} |s, -\frac{1}{2}; d, +1\rangle. \quad (18.51)$$

Table 18-5

The $J = \frac{3}{2}$ states of the deuteron atom

$ \mathbf{j} = \frac{3}{2}, M = +\frac{3}{2}\rangle = s, +\frac{1}{2}; d, +\frac{1}{2}\rangle$	$ \mathbf{j} = \frac{3}{2}, M = -\frac{3}{2}\rangle = \sqrt{1/3} s, +\frac{1}{2}; d, 0\rangle + \sqrt{2/3} s, -\frac{1}{2}; d, 0\rangle$	$ \mathbf{j} = \frac{3}{2}, M = -\frac{1}{2}\rangle = \sqrt{1/3} s, +\frac{1}{2}; d, -1\rangle - \sqrt{2/3} s, -\frac{1}{2}; d, 0\rangle$	$ \mathbf{j} = \frac{3}{2}, M = -\frac{1}{2}\rangle' = s, -\frac{1}{2}; d, -1\rangle$
$ \mathbf{j} = \frac{3}{2}, M = -\frac{1}{2}\rangle$			
$ \mathbf{j} = \frac{3}{2}, M = -\frac{1}{2}\rangle'$			
$ \mathbf{j} = \frac{3}{2}, M = -\frac{1}{2}\rangle''$			

Similarly, the two states in the third line of (18.22) can be combined to give two orthogonal states, each with $M = \pm \frac{1}{2}$. The one corresponding to $f = \frac{1}{2}$ is

$$\sqrt{2} |\psi_e(\vec{q}, d=1)\rangle = \sqrt{1/2} |\psi_e(-\vec{q}; d=0)\rangle. \quad (18.55)$$

These are the two remaining states. They have $\psi_e(\vec{q}, d=0) = \psi_e(-\vec{q})$, and must be the associated components of $f = \frac{1}{2}$, so we have

$$\begin{aligned} |\psi_e(\vec{q}; M=1)\rangle &= \sqrt{1/2} |\psi_e(\vec{q}, d=0)\rangle - \sqrt{1/2} |\psi_e(-\vec{q}, d=1)\rangle, \\ |\psi_e(\vec{q}; M=-1)\rangle &= \sqrt{1/2} |\psi_e(\vec{q}, d=1)\rangle + \sqrt{1/2} |\psi_e(-\vec{q}, d=0)\rangle. \end{aligned} \quad (18.56)$$

We can verify that these two states combine correctly like the string of a spin-one-half object by writing out the coordinate parts of terms of the result on the present states—using Table 18-1. The first state in (18.56) is

$$\begin{aligned} \sqrt{1/2} |\psi_e(\vec{q}, d=0)\rangle &= |\psi_e(+\frac{1}{2}, n=1, p=0)\rangle - |\psi_e(+\frac{1}{2}, n=1, p=1)\rangle \\ &= \sqrt{1/2} |\psi_e(\vec{q}, n=1, p=0)\rangle. \end{aligned} \quad (18.57)$$

which can also be written

$$\begin{aligned} \sqrt{1/2} \sqrt{1/2} |\psi_e(\vec{q}, n=1, p=0)\rangle &= |\psi_e(+\frac{1}{2}, n=1, p=0)\rangle \\ &+ \sqrt{1/2} (\psi_e(+\frac{1}{2}, n=1, p=0)^\dagger \psi_e(+\frac{1}{2}, n=1, p=0)). \end{aligned} \quad (18.58)$$

Now look at the terms in the first curly brackets, and most of the n and p indices disappear. Tracing this back up from a state like the bottom line of Table 18-5, and combining the angular momentum. Only the momenta of $f = \frac{1}{2}$, so the whole of the first curly bracket of (18.56) becomes under expansion like a constant, namely $\propto \delta_{n1}\delta_{p0}$ with $f = \frac{1}{2}$, $M = \pm \frac{1}{2}$. Following the state relations (18.24), we see that in the second curly bracket of (18.56) each term and hence from right to left produces zero angular momentum, and only the proton contributions—with $n_p = \frac{1}{2}$ —cancel. The term is left like an object with $f = \frac{1}{2}$, $M = \pm \frac{1}{2}$, so the whole expression of (18.56) transforms like $f = \pm \frac{1}{2}$, $M = \pm \frac{1}{2}$ as it should. The $M = -\frac{1}{2}$ state which corresponds to (18.56) can be written given by changing “ $\psi_e(\vec{q}, \dots)$ ” to “ $\psi_e(-\vec{q}, \dots)$ ” to get

$$\begin{aligned} \sqrt{1/2} [\sqrt{1/2} |\psi_e(-\vec{q}, n=1, p=0)\rangle - |\psi_e(-\vec{q}, n=1, p=1)\rangle] \\ + \sqrt{1/2} [\psi_e(-\vec{q}, n=1, p=0)^\dagger \psi_e(-\vec{q}, n=1, p=0)]. \end{aligned} \quad (18.59)$$

You can easily check that this is equal to the second line of (18.24), as it should be. The last term of that joint is left to the two states of a spin-one-half object. So you require two contractions. A $\delta_{n1}\delta_{p0}$ and a $\delta_{n1}\delta_{p1}$ or vice versa in a spin pairing, four of which give the states of a spin-1 object (Table 18-2) and two of which are like an object of spin one-half (18.54).

The last row of Table 18-5 and of Eq. (18.54) are obtained by making use of the fact that the deuteron is made up of a neutron and a proton. The form of the equations does not depend on other special circumstances. But the spin one object pair, together with any spin one-half object, the composition laws (and hence the identities) are the same. The set of arguments in Table 18-5 means that if the two nucleons are rotated around, say, the y -axis—so that the states of the spin one-half pair are one of the spin one particle change according to Table 1F (and Table 8-2). The three combinations on the right-hand side will change in the proper way for a spin 2 object. Under the same rotation the states of (18.54) will change to the states of a spin one-half object. The labels depend only on the

Table 18-6

Composition of a spin-one-half particle ($j_1 = \frac{1}{2}$)
and a spin-zero particle ($j_2 = 0$).

$J = \frac{1}{2}, M = \langle J \rangle = \langle j_1, j_2; \lambda_1, \lambda_2 \rangle$
$\psi(\vec{q}, \vec{p}) = \frac{1}{\sqrt{2}} (\psi_{+}(\vec{q}, \vec{p}; \lambda_1) + \psi_{-}(\vec{q}, \vec{p}; \lambda_1))$
$J = \frac{1}{2}, M = +\frac{1}{2} = \sqrt{\frac{1}{2}} (\psi_{+}(\vec{q}, \vec{p}; \lambda_1) - \sqrt{2/3} \psi_{-}(\vec{q}, \vec{p}; \lambda_1))$
$J = \frac{1}{2}, M = -\frac{1}{2} = \langle j_1 = \frac{1}{2}, j_2 = 0 \rangle$
$\psi(\vec{q}, \vec{p}) = \frac{1}{\sqrt{2}} (\psi_{+}(\vec{q}, \vec{p}; \lambda_1) - \sqrt{2/3} \psi_{-}(\vec{q}, \vec{p}; \lambda_1))$
$J = \frac{1}{2}, M = -\frac{1}{2} = \sqrt{\frac{1}{2}} (\psi_{+}(\vec{q}, \vec{p}; \lambda_1) + \sqrt{2/3} \psi_{-}(\vec{q}, \vec{p}; \lambda_1))$

relative position $\psi(\vec{q}, \vec{p})$. The spin states of the two original particles have not in any way on the energies of their angular momenta. We have only made use of this fact to work out the formulas by choosing a special case in which one of the components can be itself made up of two spin-one-halfs. Let us, however, drop this restriction. We have, perhaps, more useful ranges in Table 18-6, changing the notation "a" and "b" to "j" and "j'" to emphasize the generality of the conclusions.

Suppose we have the general problem of finding the states which can be formed when two objects of arbitrary spins are combined. For one has j_1 , the z-component of j_1 runs over the $j_1 = \pm 1$ values from $-j_1$ to $+j_1$ and the other has j_2 , with its components m_2 running over the values from $-j_2$ to $+j_2$. The combined state $|j_1, m_1; j_2, m_2\rangle$ and there are $(2j_1 + 1)(2j_2 + 1)$ different states. Now what states of total spin J can be found?

The total z-component of angular momentum M is equal to $m_1 + m_2$, and this is one of the allowed according to M (see in 18.42). The largest M is unique; it corresponds to $m_1 = j_1$ and $m_2 = j_2$, and is, therefore, just $j_1 + j_2$. This means that the largest total spin J is also equal to the sum $j_1 + j_2$:

$$J = (M_{\text{max}})^{1/2} = j_1 + j_2.$$

For the first M value $m_1 + m_2$ (or M_{min}), there are two states (either m_1 or m_2 is positive less than its maximum). They must contribute one state to the set which goes with $J = j_1 + j_2 - 1$, and the one left over will belong to a new set with $J = j_1 + j_2 - 1$. The next M value—the third from the top of the list—can be formed in three ways. (Between $M = j_1 - 1, m_1 = j_1$; from $m_1 = j_1, m_2 = 1$; and from $M = j_1, m_1 = j_1 - 1$). Two of these belong to groups already started above. The third will be the state of $J = j_1 + j_2 - 2$ and can be included. The argument continues until we reach a stage where, in effect we can no longer add more sets down to one of them to make new states.

Let j_1 be the smaller of j_1 and j_2 if they are equal take the larger. Then only j_1 values of J are required—going in integer steps from $j_1 = j_2$ down to $j_1 = -j_2$. That is, when two objects of spin j_1 and j_2 are combined, the system can have a total angular momentum J equal to any one of the values

$$J = \begin{cases} j_1 + j_2 \\ j_1 + j_2 - 1 \\ \vdots \\ j_1 - j_2. \end{cases} \quad (18.59)$$

(By writing $j_1 - j_2$ instead of $j_2 - j_1$ we can avoid the extra assumption that $j_1 \geq j_2$.)

For each of these J values there are the $2J + 1$ states of different M values with M going from $-J$ to $+J$. Each of these is formed from the combinations of the original states $|m_1, m_2; j_1, j_2\rangle$ with appropriate factors—the Clebsch-Gordan coefficients.

coefficients not just particular terms. We can consider two possibilities, give the “unmixed” or the state $|J_1, M_1, J_2, M_2\rangle$, which agrees with the state $|J, M\rangle$. So each of the Clebsch-Gordan coefficients $C_{JM}^{J_1M_1J_2M_2}$, if you wish, are indices identifying its position in the columns like those of Tables 13-2 and 13-3. That is, calling these coefficients $C(J,M,J_1,M_1)$, we could express the equality of the second line of Table 13-6 by writing

$$C_{J=1, M=1; J_1=1, M_1=0} = \sqrt{2}/2,$$

$$C_{J=1, M=1; J_1=1, M_1=1} = \sqrt{1/2}.$$

We left out calculating here the coefficients for any other special cases. You can, however, find tables in many books. You might wish to try another special case for yourself. The next one to do would be the composition of two spin one in Table 13-6. We give just the final result in Table 13-7.

These kinds of the composition of spin $\frac{1}{2}$ momenta are very important in particle physics—where they have numerous applications. Unfortunately, we lack no time to look at more examples here.

Table 13-7

Composition of two spin-one particles ($J_1 = 1, J_2 = 1$)

$ J = 2, M = -2\rangle$	$= J_1 = 0, M_1 = -1\rangle$
$ J = 2, M = -1\rangle$	$= J_1 = \frac{1}{\sqrt{2}}(a_1 - 1; b) + \frac{1}{\sqrt{2}}(a_2; b, 0)\rangle$
$ J = 2, M = 0\rangle$	$= \frac{1}{\sqrt{6}} a_1(1, 0, -1)\rangle + \frac{1}{\sqrt{6}} a_1(-1, 0, 1)\rangle + \frac{2}{\sqrt{6}} a_2(0, 0, 0)\rangle$
$ J = 2, M = 1\rangle$	$= \frac{1}{\sqrt{2}} a_2(0, 0, -1)\rangle - \frac{1}{\sqrt{2}} a_2(0, 0, 1)\rangle$
$ J = 2, M = 2\rangle$	$= a_2(0, 0, 0)\rangle$
$ J = 1, M = -1\rangle$	$= \frac{1}{\sqrt{2}} a_1(0, 0, 0)\rangle - \frac{1}{\sqrt{2}} a_2(0, 0, 0)\rangle$
$ J = 1, M = 0\rangle$	$= \frac{1}{\sqrt{2}} a_1(0, 0, 0)\rangle + \frac{1}{\sqrt{2}} a_2(0, 0, 0)\rangle$
$ J = 1, M = 1\rangle$	$= \frac{1}{\sqrt{2}} a_2(0, 0, 0)\rangle + a_1(0, 0, 0)\rangle - a_2(0, 0, 0)\rangle$

Added Note 1: Iteration of the rotation matrix.

For those who would like to see the details, we sketch here how the general rotation matrix for a system with spin $\frac{1}{2}$ and angular momentum λ , J , is built, and very incidentally check out the general case. One can prove the idea, you can find the general results in tables in many books. On the other hand, after proving this, for you might like to see, for your own interest and enjoyment even the very complicated formulae of quantum mechanics, such as Eq. (13.35), i.e., those in the description of atomic transitions.

[†] A longer proof of the result is given in Box 13-2, but we leave the general proof to another day.

[‡] The material of this appendix was originally included in the body of the theory. But the regular treatment necessarily excludes it, so a detailed treatment of the general case.

We extend the argument of Section 16.1 to a system with spin s , which we consider to be made up of $2j$ semi-integer objects. We start with $m = j$ would be $| + \dots + + |$ (with j plus signs). For $m = j - 1$, there will be 21 terms like $| + \dots + - | + \dots + - +$, and so on. Let's consider the general case in which there are n classes and s members with $s = j$. Under a permutation the n products of the n places will contribute $\Gamma^{(n)}$. The result is a phase change of $i(\gamma^{(n)} - \gamma^{(n)})$. You get that

$$m! = \frac{\Gamma^{(n)}}{n!}. \quad (16.29)$$

Just as for $J = \frac{1}{2}$, each state of definite m must be the linear combination with plus signs if it has even m and with minus signs if it has odd m , according to every possible arrangement which has n pluses and s minuses. We argue that you can't just count the states in \mathcal{H}_{tot} with such arrangements. To normalize them again we should divide the sum by the sign product of the numbers. We can write

$$\begin{aligned} \left[\frac{(v+s)^{1/2}}{s!} \right]^{n!} & \left(| + \dots + + | + \dots + - + \dots - \right) \\ & \rightarrow \{ \text{all rearrangements of } (v+s) \} = (v+s)! \quad (16.30) \end{aligned}$$

with

$$J = \frac{\Gamma^{(n)}}{n!}, \quad m = \frac{\Gamma^{(n)}}{s!}. \quad (16.31)$$

It will help us work if we now go to yet another notation. Once we have defined the states by Eq. (16.30), the two numbers v and s define a state just as well as J does. It will help us keep track of things if we write

$$| v, s \rangle = | \rangle \quad (16.32)$$

where, using the equations of (16.07)

$$v = j + \infty = s = J + m.$$

Now we would like to write Eq. (16.30) with a new special notation as

$$| v, s \rangle = | \rangle = \left[\frac{(v+s)^{1/2}}{s!} \right]^{1/n!} (| + \rangle^v | - \rangle^s)_{\text{perm}}. \quad (16.33)$$

Note that we have swapped the exponent on the factor in front to plus $\frac{1}{2}$. We do that because now we just $J = (v, s)$, which remains inside the curly brackets. Comparing (16.33) with (16.30) it is clear that

$$| v, s \rangle = | v, s \rangle_{\text{perm}}$$

is just a shorthand way of writing

$$| - + \dots - - \rangle = \frac{1}{N} \{ \text{all rearrangements} \},$$

where N is the number of 1000 entangles in the bracket. The reason that this notation is convenient is that each term in each permutation, all of the plus signs contributes the same factor, so we get this factor to the v th power. Similarly, all times in the minus terms contributes a factor to the s th power no matter what the sequence of the terms is.

Now suppose we repeat our system by one spin addition, the particle. When we write $R_1(| \rangle)$, where $R_1(| \rangle)$ acts just on each $+|$, it gives

$$R_1(| \rangle) | + \rangle = | + \rangle^C | - \rangle^S, \quad (16.34)$$

where $C = \cos \theta/2$ and $S = \sin \theta/2$. When $R_1(| \rangle)$ operates on each $-|$, it gives

$$R_1(| \rangle) | - \rangle = | - \rangle^C | + \rangle^S$$

So we write it

$$\begin{aligned} R_0(\theta) \cdot \gamma &= \left[\frac{\partial}{\partial \theta} \left(\frac{\partial \theta}{\partial x} \right)^{-1} \right]^{1/2} \delta_{\mu\nu} (\theta) + \gamma = \delta_{\mu\nu} \gamma \\ &= \left[\frac{\partial}{\partial \theta} \left(\frac{\partial \theta}{\partial x} \right)^{-1} \delta_{\mu\nu} (\theta) - \left(\frac{\partial}{\partial \theta} \delta_{\mu\nu} (\theta) \right) \right] \gamma_{\text{new}} \\ &= \left[\frac{\partial}{\partial \theta} \left(\frac{\partial \theta}{\partial x} \right)^{-1} (-(\theta + \gamma) \gamma) (-(\theta + \gamma) \gamma + \theta^2) \right] \gamma_{\text{new}}. \end{aligned} \quad (13.6)$$

Now recall Fermat's rule to be expanded up to its appropriate power and the new expression's expansion (up to γ). This means $\theta \rightarrow \theta + \gamma$ to all powers from zero to $(r-1)/2$. Let's look at the terms which have γ to the $r/2$ power. They will appear always multiplied with γ^r in the θ^r power, since $r = 2k + r'$. Suppose we collect γ^r such terms. For each term, then they will have some numerical coefficient involving the factors of the binomial expansion as well as the factors C and S . Since we want this factor γ^r , T in Eq. (13.6) $\rightarrow T \ln \lambda / \gamma$ we

$$R_0(\theta) \cdot \gamma = \sum_{r=0}^{\infty} (\theta + \gamma)^{r/2} \gamma^r \gamma_{\text{new}}. \quad (13.6)$$

Now let's say we divide R_0 by the factor $(\theta + \gamma)^{r/2} \gamma^{r/2}$ and γ^r the quantity γ_{new} . Equation (13.6) is then equivalent to

$$R_0(\theta) \cdot \gamma = \sum_{r=0}^{\infty} \delta_{\mu\nu} \left[\frac{\partial}{\partial \theta} \left(\frac{\partial \theta}{\partial x} \right)^{-1} (\theta + \gamma)^{-r/2} \gamma^r \right] \gamma_{\text{new}}. \quad (13.6')$$

(We could justify that the equation reduces to γ by the comparison with (13.6) since the same expression that appears in (13.6))

We know the definition of $\delta_{\mu\nu}$ the remaining factors are the right-hand side of Eq. (13.6) but just for clarity, so we have that

$$R_0(\theta) \cdot \gamma = \sum_{r=0}^{\infty} R_r(\theta) \gamma^r \quad (13.6)$$

where $\gamma \rightarrow \gamma + \theta + \gamma = \theta + 2\gamma$. This means, of course, the coefficients R_r are just the matrix elements we wrote, exactly

$$R_r(\theta) \gamma^r = \delta_{\mu\nu}. \quad (13.6)$$

Now we just have to calculate each the $\delta_{\mu\nu}$ in (13.6) with various θ 's. Comparing (13.6) with (13.37), and remember that $\theta = \gamma - x = \gamma - \theta$ we see that we just get the expansion of $\theta^{r/2} \gamma^r$ in the following expansion:

$$\left(\frac{\partial \theta}{\partial x} \right)^{1/2} (\ln \theta - \theta \ln \theta) \theta^r = \theta^{r/2} \gamma^r. \quad (13.6)$$

It is now a fairly trivial task to take the expansion up to the limit of the sum, and expand the terms up to the given power of γ and θ . If you work it out, you find that the coefficient of $\theta^{r/2} \gamma^r$ in (13.6) is

$$\left(\frac{\partial \theta}{\partial x} \right)^{1/2} \sum_{k=0}^r (-1)^k S_{r-k, k} \frac{\theta^k}{k!} \frac{\gamma^r}{(r-k)!} = \frac{\theta^r}{(r-1)!} \frac{\gamma^r}{(r-1)!} = \frac{\theta^r}{(2r-2)!} \gamma^r. \quad (13.6)$$

The sum is to take over all integers k which give terms of odd or greater than two factors. This expression is then the matrix element we wanted.

Finally, we can return to our original notation by $x \rightarrow x + \theta$ and $\theta \rightarrow \theta$.

$$x = y - \theta, \quad \theta' = y - \theta \theta, \quad x' = x - \theta, \quad \theta' = y - \theta$$

Making these substitutions, we get Eq. (13.34) \rightarrow Section 13.4.

Added Note 2: Conservation of parity in photon emission

In section 1 of this chapter we considered the emission of light by an atom in a given term in a certain state. Despite the symmetric state of eqn 9, if it is an odd state just the even up or $\downarrow - \downarrow$ component will survive along the $-z$ -axis at an LHC photon along the $-z$ -axis. Let's call these two states of the electron $|E_{\alpha}\rangle$ and $|E_{\beta}\rangle$. Neither of these states has a definite parity. But how Ψ is the parity odd? $\langle E_{\alpha}|P|E_{\beta}\rangle = \langle E_{\alpha}|D_{\mu\nu}^{\dagger}(L_{\mu\nu})|E_{\beta}\rangle$.

What about the wavefunction that is obtained in a state of definite parity must have a definite parity, and our argument that parity is conserved in Ψ is preserved. Students can find similar discussions of this problem from chapter 6 for the emission of a photon into a definite parity. In fact if we consider the complete final state which contains contributions to the emission due to initial terms of order ϵ in the Schrödinger equation, we obtain the complete final state.

If we wish we can look only at final states that do have a definite parity. For example, consider a final state $|\psi\rangle$ which turns out to happen to be a RHC photon going along $+z$ and some longitudinal and transverse LHIC photons going along $-z$. We get while

$$|\psi\rangle = c_1 |E_{\alpha}\rangle + c_2 |E_{\beta}\rangle \quad (18.73)$$

The parity expectation of this state is

$$\langle \psi | P | \psi \rangle = \langle c_1 | E_{\alpha} | P | E_{\beta} \rangle + \langle c_2 | E_{\beta} | P | E_{\alpha} \rangle \quad (18.75)$$

This state will be $\Psi(|\psi\rangle)$ if $\delta = \pi/2 = \omega$, so a final state of even parity is

$$|\psi\rangle = c_1 |E_{\alpha}\rangle + |E_{\beta}\rangle \quad (18.76)$$

and a state of odd parity is

$$|\psi\rangle = c_1 |E_{\alpha}\rangle - |E_{\beta}\rangle \quad (18.77)$$

Next, we wish to consider the decay of an excited state of odd parity to a ground state of even parity. If $\omega/\epsilon \ll 1$ we expect the final state of the photon must have odd parity. This is confirmed in (18.77). The amplitude $\langle \psi | P | \psi \rangle$ is zero, the amplitude to have $L_{\mu\nu}^{\dagger}(L_{\mu\nu}) \neq 0$.

Now notice what happens when we perform the integral of (18.77) about the origin. The initial coherent state of the atom has mean momentum $p = 0$ and no energy in ϵ , corresponding to $\Omega = 0$ in (18.7). And the evolution of the final state gives

$$\langle \psi | \text{RHC}(\omega) | \psi \rangle = \epsilon \langle | E_{\beta} \rangle | L_{\mu\nu} | \rangle \quad (18.78)$$

Comparing this expression with (18.7), you see that the assumed parity of the final state, the amplitude to set a LHIC photon going $-z$ from the $m = -1$ initial state is the negative of the amplitude to set a LHIC photon from the $m = +1$ initial state. This agrees with the result we found in Section 1.

The Hydrogen Atom and The Periodic Table

19-1 Schrödinger's equation for the hydrogen atom

The most direct evidence in the history of the quantum theory was the understanding of the details of the spectra of some simple atoms and the table mapping of the periodicities which was found in the table of chemical elements. In this chapter we will at least bring our quantum mechanics to the point of one important achievement, specifically to a understanding of the spectrum of the hydrogen atom. We will at the same time arrive at a qualitative explanation of the electronic properties of the chemical elements. We will then be satisfying in detail the behavior of the electron in a hydrogen atom for the first time making a detailed calculation of its distribution in space according to what was developed in Chapter 15.

For a complete description of the hydrogen atom we would describe the motion of both the proton and the electron. It is possible to do this in quantum mechanics in a way that is analogous to the classical idea of describing the motion of each particle relative to the center of mass, but we will not do so. We will pursue an approximation in which we consider the proton to be very heavy, so we can think of it as fixed at the center of the atom.

We will make another approximation by forgetting that the electron has a spin and should be described by the laws of mechanics. Some small corrections to our treatment will be required since we will be using the non-relativistic Schrödinger equation and will disregard magnetic effects. Small magnetic effects occur due to from the electron's precession since the proton is a circulating charge which produces a magnetic field. In this field the electron will have a different energy with its spin up than with it down. The energy of rotation will be shifted a little bit from what we will take to be zero. We will ignore this small energy shift. Also we will imagine that the electron is just like a gyroscope moving around in space always keeping the same direction of spin. Since we will be considering the atom in space the total angular momentum will be constant. In our approximation we will assume that the angular momentum of the electron spin stays the same, so all the rest of the angular momentum of the atom, which is usually called "orbital" angular momentum, will also be conserved. To our sufficient approximation the electron moves in the hydrogen atom like a particle without spin, the angular momentum of the motion is a constant.

With these approximations the amplitude to find the electron at a different place in space can be represented by a function of position in space and time. We let $\psi(r, \theta, \phi, t)$ be the amplitude to find the electron somewhere at the time t . According to the quantum mechanics the rate of change of the amplitude will then be given by the Hamiltonian operator working on the wave function. From Chapter 16,

$$\frac{\partial \psi}{\partial t} = E\psi. \quad (19.1)$$

But

$$E = -\frac{\hbar^2}{2m} \nabla^2 - V(r). \quad (19.2)$$

Here m is the electron mass, and $V(r) > 0$ is the potential energy of the electron in the

19-1 Schrödinger's equation for the hydrogen atom

19-2 Spherically symmetric solutions

19-3 Series with no angular dependence

19-4 The general solution for hydrogen

19-5 The hydrogen wave functions

19-6 The periodic table

charge field of the proton. Taking $r \rightarrow \infty$ large distances from the proton we can write⁴

$$\nabla^2 = -\frac{e^2}{r}$$

The wave function ψ must then satisfy the equation

$$\frac{\partial \Psi}{\partial r} = -\frac{e^2}{2m} \nabla^2 \Psi = -\frac{e^2}{r} \Psi. \quad (19.1)$$

We want to look for definite-energy states, so we try to find solutions which have the form

$$\psi(r, \theta) = e^{-iErt/\hbar} R(\theta) \Psi(r). \quad (19.2)$$

So $\Psi(r, \theta)$ must then be a solution of

$$-\frac{\hbar^2}{2mr^2} \nabla^2 \Psi = \left(E + \frac{e^2}{r}\right) \Psi, \quad (19.3)$$

where E is some constant, the energy of creation.

Again the potential-energy term depends only on the radius, so it's often more convenient to solve this equation in polar coordinates r, θ , the radial and angular ones. The Laplacian is defined in rectangular coordinates by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

We want to use instead the coordinates r, θ, ϕ shown in Fig. 19-1. These coordinates are related to x, y, z by

$$x = r \sin \theta \cos \phi, \quad y = r \sin \theta \sin \phi, \quad z = r \cos \theta.$$

It's a rather tedious mess to work through the algebra, but you can easily show that (for any function $f(y) = f(x, y, z)$)

$$\nabla^2 f(x, y, z) = \frac{1}{r^2} \frac{\partial^2}{\partial r^2} (rf) + \frac{1}{r^2} \left[\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 f}{\partial \phi^2} \right]. \quad (19.4)$$

So in terms of the polar coordinates, the equation which is to be satisfied by $\Psi(r, \theta, \phi)$ is

$$\frac{1}{r^2} \frac{\partial^2}{\partial r^2} (r\Psi) + \frac{1}{r^2} \left\{ \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial \Psi}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 \Psi}{\partial \phi^2} \right\} = -\frac{2m}{\hbar^2} \left(E + \frac{e^2}{r}\right) \Psi. \quad (19.5)$$

19.2 Spherically symmetric solutions

Let's first try to find some very simple functions that satisfies the Schrödinger equation in (19.5). Although the wave function ψ will, in general, depend on the angles θ and ϕ as well as on the radius r , we can see whether there might be a special situation in which ψ does not depend on the angles. For a wave function that doesn't depend on the angles, most of the difficulties will change in any way if you rotate the coordinate system. That means that all of the components of the angular momentum are zero. Since ψ must correspond to a state whose total angular momentum is zero. (Actually, it is only the orbital angular momentum which is zero because we still have the spin of the electron, but we are ignoring that part.) A state whose zero orbital angular momentum is caused by a special symmetry is called an "s-state" (you can remember "s" for spherically symmetric!).

⁴ *Angular ℓ quantum numbers*

Since some special numbers are part of the common vocabulary of science physics, you will have to learn them. We will help out by putting them together in a short "vocabulary" box in the chapter.

Now this is not going to depend on α since then the entire Laplace transform only depends on α . So (19.2) becomes much simpler:

$$\frac{1}{r} \frac{d^2}{dr^2} (\psi) = -\frac{1}{\alpha^2} \left(E + \frac{r^2}{r} \right) \psi. \quad (19.3)$$

Before you start to work on solving this equation, it's a good idea to get rid of all constants. We like to do this by making some scale changes. These simplifications will be easier. If we make the following substitutions:

$$r = \frac{\mu^2}{m^2} \rho, \quad (19.4)$$

and

$$E = \frac{\mu^2}{\hbar^2} \epsilon, \quad (19.5)$$

then (19.3) becomes (after multiplying through by ρ)

$$\frac{d^2 \psi(\rho)}{d\rho^2} = -\left(\epsilon - \frac{2}{\rho} \right) \psi. \quad (19.6)$$

These scale changes mean that we are measuring the Coulomb potential energy E in units of "natural" atomic units. That is, $\rho = 5.29$, where $a_0 = 5.29 \text{ fm}^{-1}$ is called the "Bohr radius" and is about 0.028 fm inverse. Similarly, $\epsilon = E/E_\infty$ with $E_\infty = m^2 c^2 / \hbar^2$. The energy is called the "Rydberg" and is about 13.6 electron volts.

Since the ρ variable appears on both sides, it is convenient to work with it rather than with ψ itself. Doing

$$\phi(\rho) = \psi(\rho), \quad (19.7)$$

we have the more complicated equation

$$\frac{d^2 \phi}{d\rho^2} = -\left(\epsilon - \frac{2}{\rho} \right) \phi. \quad (19.8)$$

Now we'd like to find some function ϕ which satisfies Eq. (19.8). In other words, we just want to solve a differential eqn. i.e. linear theory. There is no general, generic, method for solving any given differential equation. You just have to fiddle around. Our equation is not easy, but people have found that it can be solved by the following procedure. First we replace ϕ , which is some function of ρ , by a product of two functions

$$\phi(\rho) = \rho^{-2} \psi(\rho). \quad (19.9)$$

The justification is that you are factoring a "out" of $\phi(\rho)$. You can verify this by taking the first derivative of ϕ plus shifts one problem forward by the right hand side of (19.8).

Substituting (19.9) into (19.8), we get the following equation for ψ :

$$\frac{d^2 \psi}{d\rho^2} - 2\rho \frac{d\psi}{d\rho} - \left(\frac{2}{\rho} + \epsilon - \frac{2}{\rho^2} \right) \psi = 0. \quad (19.10)$$

Since we are free to choose ϵ , let's choose

$$\epsilon^2 = -\alpha. \quad (19.11)$$

and get

$$\frac{d^2 \psi}{d\rho^2} - 2\rho \frac{d\psi}{d\rho} + \frac{1}{\rho^2} \psi = 0. \quad (19.12)$$

You may think we've made life difficult when we wrote (19.11). But here's a happy thing about our new equation is that it can be solved easily in terms of a power series in ϵ . It is possible, in principle, to solve (19.12) that way too. But it is

much harder.) We are saying that Eq. (19.17) is satisfied by some $\phi(\mu)$ which can be written as a series

$$\phi(\mu) = \sum_{k=0}^{\infty} a_k \mu^k, \quad (19.18)$$

in which the a_k are constant coefficients. Now all we have to do is find a suitable infinite set of coefficients a_k which satisfy this equation. The first derivative of this $\phi(\mu)$ is

$$\frac{d\phi}{d\mu} = \sum_{k=1}^{\infty} k a_k \mu^{k-1},$$

and the second derivative is

$$\frac{d^2\phi}{d\mu^2} = \sum_{k=2}^{\infty} k(k-1) a_k \mu^{k-2}.$$

Using these expressions in (19.17) we find

$$\sum_{k=0}^{\infty} \mu^k - 1 + a_0 \mu^{-2} - \sum_{k=1}^{\infty} k a_k \mu^{k-1} + \sum_{k=2}^{\infty} k(k-1) a_k \mu^{k-2} = 0. \quad (19.19)$$

Now, we can't tell if we have succeeded, but we hope so! Let's add better if we replace the first sum by an integration. Since the first term of the sum is zero, we can replace μ^k by $\mu - 1$ without changing anything in the infinite series. Just the change the first sum can equally well be replaced as

$$\sum_{k=0}^{\infty} \mu^k + (-1)^k a_k \mu^{k-1}$$

Now we can put all the sums together to get

$$\sum_{k=0}^{\infty} ((k+1)a_{k+1} - 2ka_k + 2a_0) \mu^{k-1} = 0. \quad (19.20)$$

This expression must vanish for all powers of μ , except μ^0 . It can do that only if the coefficient of each power of μ is separately zero. We will have a solution, for the given equation if we can find a set a_k for which

$$(k+1)a_{k+1} - 2ka_k + 2a_0 = 0. \quad (19.21)$$

For $k=0$ or 1 , this is certainly easy to arrange. Pick any a_0 you like. Then generate all of the other coefficients from

$$a_{k+1} = \frac{2ka_k - 2a_0}{k+1}. \quad (19.22)$$

With this you see a_0 , a_1 , a_2 , a_3 , and so on, and each pair a_k , a_{k+1} certainly satisfy (19.21). We get a series for $\phi(\mu)$ which satisfies (19.17). With it we can make a function which satisfies Schrödinger's equation. Notice that the solution depends on the external energy through the last two values of a ; there is no general a_0 .

You have a solution, but what does it represent physically? We can get an idea by seeing what happens for the positive-negative values of μ . On these, the high-order terms of the series are the most important, so we should look at what happens for large k . When $k \gg 1$, Eq. (19.22) is approximately of the form

$$a_{k+1} \approx \frac{2a}{k} a_k,$$

which means that

$$a_{k+1} \approx \frac{(2a)^k}{k!} a_0. \quad (19.23)$$

But these are just the coefficients of the series for $e^{2a/\mu}$, the function of μ is a rapidly increasing exponential. Every couple of bits μ^2 to produce $e^{2a/\mu}$ see Fig. 19.4.

Eq. (19.14) still gives a solution for $\psi(r)$ which goes like e^{kr} for large r . We have found some bound state solutions but not a physical one. It represents a situation in which the electron is most likely to be near the proton. It is always more likely to be found at a very large radius r . A wave function for a bound electron must go to zero for large r .

We have to think what the best possible way to have the game, and that is obvious. It is just happened by luck that α were equal to $1/3$, where α is only inverse, then Eq. (19.12) would make ω_{fr} = 0. All hope is lost would also be zero. We want ω_{fr} to have an infinite series in a finite polynomial. Any polynomial increases more slowly than e^{kr} , so the term e^{kr} will eventually beat it down, and the function ψ will go to zero for large r . The only bound-state solutions are those for which $\alpha = 1/3$, with $n = 1, 2, 3, 4$, and so on.

Looking back to Eq. (19.16), we see that the bound state solutions to the spherically symmetric wave equation can be stated when

$$-\frac{1}{r^2} - \frac{1}{r^3} + \frac{1}{16} \cdots \cdots \cdots \cdots \cdots \cdots$$

The allowed energies are just those that lie below the Rydberg, $E_R = -m^4/2k^2$, or the energy of the zero-energy level is

$$E_0 = -E_R \frac{1}{n^2}. \quad (19.20)$$

There is, unfortunately, nothing mysterious about negative numbers for the energy. The energies are negative because when we write down its $V = -e^2/r$, we picked our zero point as the energy of an electron free floating far from the proton. When it is close to the proton, its energy is less, so somewhat below zero. The energy is lowest (most negative) for $n = 1$, and increases toward zero with increasing n .

Before the discovery of quantum mechanics, it was known from experimental studies of the spectrum of hydrogen that the energy levels could be described by Eq. (19.21), where R_H was found over the observations to be about 13.6 electron-volts. Bohr then derived a model which gave the same equation and predicted that E_0 should be $m^4/2k^2$. But it was the greatest success of the Schrödinger theory that it could reproduce this result from a basic equation of motion for the electron.

Now that we've resolved our first concern, let's look at the nature of the solutions we get. Putting all the pieces together, each solution looks like this:

$$\psi_n = \frac{C_n(r)}{r} = \frac{e^{-\alpha r}}{r} \psi_n(r) \quad (19.21)$$

where

$$\psi_n(r) = \sum_{i=1}^n a_i r^i \quad (19.22)$$

and

$$a_{n+1} = \frac{n a_n}{(n+1) \alpha}, \quad (19.23)$$

So long as we are entirely interested in the relative probabilities of finding the electron at various places we can pick any number of such terms. We may as well do $a_{n+1} = 1$ (people often choose a_1 so that the wave function is "normalized," that is, so that the integrated probability of finding the electron anywhere in the atom is equal to 1. We have no need to do that just now.)

For the lowest energy state, $n = 1$, and

$$\psi_1(r) \sim e^{-\alpha r}. \quad (19.24)$$

For a hydrogen atom in its ground (lowest-energy) state, the amplitude of finding the electron at any point drops off exponentially with the distance from the proton. It is most likely to be found right at the proton, and the cut-off distance at which the amplitude is about one unit in ρ , or about one Bohr radius, is

Energy $E = 2$ goes to a higher level. The wave function for this state will have two terms. It is

$$\psi_{200} = \left(1 - \frac{r}{a}\right)e^{-r/a^2} \quad (19.29)$$

The wave function for the $1s$ level is

$$\psi_{100} = \left(-\frac{10}{a^3} + \frac{2}{a^2}r^2\right)e^{-r/a^2} \quad (19.30)$$

The wave functions for these first three levels are plotted in Fig. 19.2. You can see the general trend. All of the wave functions approach zero rapidly for very large r values, oscillating a few times. In fact, the number of "nodes" is just equal to $n - 1$. For example, the one $1s$ node corresponds to $n_1 = n - 1$.

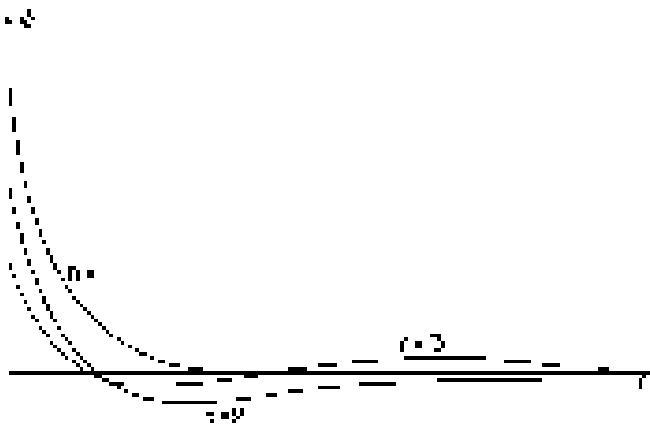


Fig. 19.2. The wave functions for the first three n -values of the hydrogen atom. (The scales are chosen so that the total probabilities are equal.)

19.3 States with an angular dependence

In the wave description of the $1s$, we have found that the probability amplitude to find the electron is spherically symmetric, depending only on r , the distance from the proton. There is no net overall angular momentum. We would like to know about states which may have some angular dependence.

We could, I'm afraid, just investigate the already mathematical problem of finding the functions $\psi(r, \theta, \phi)$, which satisfy the differential equation (19.7), putting in the additional physical conditions that they only represent l numbers ($l = 0, 1, 2, \dots$) for large r . You will find this in some of many books. We are going to take a short cut by using the knowledge we already have about how angular momentum is represented.

The hydrogen atom at any $2l$ level n is a particle with a certain "spin" l , the quantum number of the angular momentum. But if the spin comes from the electron's internal spin, and perhaps its electronic motion, Ψ for each of these two components can independently form an excellent system, and yet still retain both the spatial and time-only about the "orbital" together information. The orbital motion behaves, however, just like a coin. For example, if the orbital spin is turned to zero, the component of angular momentum can be $l = 0, 1, 2, \dots, l$ (for up, or down, respectively, measuring it, $m_l = 0, \pm 1, \dots, \pm l$). Also, all the rotation in addition to other properties we have worked out will apply. (In our discussion we will be ignoring the electron's spin; if we speak of "angular momentum" we will mean only the orbital term.)

Since the potential V at which the electron moves depends only on r , the two components of the Hamiltonian is constant under all rotations. It follows that the angular momentum and its components are conserved. (This is the famous reason why "central field" magnetohydrodynamics may represent not a basic feature of the Coulomb's law problem.)

Now let's think of some possible state of the economy; its income angular structure will be characterized by the quantum number ℓ . Depending on the "orientation" of the total angular momentum with respect to the z -axis, the z-component of angular momentum will be m , which is one of the $2\ell + 1$ possibilities between $-\ell$ and ℓ . Let's say $m = 1$. What what amplitude will the electron be found at some distance r from the origin? At a distance r , the basis cannot have any radial angular momentum around the z -axis. Alright, suppose it is zero, then there can be some nonzero amplitude to find the electron at each distance from the position. We'll call this amplitude $F(r)$. It is the amplitude to find the electron at the distance r up along the z -axis, when the system is in the state $|\ell, m\rangle$, by which we mean orbital spin 1 and a component $m = 1$.

If we know $A_r(r)$ everything is known. For any state $|\ell, m\rangle$ we know the amplitude $A_{\ell m}$ to find the electron anywhere in the atom. Now! Watch this. Suppose we have the atom in the state $|\ell, m\rangle$, that is the amplitude to find the electron at the angle θ , and the distance r from the origin. Put a new z' -axis, say z' , at the angle α (Fig. 19-3), and ask what is the amplitude that the electron will be at the distance r along the new axis z' ? We know that it cannot be found along z' unless its z-component of angular momentum, say m' , is zero. When m' is zero, however, the amplitude to find the electron along z' is $A_{\ell m}$. Therefore the result is the product of two factors. The first is the amplitude that an atom in the state $|\ell, m\rangle$ along the z -axis will be in the state $|\ell, m' = 0\rangle$ with respect to the z' -axis. Actually this amplitude by $R_{\ell m}(\theta)$ and you have the amplitude $A_{\ell m}(r)$ to find the electron at (r, θ, ϕ) with respect to the original axes.

Let's write it out. We have worked out earlier the transformational matrices for rotations. Going from the frame x, y, z to the frame x', y', z' of Fig. 19-3, we have once last rotated the z' -axis by the angle α , and then rotated about the new x' -axis (ℓ) by the angle θ . The combined rotation is the operator

$$R_{\ell m}(\theta) R_{\ell m}(\alpha)$$

The amplitude to find the state $|\ell, m' = 0\rangle$ after the rotation is

$$|\ell, m | R_{\ell m}(\theta) | \ell, m' \rangle. \quad (19.37)$$

Our result then is

$$A_{\ell m}(r) = |\ell, 0 | R_{\ell m}(\theta) | \ell, m' \rangle \langle \ell, m' | r \rangle. \quad (19.38)$$

Two orbital motion can have only integral values of ℓ . If the electron can be found anywhere $m \neq 0$, there is some amplitude to have $m = 0$ in that direction. And $m = 0$ occurs exactly in integral spins. The selection rules for $\ell = 1$ are given in Table 17-2. For larger ℓ you can use the general formulae we worked out in Chapter 18. The matrices for $S_x(\theta)$ and $S_y(\theta)$ appear again, but you know how to combine them. For the general case you would operate the matrix $|J, m |$ and operate with $R_{\ell m}(\theta)$ to get the new state $R_{\ell m}(\theta) | J, m \rangle$. Then you operate on this state with $S_x(\theta) + S_y(\theta) + S_z(\theta)$ and $S_z(\theta)$ which is just $\ell = 1, m = 0$. Multiplying by $|J, m |$ gives the answer already (19.11).

The various elements of the rotation operator have algebraic differences of i and $-i$. The joint distribution function which appears in (19.38) also shows up in many kinds of problems which involve waves in spherical geometry, and in fact have a special name. Not everyone uses the same convention, but one of the most common ones is

$$|\ell, 0 | R_{\ell m}(\theta) | \ell, m' \rangle = i \ell m' Y_{\ell m'}(\theta). \quad (19.39)$$

The functions $Y_{\ell m'}(\theta)$ are called the spherical harmonics, and i is just a normalization factor which depends on the definition chosen for $Y_{\ell m}$. For the usual definition

$$\psi = \frac{e^{-\frac{r^2}{2}}}{\sqrt{3\pi}}, \quad (19.40)$$

With this notation, the hydrogen wave functions can be written:

$$\psi_{100}(r) = Y_{100}(\theta) \psi_1(r) \quad (19.41)$$

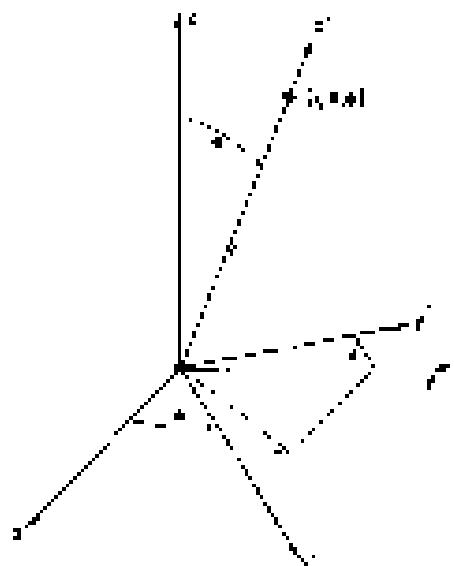


Fig. 19-3. The angle β , α , γ is an orientation of the x', y', z' coordinate frame.

The angle θ (where $\theta \neq 90^\circ$) is important not only in many quantum-mechanical problems but can in many areas of classical physics. In such the θ operator appears with no interpretation. As another example of the use of quantum mechanics, consider the deexcitation of a neutron state by $N\gamma$ radiation or thermal γ , the last chapter which decays by emitting two photons and going into $1S^1$:



Suppose that the excited state has some spin (necessarily an integer) so that the "component" of angular momentum is m . We might now ask the following question: what is the amplitude that we will find the neutron going off at a direction θ with respect to the x -axis and the angle ϕ with respect to the y -axis? as shown in Fig. 17.4.

To solve this problem we take $\langle 1S | \gamma | 1S \rangle$, the Clebsch-Gordan coefficient. A decay in which one particle goes straight and only emits energy from a source is zero. Thus $\langle 1S | \gamma | 1S \rangle = 0$ and the two particles have spin zero, and because their sum is that they may couple momentum about the x -axis. The total A is amplitude a $1S^1$ into solid angle. Then, to find the my amplitude for a decay is the infrared angle of Fig. 17.4 all we need to know is what amplitude the given initial state has zero angular momentum about the decay direction. The amplitude for the decay of S and g is then a three-dimensional unit vector ($| \psi \rangle$) with respect to the x -axis will be $\langle \psi | \psi \rangle$ the state $|\psi\rangle$ with respect to x , the decay direction. The final amplitude is just what we have written in (17.2). The probability to see the particle at θ, ϕ is

$$P(\theta, \phi) = n^2 |\langle 1S | \psi | \psi \rangle \langle \psi | 1S \rangle|^2$$

As an example, consider a situation with $\theta = 135^\circ$ and $\phi = 0^\circ$. The values of n from Table 17.2 we know the necessary amplitudes. They are

$$\begin{aligned} \langle 1, 0 | R_x(0)R_y(0) | 1, -1 \rangle &= -\frac{1}{\sqrt{2}} \sin \pi/2, \\ \langle 1, 0 | R_y(0)R_z(0) | 1, 0 \rangle &= 0.65, \\ \langle 1, 0 | R_z(0)R_y(0) | 1, -1 \rangle &= -\frac{1}{\sqrt{2}} \cos \pi/2. \end{aligned} \quad (17.26)$$

Now are the three possible angular distribution amplitudes depending on the nature of the initial nucleus.

Amplitudes such as the ones in (17.26) appear often and are sufficiently simple that they are given several names. If the angular distribution amplitude is proportional to my value of the first function, it is the combination of the two, "The system has an overall angular momentum of $m = 1$." Or we may say, "The $N\gamma$ system is a wave packet." Or we say, "The angular momentum is $m = 1$." Because there are so many ways of saying the same thing it is useful to have a dictionary. If you are going to understand what other physicists are talking about, you will just have to memorize the language in Table 17.2 for your collection of orbital angular momentum:

If the orbital angular momentum is $m = 0$, then there is no change except you rotate the coordinate system without variation with angle, the "independence" of angle is an $m = 0$ function, say $1, 0$ is an $m = 0$ function, and there is only one term in (17.26) if the angular dependence is concerned. If the orbital angular momentum is 1, then the amplitude of the angular variation may be any one of the four functions given, depending on the value of $m = 1$, it may be a linear combination, these are called "vector," and there are three of them. If the orbital angular momentum is 2, then there are the five functions shown. Any linear combination is called an " $m = 2$ " or a "tensor" amplitude. Now you can immediately guess what the next letter is—what should come after $1, 0, 2, 3$? Well, of course, it's $4, 5, 6$, and so on forever and always. The letters don't mean anything—(they just carry over something—they meant "the 2 lines," "principal" lines, "diffuse" lines and so on).

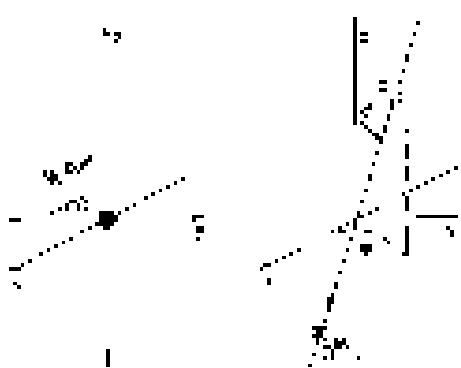


Fig. 17.4. The decay of an excited state of $N(1S)$.

Table 19-1

Multiplets of orbital angular momentum
($l = 0, 1, 2, \dots, n - 1$)

Orbital angular momentum, l	Angular component, m	Angular dependence of amplitude	Sign	Number of states	Central parity
0	0	1	+	1	-
1	-1	$\frac{1}{\sqrt{2}}(\sin \theta e^{i\phi})$	-	2	-
1	0	$\frac{1}{\sqrt{2}}(\cos \theta)$	+	2	+
1	+1	$\frac{1}{\sqrt{2}}(\sin \theta e^{-i\phi})$	-	2	-
2	-2	$\frac{\sqrt{6}}{4}(\sin^2 \theta e^{i\phi})$	-	1	-
2	-1	$\frac{\sqrt{3}}{2}(\sin \theta \cos \theta e^{i\phi})$	-	4	-
2	0	$\frac{1}{2}(3 \cos^2 \theta - 1)$	+	5	+
2	+1	$\frac{\sqrt{3}}{2}(\sin \theta \cos \theta e^{-i\phi})$	-	4	-
2	+2	$\frac{\sqrt{6}}{4}(\sin^2 \theta e^{-i\phi})$	-	1	-
3	-3	$\frac{1}{2}\sqrt{10}(\sin^3 \theta e^{i\phi})$	-	1	-
3	-2	$P_2(\cos \theta)$	-	3	-
3	-1	$\frac{1}{2}\sqrt{15}(\sin^2 \theta \cos \theta e^{i\phi})$	-	6	-1
3	0	$\frac{1}{4}(5 \cos^3 \theta - 3 \cos \theta)$	+	7	-
3	+1	$\frac{1}{2}\sqrt{15}(\sin^2 \theta \cos \theta e^{-i\phi})$	-	6	-1
3	+2	$\frac{1}{2}\sqrt{10}(\sin^3 \theta e^{-i\phi})$	-	1	-
3	+3	$\frac{1}{2}\sqrt{6}(\sin^3 \theta e^{i\phi})$	-	1	-

"three-quantum" uses of the optical spectra of atoms. But those were in the days when people did not know where the three came from. All of them were given special names, so we may just call them p_0, p_1 , and p_2 .

The angular functions in the last group above, p_{-1}, p_0 , and p_1 , are antisymmetric about the horizontal. States that repeat out in front. Sometimes, they are called "oblique harmonics," and sometimes P_{-1} , etc. Sometimes they are zero at $\theta = \pi/2$, and, if $m = 0$, simply as $P_0(m=0)$. The functions $p_2(\cos \theta)$ are called the "Legendre polynomials" in physics, and the functions $p_m(\cos \theta)$ are called the "associated Legendre functions." You will find tables of these functions in many books.

Note, particularly, that all the amplitudes p_m given have the property that they have the same parity; for odd m they change sign under π inversion and for even m they don't change. So we can write the parity of a state of orbital angular momentum $|l, m\rangle$ as

As we have seen, these angular distributions may give rise to a nuclear disintegration or some other process, as in the distribution of the amplitude to line an electron from one place to another place in the hydrogen atom. For instance, if an electron is in a plane ($\theta = \pi/2$), the amplitude to find it in the point θ is dependent on the angle ϕ in many possible ways—but all are linear combinations of the three functions for $l = 1$ at Table 19-1.

Of course the case $l = 0$ is special. That's interesting. That means that the amplitude is positive only in the upper part ($\theta < \pi/2$), negative in the lower part ($\theta > \pi/2$), and is zero when $\theta = \pi/2$. Required the amplitude to be real and nonzero only by finding the exact ϕ values with $\theta = \pi/2$ shown in Fig. 19-5, and it is dependent of ϕ ; this angle's existence is responsible for the fact that in molecular binding the electron can't be in $\theta = \pi/2$. Let's refer again to our diagram of arrangement in the atom. At the direct vertices of interest almost all

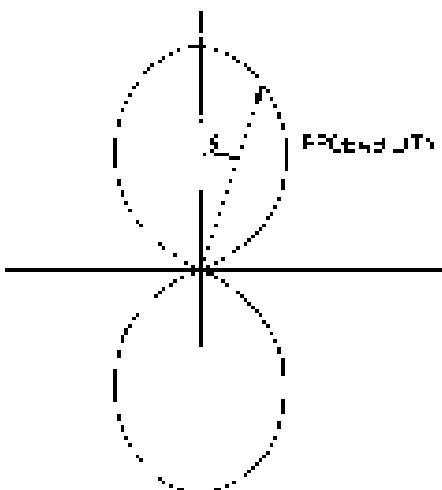


Fig. 19-5. A polar graph of $|\psi|^2/\lambda$, which is the relative probability of finding an electron at various angles from the z -axis for a given l . In each sketch, $\theta = \pi/2$ and $\phi = 0$.

14-4 The general solution for hydrogen

In Eq. (19.35) we have written the wave functions for the hydrogen atom as

$$Y_{l,m}(\theta) = R_{l,m}(r, \theta) Y_l^m(\theta). \quad (19.37)$$

These wave functions must be solutions of the Schrödinger equation (19.7). To do this with the r -wave, $R_{0,0}(r, \theta, \phi)$ in Eq. (19.37) into (19.7), you get

$$\begin{aligned} \frac{\partial^2}{\partial r^2} (r^2 R_{0,0}) - \frac{E_0}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial R_{0,0}}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 R_{0,0}}{\partial \phi^2} \\ = -\frac{2m}{\hbar^2} \left(E_0 + \frac{l^2}{r^2} \right) R_{0,0}. \end{aligned} \quad (19.38)$$

Now multiply through by r^2/r , and rearrange terms. The result is

$$\begin{aligned} \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial R_{0,0}}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 R_{0,0}}{\partial \phi^2} \\ = -\left[\frac{r^2}{\hbar^2} \left[\frac{\partial^2}{\partial r^2} (r^2 R_{0,0}) + \frac{2m}{\hbar^2} \left(E_0 + \frac{l^2}{r^2} \right) \right] \right] Y_{0,0}. \end{aligned} \quad (19.39)$$

The left-hand side of this equation depends on θ and ϕ , but not on r . We consider what value we choose for r , the left side doesn't change. This means that we can fix the right-hand side. Although the quantity in the square brackets has r 's all over the place, the whole quantity cannot depend on r , otherwise we wouldn't have an equation valid for all r . As you can see, the bracket does depend on r via E_0 . It must be zero (constant). Its value may well depend on the l -value of the wave we are applying, since the function R_0 must be the one appropriate to that state; well call the constant K_0 . Equation (19.39) is therefore equivalent to two eqs., which

$$\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial R_{0,0}}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 R_{0,0}}{\partial \phi^2} = K_0 Y_{0,0}, \quad (19.40)$$

$$\frac{1}{\hbar^2} \frac{\partial^2}{\partial r^2} (r^2 R_0) + \frac{2m}{\hbar^2} \left(E_0 + \frac{l^2}{r^2} \right) R_0 = K_0 Y_{0,0}. \quad (19.41)$$

Now look at what we've done. For any state described by l and m , we know the function $Y_{l,m}$; we can use Eq. (19.40) to determine the constant K_0 . Putting K_0 into Eq. (19.41) we have a differential equation for the function $R_0(r)$. If we can solve that equation for $R_0(r)$, we have all of the pieces to put into (19.37) to give $\Psi(r)$.

What is K_0 ? First, notice that it must be the same for all m (which go with a particular l), so we can pick any m we want for $Y_{l,m}$ and plug it into (19.40) to solve for K_0 . Perhaps the easiest one to use is $Y_{0,0}$. From Eq. (18.24)

$$\langle 0,0 | 1,0 \rangle = e^{i k_z} \langle 0,0 | 1,0 \rangle. \quad (19.42)$$

The matrix element for $R_0(r)$ is also quite simple:

$$\langle 0,0 | R_0(r) | 1,0 \rangle = \delta(l,1) \delta(m,0), \quad (19.43)$$

where δ is some number ± 1 . Combining the two, we obtain

$$Y_{0,0} = e^{i k_z} \sin \theta. \quad (19.44)$$

You can with some work show that the constant of Eq. (19.37) is not 0, but it is also easy to work out from first principles following the ideas of Section 17.4. A state $| 1,0 \rangle$ can be made out of 21 spin-one-half particles all with spins up; since the state $| 1,0 \rangle$ would have 1 up and 1 down. Under σ_z rotations, the amplitude that all up-spins are up is $+1/\sqrt{2}$, and that all up-spins goes down is $-i/\sqrt{2}$. We are looking for the amplitude that all up-spins stays up, while the other two up-spins go down. The amplitude for that is $(+i/\sqrt{2})(-i/\sqrt{2})$, which is the same as $\delta(l,1)$.

Picking out function ψ_0 from Eq.(19.3) gives

$$E_0 = \frac{1}{2} \theta^2 - \frac{1}{r_0}. \quad (19.4)$$

Now that we have computed E_0 , Eq.(19.4) tells us about the total energy $E_0(\theta)$. And, of course, just like Schrödinger's equation with the angular component, by its equivalent $K(E)/r^2 = \nabla^2 \psi_0 + E_0(\theta) \psi_0$ (Eq.(19.3)) in the form of Eq.(19.4), we have:

$$\frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} \psi_0 = -\frac{2m}{\hbar^2} \left[\frac{1}{r^2} \frac{\partial^2}{\partial r^2} \psi_0 + \frac{E_0(\theta) - 1/r^2}{2mr^2} \right] \psi_0. \quad (19.5)$$

A centrifugal term has been added to the potential energy. Although at first this term may seem mathematical, it does have a simple physical origin. We will give you an explanation where it comes from in terms of conservation arguments. But perhaps you will not find it quite so mysterious.

Think of a classical particle moving around an elliptical orbit. The total energy is conserved and is the sum of the potential and kinetic energies:

$$E = E(r) + \frac{1}{2} m v^2 = \text{constant}$$

In general, since the radial motion is radial or spherically symmetric, we can write

$$v^2 = r^2 \dot{\theta}^2 + \frac{r^2 \dot{r}^2}{r^2}.$$

Now the angular momentum vector is conserved and is equal to $L = m r^2 \dot{\theta}$. So we can write

$$m r^2 \dot{\theta} = L, \quad \text{or} \quad \dot{\theta} = \frac{L}{m r^2}$$

and the energy is

$$E = \frac{1}{2} m r^2 \dot{r}^2 + \frac{L^2}{2m r^2} + \frac{L^2}{2mr^2}.$$

If there were no angular momentum we would have just the first two terms. Adding the angular momentum L does two things: first what's left of the term L^2/mr^2 is the potential energy contribution. But this is almost exactly the same term in Eq.(19.4) (the only difference is that (L^2/mr^2) appears as the angular momentum term instead of (L^2/mr^2) as in Eq.(19.3)). But we have seen, back in Chapter 10, Volume 1, Section 10.7.1 that this is not the situation that is usually referred to when a quantum-mechanical argument with a correct classical-mechanical calculation. We can then understand the new term as a "correction" (or "perturbation") which goes like the coupling force in the equation of the equations of motion for a classical system. (See the discussion of "perturbations" in Volume 1, Section 10.7.)

We are now ready to solve Eq.(19.4) for $\psi_0(\theta)$. It is very much like Eq.(19.3); the same technique will work again. Expanding ψ_0 as before, and you get in Eq.(19.4) what will look the odd-shaped term

$$(V + 1) \sum_{n=0}^{\infty} a_n \theta^{2n}. \quad (19.6)$$

This can now also be written as

$$(V + 1) \left\{ \frac{a_0}{\theta} + \sum_{n=1}^{\infty} a_n \theta^{2n-1} \right\}. \quad (19.7)$$

(We have taken out the a_0/θ term and rearranged the remaining terms as shown by Eq.(19.6) instead of Eq.(19.4) we have:

$$\sum_{n=0}^{\infty} (V + 1) \theta^{2n-1} a_n = -2ma^2 - 1/a^2 \theta^2. \quad (19.8)$$

See Appendix A for the solution.

There is only one term in ψ_{nlm}^+ that is odd by m . The coefficient of $m=0$ is zero (unless $l=0$ and we have our previous solution). Each of the other terms is made even by having the x_2 and x_3 terms come out even for even m . This leads to Eq. (19.33) by

$$\alpha_{nlm} = \frac{2\pi l+1}{2l+1} - \frac{\pi m}{l(l+1)} = \frac{\pi}{l(l+1)}. \quad (19.33)$$

This is the only significant change from the spherically symmetric case.

As before the series must terminate if we are to have solutions which can represent R and ψ_{nlm} too. The series will end at $k = n$ if $n > l$. We get again the same separation eq. (19.1), thus it must be equal to $1/k$, where k is some integer. However, Eq. (19.33) also gives a new restriction. It requires k to be equal to l , so the denominator becomes zero and α_{nlm} is infinite. That is, unless $n = l$, Eq. (19.33) implies that all successive α_{nlm} are zero until we get to $k = l$, which can't be done. This means that k must start at $k = l$ and end at n .

The final result is that for any l there are many possible solutions which are given by R_{nlk} , where $k \geq l + 1$. Each solution has its energy

$$E_{nlk} = \frac{me^2}{2k^2} \left(\frac{1}{n^2} \right). \quad (19.34)$$

The wave function for the state of this energy with the angular quantum number k and m is

$$\psi_{nlkm} = R_{nlk}(r) Y_{lm}(x), \quad (19.35)$$

with

$$dE/dk = e^2 \sum_{n=1}^{\infty} \frac{4\pi^2}{n^2}. \quad (19.36)$$

The coefficients a_k are obtained from (19.30). We have finally a complete description of the states of a hydrogen atom.

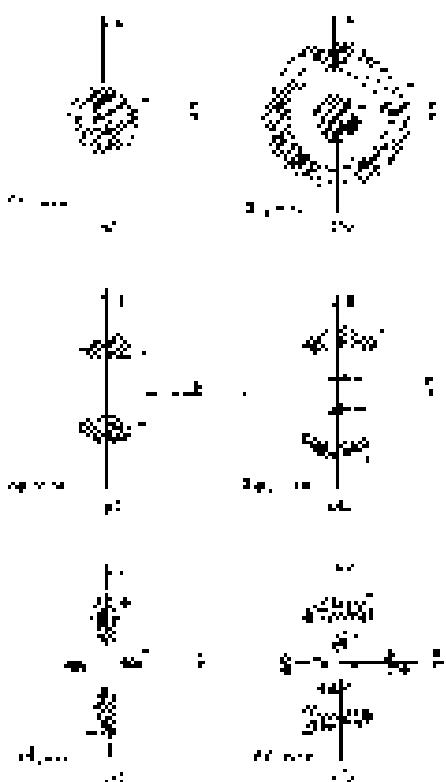


Fig. 19-6 Energy levels showing the angular nature of some of the hydrogen wave functions. The second column, top down, are the methylene wave functions. The plus and minus signs show the relative sign of the amplitude in each region.

19-5 The hydrogen wave functions

We're now in a position to discuss the Schrödinger's equation for an electron in a hydrogen atom characterized by three quantum numbers n, l, m , all integers. The angular distribution of the electron amplitude can't be arbitrary since it must satisfy the condition $\int \psi^2 d\tau = 1$. This is satisfied by the quantum number of total angular momentum and is the "magnitude" quantum number, which can take the values $-l, -l+1, \dots, l-1, l$. For each angular configuration, various possible radial distributions R_{nlk} of the electron amplitude are possible, those calculated by the formula given earlier, which can happen if $k \geq l + 1$. The energy of the state depends only on n , the n regions were nonoverlapping.

The lowest energy, or ground state is one state. It has $l = 0$, $n = 1$, and $m = 0$. It is a "degenerate" state: there's only one $k = 1$ energy, and it's spherically symmetric. The amplitude is finite the electron is a maximum at the center, and falls off monotonically with increasing distance from the center. We can visualize the electron amplitude as a blob as shown in Fig. 19-6(a).

There are other states with higher energies, for $n = 1, 2, 3, \dots$ the next energy there's only one version ($k = 0$), and they are all spherically symmetric. These states have quadrupoles which alternate in sign and we have three with nonvanishing m : the same $n = 1$ hydrogen model will form three planes passing through zero. One possibility of $n = 2$, $l = 1$, for example, will look as sketched in Fig. 19-6(b). (The $l = 1$ states indicate regions where the amplitude is large, while the plus and minus signs indicate the relative phases of the quadrupole.) The energy levels of the system are shown in the first column of Table 19-1.

There are also the l states which $n = 2$. For example, which could be $l = 2$ or greater. There are three states of the same size $k = 2$, and the $m = -1, 0, 1$, and $m = -2$. The energy levels are as shown in Fig. 19-2. The angular dependences of these states are given in Table 19-2. For instance, for $m = 0$, if $n = 3, l = 1$,

amplitude is positive for $\sin \theta = 0$, it will be negative for $\theta = 90^\circ$. There is a nodal plane consistent with the λ -spins. For $m > 2$ there are also spherical nodes. The $n = 2, m = 0$ amplitude is sketched in Fig. 19-6(a), and the $n = 2, m = 2$ wave function is sketched in Fig. 19-6(b).

Now in the three first states we represent a kind of "quantum" of spin, there should be similar contributions with the peaks of amplitude along the x axis or along the y -axis. Are these perhaps the $m = +1$ and $m = -1$ states? Now obviously we have three states with m values plus, say three combinations of the λ 's. These will also be stationary states of the stand energy. It turns out that the "x" state—which corresponds to the "elliptical" $m = 0$ state of Fig. 19-6(a)—is a linear combination of the $m = +1$ and $m = -1$ states—the corresponding "y" state is another combination. Specifically, we mean that

$$\begin{aligned} \psi_{\text{tot}}^{\text{x}} &= 1.00 \\ \psi_{\text{tot}}^{\text{y}} &= \frac{1}{\sqrt{2}}(1.00 + i) \\ \psi_{\text{tot}}^{\text{z}} &= \frac{1}{\sqrt{2}}(1.00 - i) \end{aligned}$$

These stages I took the same when referring to other particular cases.

The states $n = 2$ have five possible values of m for each energy. The lowest-energy levels ($n = 1$) in the hydrogen atom shown in Fig. 9-3. The angular dependences get more complicated. Just as you can see, the $m = 0$ state has two nodal planes in the wave function reversed phase from $m = +1$ to $m = -1$ as you go around from the north pole to the south pole. The rough form of these situations is sketched in (a) and (c) of Fig. 19-6 for the $m = 0$ states with $n = 3$ and $n = 4$. Again, the larger n 's have spherical nodes.

We will now try to describe why there are the possible states. You will find the hydrogen wave functions described in more detail in many books. Two good references are L. Pauling and J. A. Wilson, *Introduction to Quantum Mechanics*, McGraw-Hill (1935); and R. B. Leighton, *Principles of Modern Physics*, McGraw-Hill (1959). You will find in the chapters of some of the functions and pictorial representations of many states.

We would like to mention one singular feature of the wave functions the figure 19-6 for $n > 0$ the amplitudes are zero at $r = 0$ cm. This is not surprising, since it is hard for an electron to have enough momentum when its radius r is very small. As this process, r higher and, the more the amplitudes are "pushed away" from the center. If you look at the way the radial functions $R(r)$ vary for small r , you find from (19.5) that

$$R_{n,l}(r) \sim r^l.$$

Such a dependence on r means that for larger r 's you have to go further from $r = 0$ before you get an appreciable amplitude. This behavior is naturally determined by the centrifugal force given in the radial equation, but something will apply for any potential $V(r)$ values lower than E_n^2 for small r , which does obtain potentials $V(r)$.

19-4 The periodic table

We would like now to apply the theory of the hydrogen atom in an approachable way to get some understanding of the chemist's periodic table of the elements. For an element with atomic number Z there are Z electrons held together by the electric attraction of the nucleus but with mutual repulsion of the electrons. The general case is given below, it has been given Schrödinger's equation for Z electrons in a Coulomb field. For helium the equations is

$$-\frac{\hbar^2 \partial^2}{2m} \psi = -\frac{e^2}{4\pi \epsilon_0 r_1} (1/r_1) + \left(-\frac{2e^2}{r_1} - \frac{2e^2}{r_2} - \frac{e^2}{r_{12}}\right) \psi$$

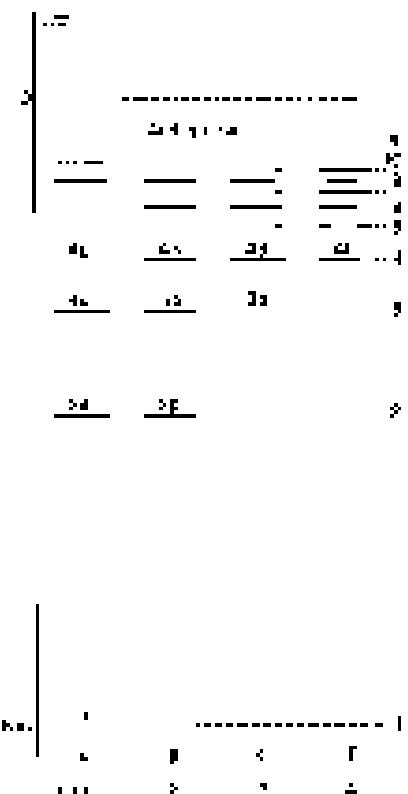


Fig. 19-7. The energy level diagram for the hydrogen atom.

where $\hat{V}_1^2 = -\frac{e^2}{r_1}$ is the potential which operates on r_1 , the distance of one electron. \hat{V}_2 depends on r_1 and $r_{12} = r_1 - r_2$. This is the part reflecting the spin of the electrons. To find the stationary states and energy levels we would have to find ψ in terms of $\hat{\psi}$ in form

$$\psi = \hat{\psi}(r_1, r_{12}) e^{-iE t / \hbar}$$

The general form depends on $\hat{\psi}$, which is a function of six variables—the coordinates and spins of the two electrons. Numerical methods for solving this Schrödinger equation for the lowest energy states have been developed by a myriad methods.

With 1, 2, 3, 4, 5 each one it is hopeless to try to obtain exact solutions without using computers to see that quantum mechanics has given us some understanding of the parameters. It is possible, however, even with a stopgap approximation—and some guesswork—understanding, at least qualitatively, many chemical properties which show up in the periodic table.

The chemical properties of atoms are determined primarily by their bound energy states. We can use the coupling appropriate theory to find these states and their energies. First, we neglect the effect of spin, saying that we neglect the exclusion principle and say that any particular electronic state can be occupied by only one electron. This means that any particular orbital configuration can have up to two electrons—one with spin up, the other with spin down. Next we disregard the details of the interactions between nucleons in our field approximation, and we just take a nuclear mass as a constant plus the combined field of the nucleus and all the other electrons. For them, with the 10 electrons, we say that one increases an average potential due to the nucleus plus the other nine electrons. We are given then that in a hydrogen-like atom for each proton we get a E_p which is a $1/r$ -field modified by a spherically symmetric charge density coming from the other electrons.

In this model each electron acts like an independent particle. The angular dependence of its wave function will be just the same as the ones we had for the hydrogen atom. There will be various orbitals, and so on, and they will have the same probabilities as. Since k_F is no longer given as $1/a$, the radial part of the wave functions will be somewhat different, but as far as qualitatively the same, so we will use the same radial quantum numbers n . The energies of the states will also be somewhat different.

II

With these ideas, let's see what we get. The ground state of hydrogen has $l = m = 0$ and $s = 1/2$ (so the electron configuration is $1s$). The energy is -13.6 eV . This means that it takes 13.6 electron volts to pull the electron off the atom. We call this the "ionization energy", E_I . A more reasonable energy seems that it is harder to pull the electron off. And, in general, that the material is chemically less active.

III

Now take helium. Both electrons can be in the same lowest state (one spin up and the other spin down). In this lowest state the $\hat{V}_1^2 + \hat{V}_2$ is a potential which is for small r_{12} a sum of two fields $1/r_1$ and for very large r_{12} a Coulomb field for $z = 2$. The result is a "hydrogen-like" result with a somewhat lower energy. But electron escape identical to states ($l = 0, m = 0$). The observed ionization energy for helium (one $1s$) is 24.6 electron volts. And the "shell II" is now full—so it only two occurs—there is practically no tendency for an electron to be attracted from another atom. Hence ∞ chemically inert.

IV

The ground hydrogen has a charge of $-e$. The electron charge is $+e$ so hydrogen-like, and the three electrons all occupy the lowest three energy levels. Two will go into $1s$ and one third will go into $2s$ state. But $m_l = 0$ or $l = 0$ in hydrogen these states have the same energy, but in this atom they

don't, for the following reason. Remember that a $2s$ orbital has some emptiness to be near the nucleus while the $2p$ state does not. That means that a $2s$ electron will feel some of the triple-electron charge of the L electrons but that a $2p$ electron will stay out where the field looks like the Coulomb field of a single charge. The extra attraction lowers the energy of the $2s$ state relative to the $2p$ state. The energy levels will be roughly as shown in Fig. 19-8, which you should compare with the corresponding diagram for hydrogen in Fig. 19-7. Be^{+1} from now will have two electrons in $1s$ states and one in a $2s$. Since the $2s$ electron has a higher energy than a $1s$ electron, it is relatively easily removed. The ionization energy of lithium is only 5.3-electron volts, and it is quite reactive chemically.

So you can see the patterns which develop; we have given in Table 19-2 a list of the first 36 elements, showing the states occupied by the electrons in the ground state of each atom. The Table gives the ionization energy for the most loosely bound electron, and the number of electrons occupying each "shell"—by which we mean states with the same n . Since the different isotopes have different

Table 19-2
The electron configurations of the first 36 elements

Z	Element	IP (eV)	n	Electron Configurations									
				1s	2s	2p	3s	3p	3d	4s	3p	4d	5s
1	H: hydrogen	13.6	1										
2	He: helium	24.6	2										
3	Li: lithium	5.3											
4	Be: beryllium	9.3											
5	B: boron	11.2											
6	C: carbon	11.2		FILLED	2	2							
7	N: nitrogen	14.5		(2)	2	3							
8	O: oxygen	13.6				2	4						
9	F: fluorine	17.4				2	5						
10	Ne: neon	21.6				2	8						
11	Na: sodium	5.1											
12	Mg: magnesium	7.6											
13	Al: aluminum	6.0											
14	Si: silicon	8.1											
15	P: phosphorus	10.5											
16	S: sulfur	10.0											
17	Cl: chlorine	12.9											
18	Ar: argon	15.8											
19	K: potassium	4.1											
20	Ca: calcium	6.1											
21	Sc: scandium	6.5											
22	Ti: titanium	6.6											
23	V: vanadium	6.7											
24	Cr: chromium	6.8											
25	Mn: manganese	7.4											
26	Fe: iron	7.9											
27	Co: cobalt	7.9											
28	Ni: nickel	9.6											
29	Cu: copper	7.7											
30	Zn: zinc	9.3											
31	Ga: gallium	6.1											
32	Ge: germanium	7.9											
33	As: arsenic	9.3											
34	Se: selenium	9.7											
35	Br: bromine	11.9											
36	Kr: krypton	14.0											

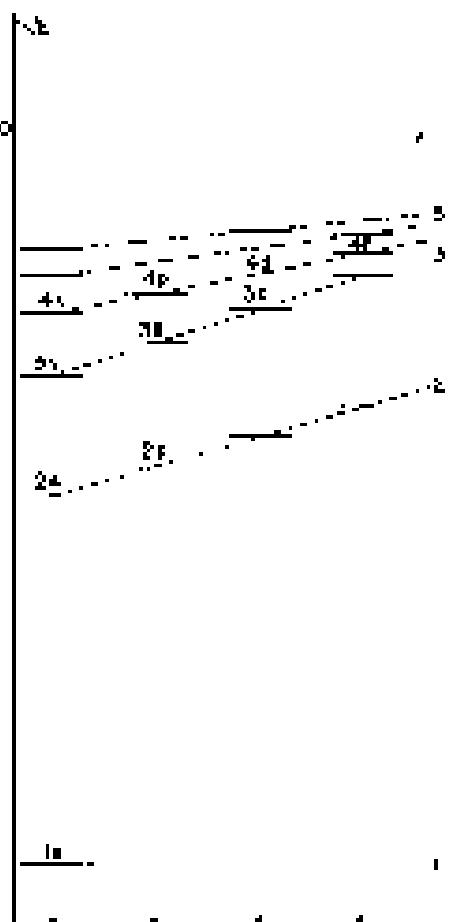


Fig. 19-8. Schematic energy-level diagram for an atomic electron with other electron shells present. The scale is not the same as in Fig. 19-7.

strength, and Rydberg corresponds to a value of 1 of 2021. It provides the same different radial electron orbitals, though all have the same energy—except for some very small effects we are neglecting.

B

"B" can be like A from except that it has two electrons in the $2p_{z}$ state, as well as two in the $2p_{x,y}$ states.

B vs. A

Boron has 3 valence. The lithium ground-state state. The case $1 \times 1 = 1$ is a different type state, so we can keep adding electrons until we get more than one. This does not happen. As we add these electrons we get an increasing Z, so the ψ_{1s} is electron distribution is pushed closer and closer to the nucleus and the energy of the $1s$ state goes down. By hydrogen example, the ionization energy goes to $11 + 9 + 8 + \dots$. Boron does not easily give up an electron. Also there are no two lone-pair electrons to be filled, so it's unlikely to gain an extra electron. Boron is thermally inert. On the other hand, due to how strongly one electron can drop into a state of low energy, so it is quite active in chemical reactions.

N vs. A

With nitrogen, the electron must start a new shell—going into a $2p$ state. The energy level of this state is much higher; the ionization energy, 102.5 kJ/mol, and so it can't easily be filled. From sodium to oxygen the $n=1$ states with $m_l = 0$ are occupied in exactly the same sequence as for lithium— $1s^2$. Angular configurations of the electrons in the outermost valence shell have the same sequence, and the progression of ionization energies is quite similar. You can see why the chemical properties repeat with increasing atomic number. Magnesium is very chemically much like boron, nitrogen, silicon, and chlorine like oxygen. Argon is inert like neon.

You may have noticed that there is a slight irregularity in the sequence of ionization energies between nitrogen and oxygen, and a similar one between sulfur and oxygen. The last electron is bound to the oxygen atom somewhat less than we might expect. And sulfur is similar. Why should that be? We can understand—if we put in just a little bit of the effects of the interactions between individual electrons. Think of what happens when we put the last electron onto the oxygen atom. It has six possibilities—three possible parities, each with two spins. Imagine that the electron goes with spin up into the $n=2$ state, which we have described as the $2p_z$ state because it looks "vertical." Now what will happen to carbon—which has two $2p$ electrons? One of them goes into the $2p_z$ state, where will the second one go? It will want lower energy if it stays away from the first electron, which it can do by going into, say, the $2p_x$ state of the $2p$ shell. (The electron, however, gets three numbers: $l=1$, $m_l = +1$ and $m_s = -1$ states.) Next, when we go to nitrogen, the three $2p$ electrons will have the lowest energy if mutual repulsion if they join the each into the $2p_x$, $2p_y$ and $2p_z$ configurations. For oxygen, however, the j_1 is 2π . The $2p_x$ and $2p_y$ configurations are at the $2p_z$ state—with opposite spins—but strongly repelled by the electron already in that state, so its energy will not be as low as it might otherwise be, and it is more easily removed. That explains the break in the sequence of binding energies which appears between nitrogen and oxygen, and between phosphorus and silicon.

N vs. Zn

After oxygen, you would expect, there's that the next electrons would start to fill up the $3d$ states. But they don't. As we described earlier—and illustrated in Fig. 19.7—the higher angular momentum states get pushed out in energy. By the time we get to the $3d_{10}$ state, they are pushed to an energy a little bit above the energy of the $4s$ state. So it's possible in the last electron, just like the $4s$ state. After zinc, 30, 31.

well as had two electrons or valence, the $3d$ shells begin to be filled for scandium, titanium, and vanadium.

The energies of the $3p$, $3d$, and $4s$ states are so close together that small changes can shift the balance either way. By the time we get to full four electrons in the $3d$ states, their repulsion raises the energy of the $4s$ state just enough that its energy is slightly lower than the $3d$ energy, and it stays in that order. For chromium we don't get a $3+4$ combination because we don't have a $4s$ electron, but instead a $3d$. In combination the other electrons tend to get "cooperative fills" and the shell capacities of the $3d$ -shell are then increased one by one in this way, with copper.

Since the outermost shell of magnesium, iron, cobalt, and nickel have the same configurations, however, they all tend to have similar electron properties. This effect is much more pronounced in the second transition which adds the $3d$ outer shell but a progressively filling outer shell which has such less influence on their electrical properties.

In copper the electron is pulled from the $3d$ shell, finally completing the $3d$ shell. The energy of the $3d$ configuration, however, is closer to the $3s$ configuration for copper than for the previous elements nearby, which is the balance. For this reason the two last electrons of copper are easily removable, and copper has a valence of either $+1$ or $+2$ (it sometimes acts as though its electrons were in the $3d$ configuration). Similar things happen at other places and account for the fact that other metals, such as iron, combine chemically with either of two valences. By now both the $3d$ and $4s$ shells are fully filled and empty.

Group 12

From gallium to krypton the sequence becomes increasingly regular. Thus the $4s-4f$ shell contains double the energy, and the chemical properties repeat themselves from boron to neon and aluminum to argon.

Boron, like arsenic and selenium, is known as "tribond" gas. All three are electronically "inert". This means only that, by using filled shells of electrons for energy storage and for covalence in which there is no energy exchange for them to join in a single combination with other atoms. Having a filled shell is not enough. Boron has a dimensionless core filled s shell, but the energy of these shells is not quite enough to stabilize. Similarly, one would have expected another "noble" element to make $\frac{1}{2}$ the energy of the $4s$ shell (the best ever for the $4s$ light). On the other hand, boron is not completely inert; it will form weak diatomic compounds with fluorine.

Since our simple tree-like approach of the main features of the periodic table, we stop our examination at element number 36. There are 36 elements so far numbered.

We would like to bring up only one more point. Let us not only examine the valence, but also see why something about the determining properties of the chemical bonds. Take an atom like oxygen. Oxygen has four $2s$ electrons. The first three go into the " x ", " y ", and " z " shells and the fourth will eventually fit in the $2p$ shells having between " x " and " y "—except consider how xy happens in that! Each of the two hydrogens are willing to share an electron with the oxygen, helping the oxygen to fill a shell. These electrons will tend to occupy the " x " and " y " positions. So the water molecule holds the two oxygens close making a right angle with respect to the center of the oxygen. The angle is actually 107° . We can now understand why the angle is large, almost 100° , indicating the electrons are occupying the x and y directions. The electron repulsion "shame" has made "water" molecules push the angle out to less than the covalent radius of H_2S . But because the sulfur atom is larger, the two hydrogen atoms are forced apart, there is less repulsion, and the angle is only pushed out to about 92° . Sulfur has been larger, so in H_2Se the angle is very close to 97° .

We can use the same arguments to understand the geometry of ammonia, H_3N . Nitrogen has four $2s$ plus three more " x ," " y ," and " z " type shells. The three hydrogens spend time in a right angle to each other. The angles are about 100° —again from the electron repul-

son, but at least we see why the molecule of H_2N is not flat. The angles in phosphine, H_3P , are close to 90° , and in H_3As are still closer. We assumed that H_3N was not flat when we described it as a two-state system. And the nonflatness is what makes the ammonia more polarizable. Now we see that also that shape can be inferred from our quantum mechanics.

The Schrödinger equation has been one of the great triumphs of physics. By providing the key to the underlying machinery of atoms in nature it has given an explanation for atomic spectra, for chemistry, and for the nature of matter.

Operations

20-1 Operations and operators

All the things we have seen so far in quantum mechanics could be handled with ordinary algebra, although we did from time to time ‘borrow’ some features of working quantum mechanics, quantities and equations. We would like now to look some more closely, since it is useful and useful mathematical ways of describing quantum-mechanical things. There are many ways of approaching the subject of quantum mechanics, and other books use a different approach from the one we have taken. As you go on to read other books you might not see right away the connection of what you will find in them with what we have been doing. Although we will also be able to get a few useful results, the main purpose of this chapter is to tell you about some of the different ways of writing the same physics because then you should be able to understand better what other people are saying. When people were first working out classical mechanics they always wrote all the equations in terms of x , y , and z components. Then someone decided to work with just one vector, but all of the writing would be made much simpler by inventing the vector calculus. This is true, but when you come down to figuring something out, you often have to convert the vector back to their components. But it's generally much easier to just start writing out what you work with vectors and also easier to do many of the calculations. In quantum mechanics we want this to write many things in a simpler way by using the basis of the quantum system. The wave vector $|\psi\rangle$ has, of course, according to the wave mechanics vectors in three dimensions but is an abstract symbol that stands for a physical state, identified by the ‘label’, or ‘name’, ψ . The label is used because the laws of quantum mechanics can be written as algebraic equations in terms of these symbols. For instance, the fundamental law that any state can be made up from a linear combination of other states is written as

$$|\psi\rangle = \sum_i C_i |i\rangle, \quad (20.1)$$

where the C_i are real or complex numbers and anything $|i\rangle = |i_1 i_2 i_3\rangle$ means $|1\rangle |2\rangle |3\rangle$, and so on. I stand for the three digits in some base, or object, whatever.

If you take some physical state and do something to it, like rotating it, or idle waiting for the time Δt , you get a different state. We say ‘performing an operation on a state produces a new state.’ We can express the same idea by an equation:

$$|x\rangle = A |\psi\rangle. \quad (20.2)$$

An operation makes something another state. The operator stands for some particular operation. When the operation is performed on the state, say, $|\psi\rangle$, ... produces some state $|x\rangle$.

What does Eq. (20.2) mean? We think it this way. If you multiply the equation by $\langle x|$ and expand $|\psi\rangle$ according to Eq. (20.1), you get

$$\langle x|\psi\rangle = \sum_i C_i \langle x|i\rangle. \quad (20.3)$$

(The states $|i\rangle$ are from the same basis as $|\psi\rangle$). This is now just an algebraic equation. The numbers $\langle x|i\rangle$ give the measure of each basis state you will find in $|x\rangle$, and this is given in terms of a linear superposition of basis amplitudes (C_i) that sum fine

20-1 Operators and operations

20-2 Average energies

20-3 The average energy of an atom

20-4 The position operator

20-5 The momentum operator

20-6 Angular momentum

20-7 The change of averages with time

$|\psi\rangle$ is each base state. The numbers a_1, a_2, \dots are just the coefficients which tell how much of $|\psi\rangle$ goes in each sum. The operator A is referred to as *mixed* by the set of numbers, or "coefficients".

$$A_{ij} = a_i \cdot A_j |\psi\rangle \quad (20.2)$$

So Eq. (20.2) is a slight class way of writing Eq. (20.3). Actually it is a little more than that: something more is implied. In Eq. (20.3) we did not make any reference to a set $\{j\}$; however, implicitly, all j is an image of Eq. (20.3) in terms of some set of base states. But if you know, you may take any set you want. And this idea is implied in Eq. (20.2). The operator way of writing avoids making any particular choice. Of course, when you want to specify, you have infinite sets. When you take your choice, you use Eq. (20.2). So the operator equation (20.3) is a more abstract way of writing the algebraic equation (20.3). It's similar to Fermi's way of *overwriting*.

$$c = a \propto b$$

instead of

$$c_1 = a_1 b_1 - a_2 b_2$$

$$c_2 = a_2 b_1 + a_1 b_2$$

$$c_3 = a_3 b_1 - a_4 b_2$$

The "new way" is much handier. The convenient results however, you will eventually have to give the components with respect to some set of bases. Similarly, if you want to be able to say what you really mean by A , you will have to be ready to give the matrix A_{ij} , i.e. basis of some set of base states. So long as you don't do it, most sensible. E.g., Eq. (20.2) may be just the same as Fig. (20.3). (You should remember also that *unless* you know a matrix for this particular set of base states you can always calculate the corresponding matrix that goes with any other basis. You can re-form the matrix from the "projection operator" in another.)

The operator version in (20.2) is to allow a new way of thinking. If we imagine some operator A , we can use it with any states $|\psi\rangle$ to create a new state $A|\psi\rangle$. Sometimes a "state" we get. It's won't be very familiar; it may not represent any physical situation we are likely to encounter in nature. (For instance, we may get a state that is not normalized to represent one electron.) In other words, we may at times get "states" that don't make much sense. Such artificial "states" may not be useful, perhaps to the point of being irrelevant.

We have already shown you many examples of quantum-mechanical operators. We have the rotation operator $R(\theta)$ which takes a state $|\psi\rangle$ and produces a new state, which is the old state as seen in a rotated coordinate system. We have had the parity operator P , which makes a new state by reversing x -coordinates. We have had the operators σ_x , σ_y , and σ_z for spin one-half particles.

The operator \hat{J}_z was defined in Sec. 17 as one of the rotation operators for small angles:

$$\hat{R}(\theta) = 1 + \frac{i}{\hbar} \int d\mathbf{r} J_z \quad (20.3)$$

In just minutes, of course, that

$$\hat{R}(\theta)|\psi\rangle = |\psi\rangle + \frac{i}{\hbar} \int d\mathbf{r} J_z |\psi\rangle \quad (20.4)$$

In this expression J_z is to be found the value you get from $\sin(\theta/\hbar)$ for the small angle θ and then subtract the original value. This represents a "state" $|\psi'\rangle$ which is the difference of two states.

One more example. We can an operator \hat{p}_x , called the momentum operator (one-component defined in an equation like (20.6)). In (20.6) is the operator which does

displace a state along \hat{q} by the distance b , then \hat{q} is defined by

$$\hat{q}|\psi\rangle = b|\psi\rangle + \frac{i}{\hbar}\hat{p}|\psi\rangle \quad (20.3)$$

where b is a real displacement. Displacing the state $|\psi\rangle$ along \hat{q} by a small amount b gives a new state $|\psi'\rangle$. We can say that the new state is obtained with plus a small kick:

$$\int_b^B V_{\hat{q}}(q) dq$$

For example we are talking about work on a state vector like $|\psi\rangle$, which is an abstract description of a physical situation. They are quite different from algebraic operators which work on mathematical functions. For instance, \hat{q} is an operator that works on $|\psi\rangle$ by changing it to a new function just $\hat{q}|\psi\rangle$. Another example is the momentum operator \hat{p}^2 . You can see that the same word is used in both cases, but you should keep in mind that the two kinds of operators are different. As a measurement operator others can work on an algebraic function, our on a state vector like $|\psi\rangle$. Both kinds of operators are used in quantum mechanics and one is in other kinds of equations. As you will see in the later chapters you are free to name the object it is well to be sure the definition always remains the same. Later on, when you are more familiar with the subject, you will find that it is less important to keep the distinction between the two kinds of operators. You will usually find that they both generally require some notation for both!

We'll go on now and look at some useful things you can do with operators. But first, one special chapter. Suppose we have a state $|\psi\rangle$ whose number in some basis $|\psi\rangle = a_1|\psi_1\rangle + a_2|\psi_2\rangle$. The amplitude for a state $|\psi\rangle$ to obtain some other state $|\phi\rangle$ is $a_1 \langle \psi_1 | \phi \rangle$. Is there some relation to the complex conjugate of this amplitude? You should be able to show that

$$a_1 \langle \psi_1 | \phi \rangle^* = \langle \psi_1 | \phi^\dagger | \psi \rangle \quad (20.4)$$

where \cdot^* (read "A dagger") is an operator whose matrix elements are

$$a_\phi = \langle \phi | \psi^* \rangle \quad (20.5)$$

To get the element $\langle \phi | \psi^* \rangle$ for any $|\phi\rangle$, elements of ψ^* are given and take its complex conjugate. This implies that the state $\hat{q}^\dagger |\psi\rangle$ is in $|\psi\rangle$ the complex conjugate of the amplitude $\langle \psi | \hat{q}^\dagger | \psi \rangle$. The operator \hat{q}^\dagger is called the "Hermitian adjoint" of \hat{q} . Many important operators of quantum mechanics have the special property that when you take the Hermitian adjoint, you get the same operator back. It's called an operator. I am

$$\hat{q}^\dagger = \hat{q}$$

and it is called a "self-adjoint" or "Hermitian" operator.

20.2 Average energies

So far we have learned how to calculate expectation values. Now we would like to calculate an average value. How would you find the average energy of a system—say, in atom i ? In atom i in a particular state of definite energy, and you measure the energy, you will find a certain energy E_i . If you keep repeat the measurement on each one of a whole series of atoms, if all atoms are in the same state, all the measurements will give E_i , and the "average" to your measurement will, of course, be E_i .

Now, however, what happens if you take the atom which is state $|\psi\rangle$ which is not a stationary state? Since the system does not have a definite energy, any measurement would give energy, the same measurement on another atom in the same state would give a different energy, and so on. What would you get for the average of a whole series of energy measurement results?

We can answer the question by projecting the wavefunction onto the set of states of definite energy. To remember that this is a special reason, we'll use the symbol $\langle \psi |$. Then if the state $|\psi\rangle$ has a definite energy E_1 , in the representation,

$$\langle \psi | = \sum_i C_i |\psi_i\rangle \quad (20.13)$$

When you make an energy measurement, and get some value E_1 , you have found that the system was in the state $|\psi_1\rangle$. But you may get a different number for next measurement. Sometimes you will get E_2 , sometimes E_3 , sometimes E_4 , and so on. Let's probability that you observe the energy E_i is just the probability of finding the system in the state $|\psi_i\rangle$, which is, of course, just the magnitude of the amplitude $C_i = \langle \psi_i | \psi \rangle$. The probability of finding each of the possible energies E_i is

$$P_i = |C_i|^2 \quad (20.14)$$

How are these probabilities related to the mean value of a whole sequence of energy measurements? Let's imagine that we get a series of energy reads like this: $E_1, E_2, E_3, \dots, E_N$. The E_i 's are $E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, \dots, E_N$ and so on. We continue for, say, a thousand measurements. When we've finished we add all the energies and divide by one thousand. That's what we mean by the average. That's also called the expectation value of the number. You can count up how many times you get E_1 , say n_1 , in N , and then count n_2 the number of times you get E_2 , or n_3 the E_3 , and so on. The sum of all the energies is certainly just

$$n_1 E_1 + n_2 E_2 + n_3 E_3 + \dots + n_N E_N$$

The average energy is this sum divided by the total number of measurements which is just the sum of all the n_i 's, which we can call N :

$$\langle E \rangle = \frac{1}{N} \sum_i n_i E_i \quad (20.15)$$

We are almost there. When we replace the probability of something happening by just the number of times we expect it to happen divided by the total number of tries, the ratio n_i/N should, for large N , be very close to P_i , the probability of finding the state $|\psi_i\rangle$, although we'll never exactly P_i because of the statistical fluctuations. So it's very reasonable to replace probability by P_i and then we can say that

$$\langle E \rangle = \sum_i P_i E_i \quad (20.16)$$

The equals sign is being liberal, but that's what it means. The average value of a measured quantity A should be $\langle A \rangle$:

$$\langle A \rangle_A = \sum_i P_i A_i$$

where A_i are the various possible values of the measured quantity, and P_i is the probability of getting that value.

Let's go back to our question. Statistical state (by 1.3 average energy is

$$\langle E \rangle_E = \sum_i C_i^* E_i C_i = \sum_i C_i^* E_i C_i \quad (20.17)$$

Now since the identity $\langle \psi |$ is 1, we write the same:

$$\sum_i (C_i - \langle \psi |) C_i^* (\psi | - \langle \psi |) \quad (20.18)$$

Now we treat the left-hand $\langle \psi |$ as a constant "factor". We can then this factor out of the sum, and write it as

$$\langle \psi | \left[\sum_i (C_i - \langle \psi |) C_i^* (\psi | - \langle \psi |) \right] \psi \rangle$$

This expression has the form

$$|\psi\rangle = \sum_i c_i |i\rangle,$$

where $|i\rangle$ is some “number-up” state defined by

$$|i\rangle = \sum_j e^{iE_j t/\hbar} |j\rangle. \quad (20.16)$$

In fact, it’s otherwise, that’s why $\langle\psi|\psi\rangle$ if you expand two such states $|i\rangle$, it becomes $\langle i|i\rangle$.

Now remember what we know of the states $|i\rangle$. They are supposed to be **orthonormal** states—by which we mean that for each one

$$\delta(i,j) = \langle i|j\rangle.$$

Since δ_{ij} is just a number (an **integer**), this is equivalent to $\langle i|i\rangle = 1$, and the right in Eq. (20.16) is the same as

$$\sum_j e^{iE_j t/\hbar} |j\rangle |\psi\rangle.$$

Now compare this to the former combination and we’re led to $\langle\psi|\psi\rangle = \infty$.

$$\sum_i R(i,i) |\psi\rangle = R \sum_i c_i c_i |\psi\rangle = R |\psi\rangle.$$

Magic! Equation (20.16) is the same as

$$|\psi\rangle = R |\psi\rangle. \quad (20.17)$$

The average energy of the state $|\psi\rangle$ can be written very easily as

$$\langle E_{\text{av}} \rangle = \langle \psi | \hat{H} | \psi \rangle. \quad (20.18)$$

To get the average energy, you’d multiply $|\psi\rangle$ with \hat{H} , and then multiply by $\langle\psi|\hat{H}|\psi\rangle$. It should result

Our one formula for the average energy is not fully pretty. It’s also useful, because now we don’t need to say anything about the probabilities of two states. We don’t even have to know all of the possible energy levels. When we go to calculate, we’ll need to decompose our state in terms of states $|i\rangle$ of those states, but if we know the Hamiltonian matrix H_{ij} , for that we can use the average energy. Equation (20.18) says that for any set of basis states $|i\rangle$, the average energy can be calculated from

$$\langle E_{\text{av}} \rangle = \sum_i \delta_{ii} \langle \psi | H | \psi \rangle / \langle \psi | \psi \rangle, \quad (20.19)$$

where the components δ_{ii} ($i=0$) are just the elements of the matrix H_{ii} .

Let’s check this result. For the special case that the states $|i\rangle$ are the **atomic energy states**, for example, $H_{ii} = E_i \delta_{ii}$, or $\langle i|H|j\rangle = E_i \delta_{ij}$, and

$$\langle E_{\text{av}} \rangle = \sum_i \delta_{ii} \langle \psi | H | \psi \rangle / \langle \psi | \psi \rangle = \sum_i \delta_{ii} E_i \langle \psi | \psi \rangle,$$

which is right.

Equation (20.19) can, incidentally, be extended to other physical measurements which you can express as an operator. For instance, \hat{L}_z is the operator of the **z-component** of the **angular momentum** L . The average of the z-component for the state $|\psi\rangle$ is

$$\langle L_z \rangle_{\psi} = \langle \psi | \hat{L}_z | \psi \rangle.$$

One way to prove it is to think of some situation in which the energy is proportional to the angular momentum. Then all the arguments go through in the same way.

In summary, "if a physical observable A is related to a suitable quantum-mechanical operator \hat{A} , the average value of A for the state $|\psi\rangle$ is given by

$$\langle A \rangle_{\psi} = \langle \psi | \hat{A} | \psi \rangle. \quad (20.22)$$

This can be written as

$$A_{\psi} = \langle \psi | \hat{A} | \psi \rangle. \quad (20.23)$$

with

$$A_{\psi} = \langle \psi | \hat{A} | \psi \rangle. \quad (20.24)$$

20.3 The average energy of an atom

Suppose we want the average energy of an atom in a state described by a wave function $\psi(x)$. How do we find it? Let's first think of a one-dimensional situation with a state $|k\rangle$ defined by the amplitude or $|\psi\rangle = \psi(x)$. We are looking for the quantity $\langle \psi | \hat{H} | \psi \rangle$. This is applied to the momentum representation. Following our usual procedure, we replace the states $|\psi\rangle$ and $|\phi\rangle$ by $|\psi\rangle$ and $|\phi\rangle$, and change the sum to an integral. We get

$$\langle \psi | \hat{H} | \psi \rangle = \int \langle \psi | \psi(x) \hat{H} | \psi(x) \rangle dx. \quad (20.25)$$

This integral can, if we wish, be written in the following way:

$$\int \langle \psi | \psi(x) \hat{H} | \psi(x) \rangle dx, \quad (20.26)$$

with

$$\langle \psi | \psi \rangle = \int (\psi - \bar{\psi}) \psi^* dx. \quad (20.27)$$

The integral over ψ is nothing but the same one you had in Chapter 16—see Eq. (16.20) and Eq. (16.52)—and is equal to

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) = -\nabla^2 \psi(x).$$

We can therefore write

$$\langle \psi | \psi \rangle = \int \frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x) \psi(x) \psi^* dx. \quad (20.28)$$

Remember that $\langle \psi | \psi \rangle = \langle \psi | \psi^* \rangle = \psi^*(x)$; using this identity, the average energy in Eq. (20.27) can be written as

$$\langle \psi | \hat{H} | \psi \rangle = \int \psi^*(x) \left(-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x) \right) \psi(x) dx. \quad (20.29)$$

Given a wave function $\psi(x)$, you can get the total energy by doing this integral. You can begin to see how we can go back and forth from the state vector $|\psi\rangle$ to the wave function $\psi(x)$.

The quantity in the braces of Eq. (20.27) is an *operator*†. We will write it as \hat{E}

$$\hat{E} = -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V.$$

With the notation Eq. (20.24) becomes

$$\langle \psi | \hat{H} | \psi \rangle = \int \psi^*(x) \hat{E} \psi(x) dx. \quad (20.30)$$

The *classical* operator \hat{E} defined here is, of course, not identical to the quantum-mechanical operator \hat{H} . The new operator works on a function of position, $f(x) = \langle x | \psi \rangle$, to give a new function of x , $\hat{E}(x) = \langle x | \hat{E} | \psi \rangle$, while \hat{H}

† The “operator” \hat{E} is not “multiplied by $\psi(x)$ ”

dependence on ϕ , the scalar field. By doing so, we can state $\psi = \phi$, without mentioning the coordinate representation or any particular representation at all. Since \hat{A}^\dagger is the same as \hat{R} even in the coordinate representation, if we choose to work in the coordinate representation, we avoid inserting \hat{R} in terms of ϕ . Instead $\hat{\psi}^\dagger$, which depends on ϕ , can be written in terms of ψ and ϕ ; that is, we expect, according to Eq. (20.25), that $\hat{\psi}^\dagger$ as ψ is related to ϕ , the amplitude of ϕ at my position. On the other hand, notice that \hat{R} is a differential operator. We have already worked on, in Section 18.2, the connection between $(\hat{R}\psi)$ and the adjoint operator \hat{R}^\dagger .

We should make one qualification on our results. We have been assuming that the amplitude $\psi(x)$ — x is measured. As this is assumed, the wave has been chosen to be

$$\int \psi^*(x') dx' = 1$$

so the probability of finding the particle somewhere is unity. If you don't choose to work with a $\psi(x)$ which is not normalized, you should write

$$\langle \hat{R} \rangle_{\psi} = \frac{\int \psi^*(x) \hat{R} \psi(x)}{\int \psi^*(x) \psi(x)}, \quad (20.27)$$

It's the same thing.

Remember, however, in between Eq. (20.26) and Eq. (20.28), there are two ways of writing the successive \hat{R} -operator after you work with the ψ -representation. You can either use first form to the second wave royal, which is a local operator, where local operator is one which acts along!

$$\int \psi(x) \hat{R}(x') \psi(x') dx'$$

or in writing as it is in, where \hat{R} is a differential algebraic operator. There are, however, operators for which this is not true. For them you must work with the basic equations, i.e. (20.21) and (20.22).

You can easily extend the derivation to three dimensions. The result is just

$$\langle \hat{R} \rangle_{\psi} = \int d^3x \hat{R}(x) \psi(x), \quad (20.29)$$

and

$$\hat{R} = -\frac{\partial^2}{\partial x^2} Y^2 + V(x), \quad (20.30)$$

the wave function being ψ .

$$\int \psi^* \hat{R} \psi = 1 \quad (20.31)$$

The last equations can be extended to systems with several electrons in a fairly obvious way, but we won't bother to write down the results.

With Eq. (20.29) we can calculate the average energy of an atomic system even without knowing its energy levels. All we need is the wave function. It's an important idea. We'll tell you about a few interesting applications. Suppose you want to know the ground-state energy of some atom—say the helium atom, for instance. It's too hard to solve Schrödinger's equation for the wave function, because there are too many variables. Suppose, however, that you take a guess at the wave function—pick one function you like—and calculate the average energy. That is, you use Eq. (20.29)—generalize to three dimensions—to find what the average energy would be if the atom were really in the state described by this wave function. This energy will certainly be higher than the ground-state energy, which is the lowest

* We write $\psi(x)$ for the element of ψ at point x , but $\psi(x)$ is also the whole function $\psi(x)$ in x and t these coordinates.

possible energy the atom can have.) Now pick another function and calculate its average energy. If it is lower than your first choice you are getting closer to the true ground-state energy. If you keep on trying all sorts of different states you will be able to get lower and lower energies, which were closer and closer to the ground-state energy. If you're clever, you will try some functions which have a few adjustable parameters. When you calculate the energy it will be expressed in terms of these numbers. By varying the parameters to give the lowest enealile energy, you are trying out a whole range of functions of course. Eventually you will find that it is harder and harder to get lower energies and you will begin to suspect that you are finally close to the lowest possible energy. The parameters have been adusted in this way—not by solving a differential equation, but by making up a specific function with a lot of adjustable parameters which are eventually chosen to give the lowest possible value for the average energy.

20.4 The position operator

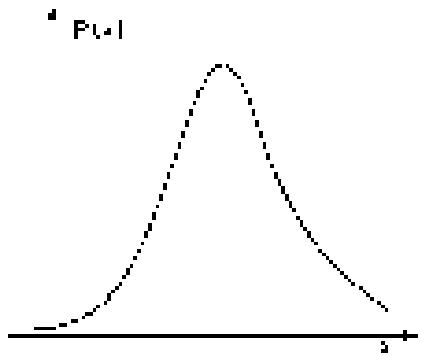


Fig. 20.1. A curve of probability density representing a localized particle

What's the average value of the position of an electron in an atom? For any particular state ψ what is the average value of the coordinate x ? Well it works in one dimension only if you extend the idea to three dimensions in 3D systems with more than one particle. We have a gauge quantum mechanics problem and we keep re-solving it over and over again. What is the average? It is

$$\int x P(x) dx,$$

where $P(x)$ is the probability of finding the electron in a little element dx . Suppose the probability density $P(x)$ looks like x as shown in Fig. 20.1. The electron is most likely to be found near the peak of the curve. The average value of x is also somewhere near the peak. It is, in fact, just the centre of gravity of the area under the curve.

We have seen earlier that $P(x) \propto |\psi(x)|^2$, $|\psi(x)|^2 = \psi^*(x)\psi(x)$, so we can write the average of x as

$$\langle x \rangle_{\psi} = \int x |\psi(x)|^2 dx \quad (20.31)$$

Our equation for $\langle x \rangle_{\psi}$ has the same form as Eq. (20.30). For the average energy, the energy operator lies somewhere between the identity operator and the position operator. (If you wish you can consider x to be the algebraic operator "imparted by x ".) We can carry the calculation still further, expressing the average position in a form which corresponds to Eq. (20.30). So suppose we are to write

$$\langle x \rangle_{\psi} = \langle \psi | x | \psi \rangle \quad (20.32)$$

with

$$|a\rangle = |x-a\rangle \quad (20.33)$$

and then see if we can find an operator A which generates the state $|a\rangle$, which will make Eq. (20.32) agree with Eq. (20.30). That is, we must find A such that

$$\langle x-a \rangle = \langle \psi | A | \psi \rangle = \int x - a |\psi(x)|^2 dx \quad (20.34)$$

First, let's expand ψ in its wave representation. It is

$$\psi |x\rangle = \int \psi(x) |x\rangle dx \quad (20.35)$$

Now compare the integrals in the last two equations. You see that in the wave representation

$$\langle x-a \rangle = \langle \psi | A | \psi \rangle \quad (20.36)$$

³ You may also look at it this way. Any function (that is, wavefunction) you choose can be written as a linear combination of the basis states which are definite energy levels. Since in this combination there is a mixture of independent states with the lowest energy states, the average energy will be higher than the ground-state energy.

Operating on $|\psi\rangle$ with \hat{x} to get $\langle\hat{x}\rangle$ is equivalent to multiplying $\psi(x) = \langle x|\psi\rangle$ by a factor of $\hat{x} = x|x\rangle$. We have a definition of \hat{x} in the coordinate representation:

[We have not bothered to try to get the x representation of the matrix of the operator \hat{x} . I leave it as an exercise you can try to show that]

$$\langle x' | \hat{x} | x \rangle = x' \delta(x - x') \quad (20.14)$$

You can then work out the coupling result that

$$\hat{x} |x\rangle = x |x\rangle \quad (20.15)$$

The operator \hat{x}^2 has the interesting property that when it works on the basis states $|x\rangle$, it is equivalent to multiplying by x^2 :

Do you want to know the average value of x^2 ? It is

$$\langle x^2 \rangle_{\psi} = \int x^2 \psi(x)^2 \delta(x) dx. \quad (20.16)$$

Or, if you prefer you can write

$$\langle x^2 \rangle_{\psi} = \langle \hat{x}^2 | \psi \rangle$$

where

$$|\hat{x}^2| = \hat{x}^2 |\psi\rangle. \quad (20.17)$$

By \hat{x}^2 we mean $\hat{x}\hat{x}$, the two operators are used one after the other. With the second form you can calculate $\langle x^2 \rangle_{\psi}$ using the representation (does anyone you what? If you want the average of x^n or of any polynomial in x , you can see how to get it).

20-5 The momentum operator

Now we would like to calculate the mean momentum of an electron again. We'll stick to one dimension – x . We might as well probability p (a measurement will give a momentum between p and $p + dp$). Then

$$\langle p_{\psi} \rangle = \int p F(p) d\psi. \quad (20.18)$$

Now we let $\langle p | \psi \rangle$ be the amplitude that the state $|\psi\rangle$ is in a definite momentum state $|p\rangle$. This is the same amplitude we called $\langle \text{pos} | p | \psi \rangle$ in Section 16.1 and is a function $\psi(p)$ just as $\langle x | \psi \rangle$ is a function of x . Let's go ahead to normalize the amplitude so that

$$F(p) = \sum_{\psi} \langle p | \psi \rangle^2. \quad (20.19)$$

We have then

$$\langle p | \psi \rangle = \left(\frac{1}{2\pi\hbar} \delta(p - k) \right) \frac{d\psi}{dk}. \quad (20.20)$$

This form is quite similar to what we had for $\langle x | \psi \rangle$.

If we want we can do exactly the same game we did with $\langle x | \psi \rangle$. First, we can write the integral above as

$$\int (p - k) \delta(p - k) \frac{d\psi}{dk}. \quad (20.21)$$

You should now recognize this equation as just the re-written form of the amplitude $\langle \psi | \psi \rangle$, expressed in terms of the new states of definite momentum. From Eq.

16.10 (in the notes we have that $\langle \psi | \psi \rangle = \langle \psi | \psi \rangle^*$). You see the "state ψ " has $\langle \psi |$, because the multiplier ψ is not in $\langle \psi | \psi \rangle$; it is a number which is different for each value of $|k\rangle$. It is not part of the construction of the state $|\psi\rangle$. See Eq. 16.10.

With the state $|\psi\rangle$ we define a ψ -representation by

$$\langle \psi | \phi \rangle = \phi_{\psi} |\psi\rangle \quad (20.47)$$

That is, we can now write

$$(\psi)_{\psi} = \langle \psi | \phi \rangle \quad (20.48)$$

$$\psi | \phi \rangle = \phi_{\psi} |\psi\rangle \quad (20.49)$$

Since the operator ϕ is defined in terms of the ψ -representation by Eq. (20.47),

Eq. (20.49) you can if you wish show that the matrix form of ϕ is

$$\langle \phi' | \phi | \phi \rangle = \phi' \phi^* \phi = \phi'^*, \quad (20.50)$$

and thus

$$\phi' \phi^* = \phi | \phi' \rangle. \quad (20.51)$$

It would not be smart to do this.

Now comes an interesting question. We can write ϕ_{ψ} , as we have done in Eqs. (20.48) and (20.49), and we know the meaning of the operator ϕ in the ψ -representation. But how specific are we about ϕ in the absolute representation? That is what we will need to know. One has some sort of function of ϕ , and we want to compute its average in quantum mechanics. Let's consider what we mean. I'll start by saying that $\langle \phi \rangle_{\psi}$ is given by Eq. (20.48), or in other words, it is equal to the expectation value of ϕ given in Eq. (20.49). If we are given the absolute value of the state $|\psi\rangle$ and the amplitude $\langle \psi | \phi \rangle$, which is an algebraic fraction of the momentum p , we can get $\langle \phi \rangle_{\psi}$ from Eq. (20.49) and proceed to evaluate the integral. The question is, what do we do if we are given a description of the state in the ψ -representation, namely the wave function $\psi(x) = \langle x | \psi \rangle$?

Well, let's start by expanding Eq. (20.48) in the ψ -representation. In a

$$\langle \psi | \phi \rangle = \int \phi(x) \psi(x) dx. \quad (20.52)$$

Now, however, we need to know what the state $|\psi\rangle$ is in the ψ -representation. If we can find it, we can carry out the integral. So our problem is to find the function ϕ_{ψ} of $x = p - \theta$.

We can find it in the following way. In Section 16-2 we saw how $\langle p | n \rangle$ was related to $\langle x | \phi \rangle$. According to Eq. (16-14),

$$\langle x | \phi \rangle = \int e^{-ipx/\hbar} \phi(x) | x \rangle dx. \quad (20.53)$$

If we know $\langle p | n \rangle$ we can use this equation for $\langle x | \phi \rangle$. Now we want, of course, to express ϕ_{ψ} in absolute form somehow in terms of $\langle x | \phi \rangle$, or $\langle x | \psi \rangle$, which we are given, and to do so we can. Start by writing Eq. (20.47) and then, using Eq. (1.6.24) to write it

$$\langle \psi | \phi \rangle = \langle \psi | \phi | x \rangle = \int e^{-ipx/\hbar} \phi(x) | x \rangle dx. \quad (20.54)$$

Since the integral is over x we can put the x outside the integral and write

$$\langle \psi | \phi \rangle = \int e^{-ipx/\hbar} \phi(x) | x \rangle dx. \quad (20.55)$$

Compare this with (20.53). You could say that ϕ is equal to $\phi(x)$. But that the wave function $\langle x | \phi \rangle = \phi(x)$ can depend only on x , not on y . That's the whole problem.

However, some ingenuity below discovered that the integral in (20.55) could be interpreted by parts. The derivative $d e^{-ipx/\hbar} / dx$ with respect to x is $-ip/\hbar e^{-ipx/\hbar}$, so the integral in (20.55) is equivalent to

$$= \frac{h}{i} \int_{-\infty}^{+\infty} \frac{d}{dx} (e^{-ipx/\hbar} \phi(x)) dx$$

If we integrate by parts we have since

$$= \frac{\hbar^2}{2} e^{-\imath kx^2} \psi(x) - \frac{\hbar}{i} \int e^{-\imath kx^2} \frac{\partial \psi}{\partial x} dx.$$

So if we consider a constant state $\psi = \psi_0$ with $\psi'(x)$ equal to zero, i.e. $\psi = \psi_0$, the derivative is zero and we have

$$\langle \psi_0 | \hat{H} \rangle = \frac{\hbar^2}{2} \int e^{-\imath kx^2} \frac{\partial \psi_0}{\partial x} dx. \quad (20.55)$$

Now compare this result with Eq. (20.54). You see that

$$\langle \psi_0 | \hat{H} \rangle = \frac{\hbar}{i} \frac{\partial}{\partial x} \langle \psi_0 | \hat{p} \rangle. \quad (20.56)$$

We have the necessary piece to be able to complete Eq. (20.52). The answer is

$$\langle \psi_0 | \hat{H} \rangle = \frac{1}{2} \langle \psi_0 | \hat{p} | \hat{p} \rangle + \frac{\hbar^2}{2} \int \psi_0(x) dx. \quad (20.57)$$

We have found that Eq. (20.57) looks like the coordinate representation.

Now we should begin to see a "natural" path to developing. When we account for the average energy of a state ψ we find it was

$$\langle \hat{E}_{\psi} \rangle = \langle \psi | \hat{E}_{\psi} | \psi \rangle \text{ with } \langle \hat{E}_{\psi} \rangle = \hat{p}^2 / 2m.$$

The same thing is written in the coordinates would be

$$\langle \hat{E}_{\psi} \rangle = \frac{1}{2} \langle \psi | \hat{p}^2 | \psi \rangle \text{ with } \langle \hat{p}^2 \rangle = \hat{p}^2 \langle \psi | \psi \rangle.$$

Here \hat{p}^2 is an algebraic operator which works as a function of x . When we take the average value of x , we found that it could also be written

$$\langle \hat{x}_{\psi} \rangle = \langle \psi | \hat{x}_{\psi} | \psi \rangle \text{ with } \langle \hat{x} \rangle = \hat{x} \langle \psi | \psi \rangle.$$

In the equivalent way the corresponding equation are

$$\langle \hat{x}_{\psi} \rangle = \int x^2 \langle \psi | \psi(x) \rangle dx \text{ with } \langle x \rangle = \langle \psi | \hat{x} | \psi \rangle.$$

When we looked about the average value of \hat{p} , we wrote

$$\langle \hat{p}_{\psi} \rangle = \langle \psi | \hat{p}_{\psi} | \psi \rangle \text{ with } \langle \hat{p} \rangle = \hat{p} \langle \psi | \psi \rangle.$$

In the coordinates words the equivalent equations were

$$\langle \hat{p}_{\psi} \rangle = \int p_{\psi}(x) \delta(x) dx \text{ with } \langle \hat{p}_{\psi} \rangle = \frac{\hbar}{i} \frac{d}{dx} \langle \psi | \psi \rangle.$$

In view of our first example we start with the state ψ and produce another (hyperoperator) with by a "coordinate-representation" operator. In the coordinate representation the generator has corresponding wave function by operating on the wave function $\psi(x)$ with an algebraic operator. There are the following approximate correspondences for one-dimensional problems:

$$\begin{aligned} \hat{H} &= \hat{p}^2 = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} = F(x), \\ x &\mapsto x, \\ \hat{p}_x &\mapsto \hat{p}_x = \frac{\hbar}{i} \frac{\partial}{\partial x}. \end{aligned} \quad (20.58)$$

Table 20.1

Physical Quantity	Operator	Coordinate Form
Position	\hat{r}	$\hat{r}_i = -\frac{\hbar^2}{2m} \nabla^2 + m\omega_i^2$
	\hat{x}	-
	\hat{y}	-
	\hat{z}	-
Momentum	\hat{p}_i	$\hat{p}_i = \frac{\hbar}{i} \frac{\partial}{\partial x_i}$
	\hat{p}_x	$\hat{p}_x = \frac{\hbar}{i} \frac{\partial}{\partial x}$
	\hat{p}_y	$\hat{p}_y = \frac{\hbar}{i} \frac{\partial}{\partial y}$

In this form our coordinate components are for the *general* space dimension, i.e.

$$\hat{p}_i = \frac{\hbar}{i} \frac{\partial}{\partial x_i}, \quad (20.30)$$

and we have inserted the x subscript i to remind you that we have been working only with the x -component of momentum.

You can readily extend the results to three dimensions. For the other components of the momentum,

$$\hat{p}_x = \hat{p}_i = \frac{\hbar}{i} \frac{\partial}{\partial x}$$

$$\hat{p}_y = \hat{p}_i = \frac{\hbar}{i} \frac{\partial}{\partial y}$$

If you want, you can even think of an operator of the *absolute* value and write

$$\hat{p} \rightarrow \hat{p} = \frac{\hbar}{i} \left(\hat{p}_x \frac{\partial}{\partial x} + \hat{p}_y \frac{\partial}{\partial y} + \hat{p}_z \frac{\partial}{\partial z} \right),$$

where \hat{p}_x, \hat{p}_y , and \hat{p}_z all work in the three dimensions. It looks a bit more elegant if we write

$$\hat{p} \rightarrow \hat{p} = \frac{\hbar}{i} \nabla. \quad (20.31)$$

Our general results for at least some spin-independent operators, *Cartesian coordinates*, are given by the operators in the coordinate representation. We summarize our results so far, extended to three dimensions, in Table 20.1. For each operator we list the corresponding formulae.

$$\hat{p}_i = \frac{\hbar}{i} \nabla_i \quad (20.32)$$

or

$$\hat{p}(r) = \hat{p}_i(r). \quad (20.33)$$

We will now give a few illustrations to the use of these ideas. The first one is just to point out the relation between \hat{p} and \hat{k} . If we use \hat{p} , we get

$$\hat{p}_i \hat{k}_i = -\hbar^2 \frac{\nabla^2}{m^2}.$$

In many books the same symbol is used for \hat{p} and \hat{k} , because they both stand for the same physical quantity, except it is convenient not to have to write different kinds of letters. You may usually find which one is intended by the context.

This means that we can write the equality

$$\hat{H} = \frac{1}{2m} \hat{p}_x^2 + \hat{p}_y^2 + \hat{p}_z^2 + V(r).$$

Or, using the vector notation,

$$\hat{H} = \frac{1}{2m} \hat{\vec{p}} \cdot \hat{\vec{p}} + V(r). \quad (20.64)$$

(It's an algebraic operator; any term without the operator symbol ($\hat{}$) means just a straight multiplication.) This equation is nice because it's easy to remember if you haven't forgotten your classical physics. Everyone knows that the energy is (approximately) just the kinetic energy $p^2/2m$ plus the potential energy, and \hat{H} is the operator of the total energy.

This result has impressed people so much that they try to teach students about classical physics before quantum mechanics. (We think differently.) But such parallels are often misleading. For one thing, when you have operators, the order of various factors is important; but that is not true for the factors in a classical equation.

In Chapter 17 we defined an operator A , in terms of the displacement operator δ , by [see Eq. (17.27)]

$$|\psi\rangle = A(x)|\phi\rangle = \left(1 + \frac{i}{\hbar} \Delta x\right)|\phi\rangle \quad (20.65)$$

where Δx is a small displacement. We should show you that this is equivalent to our new definition. According to what we have just worked out, this quantity should mean the same as

$$\delta(x) = \phi(x) + \frac{\partial \phi}{\partial x} x.$$

But the right-hand side is just the Taylor expansion of $\phi(x - \Delta x)$, which is certainly what you get if you displace the state to the left by Δx (or shift the coordinates to the right by the same amount). Our two definitions of δ agree!

Let's use this fact to show something else. Suppose we have a bunch of particles which we label 1, 2, 3, ..., in some complicated system. (To keep things simple we'll stick to one dimension.) The wave function describing the state is a function of all the coordinates x_1, x_2, x_3, \dots . We can write it as $\psi(x_1, x_2, x_3, \dots)$. Now displace the system (to the left) by Δ . The new wave function

$$\psi(x_1, x_2, x_3, \dots) = \psi(x_1 + \Delta, x_2 + \Delta, x_3 + \Delta, \dots)$$

can be written as

$$\begin{aligned} \psi(x_1, x_2, x_3, \dots) &= \phi(x_1, x_2, x_3, \dots) \\ &= \left[\phi \left(\frac{\partial \phi}{\partial x_1} + \frac{\partial \phi}{\partial x_2} + \frac{\partial \phi}{\partial x_3} + \dots \right) \right]. \end{aligned} \quad (20.66)$$

According to Eq. (20.64) the operator of the momentum of the state $|\psi\rangle$ (we call it the total momentum) is equal to

$$\hat{p}_{\text{total}} = \frac{1}{i} \left[\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2} + \frac{\partial}{\partial x_3} + \dots \right].$$

But this is just the sum of

$$\hat{p}_{\text{total}} = \hat{p}_{x_1} + \hat{p}_{x_2} + \hat{p}_{x_3} + \dots. \quad (20.67)$$

The operation of momentum obey the rule that the total momentum is the sum of the momenta of all the parts. Everything hangs together nicely, and many of the things we have been saying are consistent with each other.

20-6 Antitheta notations

Let's go for fun back at another approach—the operator approach using $\hat{\theta}$ —momentum. In Chapter 1 we defined its operator $\hat{\theta}$ in terms of \hat{p} by the operator of a rotation by the angle θ around the \hat{x} -axis. We can also have a state ψ associated directly by a single wave function $\psi(\hat{x})$, which is a function of coordinate only, and does not have any reference to the fact that the electron may have its spin either up or down. That is, we want to be able to do apart from our \hat{x} -spin interpretation and think about only one orbital part. To keep the interpretation clear, we'll call the operator $\hat{\theta}$ and define it in terms of the rotation as a rotation by an infinitesimal angle $d\theta$ by

$$\hat{\theta}_x(\hat{x}, \psi) = \left(1 + \frac{i}{\hbar} \epsilon \hat{L}_x \right) \psi.$$

(Remember, this definition applies only to a state ψ which has no internal spin variables, but depends only on the coordinates $x = x_1, x_2, x_3$.) If we look at the state ψ' in a new coordinate system rotated about the \hat{x} -axis by the small angle ϵ , we have a new state

$$\psi' = \hat{\theta}_x(\hat{x}, \psi).$$

If we choose to describe the state ψ' in the coordinate representation, that is, by no wave function ψ' , we would expect to be able to write

$$\psi'(\hat{x}) = \left(1 + \frac{i}{\hbar} \epsilon \hat{L}_x \right) \psi(\hat{x}). \quad (20.65)$$

What is \hat{L}_x ? Well, again, that's a little bit of the old quantum mechanics assumed x_1 and p_1 , but we will drop the prime now formally as $x = x_1$ and $p = p_1$, so you can see in Eq. 20-2, since the strength for the electron to be at y is $S(y)$, denoted by the rotation of the coordinates we can write

$$P(x, y, z) = \delta(x + C_1 y - S_1 z) = \delta(x, y) = \delta\left(\frac{y}{S_1} - \epsilon \frac{\partial}{\partial x}\right)$$

(remembering C_1 is a constant). This means that

$$\hat{L}_x = \frac{\hbar}{i} \left(x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} \right). \quad (20.66)$$

That's not nice. But notice it is equivalent to

$$\hat{L}_x = \hat{p}_y + \epsilon \hat{p}_x. \quad (20.67)$$

Remembering our commutation relations, we can write

$$\hat{L}_x = \hat{p}_y - \epsilon \hat{p}_x. \quad (20.70)$$

This formula is easy to remember now, since it looks like the familiar form of classical mechanics; this is the x component of

$$\hat{L}_x = \hat{r} \times \hat{p}_x. \quad (20.71)$$

One of the two parts of this operator is the vector \hat{r} . They're equal, except they're not equal over the x -constant classical form. Which ones don't? There had better be some that don't come out right, because if everything p did, then there would be nothing different about quantum mechanics. There would be no new physics. Heisenberg's equation which is different. In classical physics

$$\hat{p}_x - \epsilon \hat{p}_x = 0,$$

What is the quantum mechanical?

$$\hat{p}_x - \epsilon \hat{p}_x = \hbar,$$

Let's work it out in the representation. So first we'll know what we are doing we just assume wave function $\psi(x)$. We have

$$i\partial_x \psi(x) = \hbar \omega \psi(x)$$

or

$$i \frac{\hbar}{\imath} \frac{\partial}{\partial x} \psi(x) = \frac{\hbar}{\imath} \frac{\partial}{\partial x} \psi(x)$$

Re-arrange now with the derivatives operator on everything to the right. We get

$$i \frac{\hbar}{\imath} \frac{\partial}{\partial x} + \frac{\hbar}{\imath} \psi(x) = i \frac{\hbar}{\imath} \frac{\partial \psi}{\partial x} = i \hbar \omega \psi(x) \quad (20.19)$$

The answer is now clear. The whole equation is equivalent simply to multiplication by $-i\hbar\omega$.

$$(i\hbar \omega - f(x)) \psi = -\frac{\hbar^2}{2m} \psi. \quad (20.20)$$

If Plank's constant were zero, the classical and quantum results would be the same, and there would be no quantum mechanics to be had!

Incidentally, if any two operators A and B , when taken together, act thus:

$$AB = BA$$

then we say that "the operators A and B commute." And an equation such as (20.20) is called a "commutation rule." You can see that the commutation rule for θ and φ is

$$\theta\varphi - \varphi\theta = 0.$$

There is another very important commutation rule that has to do with angular momenta. It is

$$L_x L_y = L_y L_x - i \hbar I_z. \quad (20.21)$$

You can get some practice with θ and φ up to me by proving it for yourself.

It is interesting to notice that operators which do not commute also occur in classical physics. We have all ready seen this when we have talked about rotation in space. If you rotate something, such as a horse, by 90° around x and then 90° around y , you get something different from rotating first by 90° around y and then by 90° around x . It is, in fact, just this property of space that is responsible for Eq. (20.21).

20.7 The change of strength with time

Now we want to show you something else. How an average changes with time. Suppose for the moment that we have an operator A , which does not itself have time in it in any obvious way. We have an operator like $\theta(x, t)$. (We exclude things like, say, the potential of some external potential the wave being varied with time, such as $V(x, t)$). Now suppose we take the average, in some time $\langle \cdot \rangle_t$, which is

$$\langle A \rangle_t = \langle \psi | \hat{A} | \psi \rangle_t. \quad (20.22)$$

How $\langle A \rangle_t$ depends on time? Why should it? One reason might be that the operator itself depends explicitly on time. For instance, if it had a time-dependent time-varying potential like $V(x, t)$. But even if the operator does not depend on t , say, for example, the operator $\hat{A} = \hat{x} + \hat{y}$, the time averaging may depend on time. Certainly the average position of a particle could be moving. How does such a motion come out of Eq. (20.22) if \hat{A} has no time dependence? Well, the state $|\psi\rangle$ might be changing with time. For nonstationary states we have often shown a time-dependence explicitly by writing a state as $|\chi(t)\rangle$. We want to show that the rate of change of $\langle A \rangle_t$ is given by a new operator we will call \dot{A} . Recall that \dot{A} is an operator, so that perhaps a dot over the A does not look clear to you,

the time-dependent \hat{A} , but we just a way of writing a new operator \tilde{A} which is defined by

$$\frac{d}{dt} \langle \hat{A} \rangle_{\psi_0} = \langle \hat{B} \cdot \tilde{A} \rangle_{\psi_0}. \quad (20.77)$$

Our problem is to find the operator \tilde{A} .

First, we know that the rate of change of a state is given by the Hamiltonian. Specifically,

$$i\hbar \frac{d}{dt} \langle \psi(t) \rangle = \hat{H} \langle \psi(t) \rangle. \quad (20.78)$$

This is just the direct way of writing our original definition of the “imposition”.

$$\frac{d}{dt} \frac{\partial C}{\partial t} = \sum_k H_k C_k. \quad (20.79)$$

If we take the complex conjugate of the equation, it is equivalent to

$$-i\hbar \frac{d}{dt} \langle \psi(t) \rangle^* = \langle \psi(t) | \hat{H} | \psi(t) \rangle. \quad (20.80)$$

Now, we will compare if we take the derivative $\langle \psi(t) | \hat{H} | \psi(t) \rangle$ instead of $\langle \hat{H} | \psi(t) \rangle$. Since $\langle \psi(t) | \hat{H} | \psi(t) \rangle$ depends on t , we have

$$\frac{d}{dt} \langle \hat{H} \rangle_{\psi_0} = \left(\frac{d}{dt} \langle \psi(t) | \right) \hat{H} \langle \psi(t) | + \langle \psi(t) | \hat{H} \left(\frac{d}{dt} \langle \psi(t) | \right). \quad (20.81)$$

Finally, using the two equations in (20.79) and (20.81) to replace the derivatives we get

$$\frac{d}{dt} \langle \hat{H} \rangle_{\psi_0} = \frac{i}{\hbar} \langle \psi(t) | \hat{H} | \psi(t) \rangle - \langle \psi(t) | \hat{H}' | \psi(t) \rangle.$$

This equation is the same as

$$\frac{d}{dt} \langle \hat{H} \rangle_{\psi_0} = \frac{i}{\hbar} \langle \psi(t) | \hat{H} \hat{A} - \hat{A} \hat{H} | \psi(t) \rangle.$$

Comparing this equation with Eq. (20.77), you see that

$$\hat{A} = \frac{i}{\hbar} (\hat{H} \hat{A} - \hat{A} \hat{H}). \quad (20.82)$$

This is an interesting observation, and it is true for any operator \hat{A} .

Technically, “the operator \hat{A} does not modify the time-dependent wavefunction ψ_0 ”

$$\hat{A} = \frac{i}{\hbar} (\hat{H} \hat{A} - \hat{A} \hat{H}) = \frac{d \hat{A}}{dt}. \quad (20.83)$$

Let us try out Eq. (20.83) on some example to see whether it really makes sense. For instance, what operator corresponds to \hat{x} ? We can find it:

$$\hat{x} = \frac{i}{\hbar} (\hat{H} \hat{x} - \hat{x} \hat{H}). \quad (20.84)$$

What is this? One way to do the work is to work it through in the coordinate representation using the algebraic approach to \hat{x} . In this approach, you’ll commutate \hat{x} :

$$\hat{x}\hat{x} - \hat{x}\hat{x} = \left[\frac{\hbar^2}{m} \frac{d^2}{dx^2} + V(x) \right] \hat{x} - \left[\frac{\hbar^2}{m} \frac{d^2}{dx^2} + V(x) \right] \hat{x}.$$

If you operate with this on any wave function $\psi(x)$ and work out all of the d/dx terms where you can, you’ll end up after a lot of work with

$$\frac{\hbar^2}{m} \frac{d\hat{x}^2}{dx^2}.$$

But C is just the unit cell

$$C = \frac{\hbar}{m} \partial_{\theta} \psi$$

so we find that

$$\partial_{\theta} \psi + p\theta = -i \frac{\hbar}{m} \partial_{\theta} \psi \quad (20.56)$$

or C61

$$\psi = \frac{A}{\theta}. \quad (20.57)$$

A priori result. It means that if the mean value of ω is changing with time (result of the center of gravity is not zero), the mean momentum divided by m , **Energy-like classical mechanics**

Another result. What's the rate of change of the average momentum of a system? Same goes. The operator is

$$\hat{A} = \frac{i}{\hbar} (\partial_{\theta} \psi - \partial_{\theta} \psi). \quad (20.57)$$

Again you can write it out in the x -representation. Remember that ψ becomes $A(x)$. And this means that you will be taking the derivative of the potential energy V (in Eq. 22)—but only in the second term. That means that it is the only term which does not vanish, and you find that

$$\hat{A}\hat{p} = \hat{p}\hat{A} = -i\hbar \frac{\partial V}{\partial x}$$

or C62

$$\hat{p} = -i\hbar \frac{\partial}{\partial x} \quad (20.58)$$

Again the classical result. The right-hand side is the one we have derived Newton's law! But remember: these are the laws of the **system** which give the **energy equations**. They do not determine what happens in detail inside ψ alone.

Quantum mechanics has the essential difference that ψ is not equal to ψ_0 . These differences will be the small quantum \hbar . But the whole world quantum dynamics or interference waves, and all, result from the fact that $\hat{p} \neq 0$: it is not equal zero.

The history of this idea is also interesting. Within a period of a few months in 1925, Heisenberg and Schrödinger independently found correct laws to describe atomic mechanics. Schrödinger invented the wave function ψ (that's how he called his equation). Heisenberg, on the other hand, found that ψ must be described by classical equations, except that $p\psi = \hat{p}\psi$ should be equal to $\hbar A$, which he could make happen by defining them in terms of classical lines of curves. In fact, Heisenberg was using the **matrix representation**, with its matrices. Both Heisenberg's matrix algebra and Schrödinger's differential equation explained the hydrogen atom. A few months later Schrödinger was able to show that the two theories were equivalent, as we have seen before. But the two different mathematical forms of quantum mechanics were discovered independently.

The Schrödinger Equation in a Classical Context: A Reviewer on Superconductivity

21-1 Schrödinger's equation in a magnetic field

This lecture is only by invitation. I would like to give one lecture in a semester or two to discuss it so that, if you're part of the course —the course—that it is not supposed to be a last minute effort to teach you something new. But, rather, I imagine that I'm giving a summary or research report on superconductivity, the advanced audience, to complete what have already been available in quantum mechanics. The main difference between a seminar and a regular lecture is that the seminar speaker does not carry out all the steps, he just suggests. He says, "If you do such and such, this is what comes out," instead of showing all of the details. So in this lecture I'll describe the ideas of the work along, but just give you the results of the computations. You won't see them, you're not supposed to understand everything immediately, but believe (here or now) that they would come out, if you went through the steps.

All right again, this is a subject I used to talk about — it's even a modern one, and will be a very necessary topic to give in a research seminar. My subject is the Schrödinger equation in a classical setting — the case of superconductivity.

Primarily the wave function which appears in the Schrödinger equation applies to only one or two particles. And the wave function itself is not something that has a classical meaning — unlike the electric field, or the vector potential, or things of that kind. The wave function for a single particle is a "field" in the sense that it is a function of position — but it does not generally have classical significance. Nevertheless, there are some situations in which a quantum mechanical wave function does have classical significance, and they are the ones I would like to talk about. The peculiar quantum-mechanical behavior of the fermion at small scale doesn't always itself follow a large scale except in a global way that it produces Newton's laws — the laws of the so-called classical mechanics. But there are certain situations in which the peculiarities of quantum mechanics can become in a "global" way on a large scale.

At low temperatures, when the energy of a system has been reduced very, very low, instead of a large number of states being involved, only a very, very small number of states near the ground state are involved. Under these conditions the quantum-mechanical character of that ground state can appear on a macroscopic scale. It is the purpose of this lecture to show a connection between quantum mechanics and large-scale effects — the rough derivation of the very familiar quantum-mechanical Meissner effect, for example, but a special situation in which quantum mechanics will produce its own characteristic effects on a large-scale "macroscopic" level.

I will begin by reminding you of some of the properties of the Schrödinger equation. I want to discuss the behavior of particles in a magnetic field using the Schrödinger equation, because the superconducting phenomena are involved with magnetic fields. An external magnetic field is described by a vector potential, and the problem is: what are the laws of quantum mechanics in a vector potential? The principle that describes the behavior of quantum mechanics in a vector potential is very simple: The amplitude that a particle goes from one place to another along a certain path when there is a field present is the same as the ampli-

21-1 Schrödinger's equation in a magnetic field

21-2 The equation of continuity & probability

21-3 Two kinds of symmetry

21-4 The meaning of the wave function

21-5 Superconductivity

21-6 The Meissner effect

21-7 Flux quantization

21-8 The dynamics of superconductivity

21-9 The Josephson junction

[†] I'm not really promising you, because I haven't chosen any one of these arguments — maybe I'll remember one of this summer.

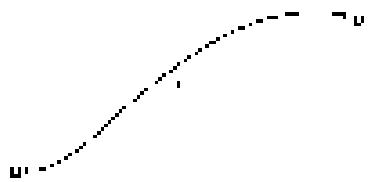


Fig. 21. The potential energy from (21.1) along the path from the origin to the point $r = R$.

more “trap-well” goes along the same route when there’s no field, multiplied by the magnitude of the line integral of the vector potential, i.e. the electric charge divided by Planck’s constant (see Chap. 11-12):

$$E_{\text{trap-well}} = \frac{q}{h} \cdot v_{\text{max}} \cdot \exp \left[\frac{q}{h} \int_{R_0}^R A \cdot dr \right]. \quad (2.11)$$

This is a short version of quantum mechanics.

Now without the vector potential the Schrödinger equation of a charged particle from classical mechanics is:

$$-\frac{\hbar^2}{2m} \nabla^2 \psi - qV = E_{\text{trap}} \left(\frac{1}{r} \cdot V \right) \left(\frac{\hbar^2}{m} \nabla^2 \psi \right) + qV. \quad (2.12)$$

Let us now include the potential V that goes to the infinite energy. Equation (21.1) is equivalent to the gradient term in a magnetized field the gradient of the Hamiltonian has reduced in value because it is a constant minus V , so that Eq. (21.12) becomes

$$\frac{\hbar^2}{m} \nabla^2 \psi = E_{\text{trap}} - \frac{q}{2m} \left(\frac{1}{r} \cdot V - qA \right) \left(\frac{\hbar^2}{m} \nabla^2 \psi + qV \right) + qV. \quad (2.13)$$

This is the Schrödinger equation for a particle moving in a constant magnetic field A , a problem we have not yet solved.

It is also not difficult to solve by analogy to a simple example in which (21.13) of classical mechanics is situation we have a limit of validity (the Schrödinger equation). We have a “resonance” — like an electron in an atom, but outside, in an “outer shell” where there is no binding force according to Eq. (21.12) if there is no vector potential (in the $\lambda = \infty$ case), i.e. the amplitude of current will be altered (so what it was before) by a factor exp $i\omega t$, the argument with $i\omega t$ times the wave number k integrated from zero to the res. Frequency ω will affect (Eq. 21.12). The effect is not at general depend on ω , so the amplitude to find the electron in the outer shell potential is called $C(\omega) = C_0$, the factor of change of that amplitude e is given by the following expression:

$$\begin{aligned} \frac{\partial}{\partial \omega} C(\omega) &= E_F C(\omega) - E_F e^{i\omega t} \exp(i\omega t) \psi = 0 \\ &= E_F e^{i\omega t} \exp(i\omega t) (\psi - e\psi) \end{aligned} \quad (2.14)$$

The two homogeneous. First, since wave energy E_F of the electron is fixed and E_F is constant, this gives the condition $C(\omega) = 0$. Next, there is the term $= E_F (e^{i\omega t} - 1)$ which is the amplitude for the electron to take unperturbed trajectory step from atom “ $n = 1$ ” to another $n = 1$. However, in a magnetized vector potential, the phase of the amplitude must be shifted according to the rule in Fig. 1.1.1. i.e. if the electron is moving clockwise in one atomic orbit, the magnetic field A turns the magnetic moment m by 180° . Since the electron is pumping back and forth, it is clear this phase shift $w = \pi$ must be given by the total phase. In the same manner there is certain amplitude to have jumps from the other side, but this time we need the vector potential at a distance $(k/2)$ to the outer wave function (because the wave k). That is exactly the reason $\Gamma = \pi \omega / \hbar$ is the capacity for the amplitude to be at $\omega = 0$ vector potential.

Now we know that in the function $C(\omega)$ growth and goes to the long wavelength limit, and if we set the mass get closer together, i.e. $\rightarrow 0$ (in approximation of classical mechanics) the solution will split. So the next step is to expand the values of (2.14) in powers of \hbar (assuming $\hbar \ll m \text{ cm}^{-1}$). For example, if $\hbar = 10^{-30}$ the right-hand side is $\sim (E_F - 10^3 E_F)(\psi)$ to the second approximation.

¹ Chapter 11, Section 13-5

By the horizon and with consideration of a free space itself

As the same argument, but was $\omega = \pi/4$ in the transition is taken away with no magnetic field, see Chapter 18.

the energy to $E_0 = \hbar\omega$. We remove the terms in \hat{p} for because the two exponentials have no more than odd even powers of $\hbar\hat{p}/m\omega$. So if you make a Taylor expansion of $C(x)$, of $C^*(x)$, and of the exponential's, and then collect the terms in $\hbar^2/m\omega^2$ you get

$$\begin{aligned} \frac{\hbar}{\hbar}\frac{dC(x)}{dx} &= E_C(x) - 2E_0/\omega \\ &= \hbar\omega^2(C^*(x) - 2p^2(x)C(x) - p^2(x)C^*(x) - f^2(x)C(x)). \end{aligned} \quad (21.5)$$

(The “ \hbar/\hbar ” here in differentiation with respect to x)

Now the final combination of terms looks quite complicated. But mathematically it is exactly the same as

$$-\frac{\hbar}{\hbar}\frac{dC(x)}{dx} = (E_0 - 2K(x)) - \hbar^2\left[\frac{\partial}{\partial x} - Q(x)\right]\left[\frac{\partial}{\partial x} + R(x)\right]C(x). \quad (21.6)$$

The second term operating on $C(x)$ gives $C'(x) + \hbar^2Q(x)C(x)$. The first bracket multiplying on these two terms give the $K''(x)$ term and terms in the last derivative of $f(x)$ and the first derivative of $C(x)$. Now remember, $C(x)$, the solutions to ψ , are magnetic so it represent a particle with an effect we have in a given b

$$E\psi = \frac{\hbar^2}{m}\psi.$$

If you then set $b_0 = -im\omega$, and put back $p(x) = ip/\hbar$, you can easily check that Eq.(21.6) is the same as the 2nd part of Eq.(21.3). (The origin of the potential energy term is well known, so I haven't bothered to include it in this discussion.) The propagation of ψ , or in fact non-relativistic quantum theory, all the symmetries by the exponential factor is the same as the rule that the momentum operator, ψ gets replaced by

$$\frac{\hbar}{\hbar}\psi = \psi,$$

as you see in the Schrödinger equation of (21.3).

21-1 The equation of continuity for probabilities

Now I turn to a second point. Another important part of the Schrödinger equation for a single particle is the idea that the probability to find the particle at a position is given by the absolute square of the wave function. It is also characteristic of the quantum mechanics that probability is conserved in a local sense. When the probability of finding the electron somewhere decreases, while the probability of the electron being elsewhere increases (keeping the total probability unchanged), something must be going on in between. In other words, the electron has a continuity in the sense that if the probability decreases at one place and builds up at another, there must be some kind of flow between. If you can't wall, for example, in this way, it will have an infinite add the probability will not be the same. So the conservation of probability there is not the empirical statement of the conservation law, just as the conservation of energy alone is not as deep and important as the local conservation of energy.⁴ If energy is dispersed into the same, how flow of energy to compensate. In the same way, we would like to find a “current” of probability such that if there is any change in the probability density (the probability of being found in a unit volume), it can be considered as coming from an inflow or an outflow due to some current. This current would be a wave which could be represented this way. On a coordinate would be the net probability per second and per unit area for a particle passes in the direction across a plane parallel to the $y-z$ plane. Positive x -axis is considered a positive flow, and passing in the opposite direction a negative flow.

⁴ Session 13.1.

⁵ Volume II, Section 27.1.

Is there such a current? Well, you know that the probability density $P(x,t)$ is given in terms of the wave function by

$$P(x,t) = \psi^*(x,t)\psi(x,t) \quad (21.2)$$

Is it according to theory a current J such that

$$\frac{\partial P}{\partial t} = -\nabla \cdot J \quad (21.3)$$

With the considerations of Eq. (21.2), I get two terms

$$\frac{\partial P}{\partial t} = \psi^* \frac{\partial \psi}{\partial t} + \psi \frac{\partial \psi^*}{\partial t}. \quad (21.4)$$

Now use the Schrödinger equation, Eq. (21.3); the ψ 's, and take the complex conjugate of it to get $\psi^*\psi^*$, and I get the following result. You get

$$\begin{aligned} \frac{\partial P}{\partial t} &= -\frac{i}{\hbar} \nabla \cdot \frac{1}{2m} \left(\frac{\hbar^2}{c} \nabla - i\epsilon \right) \cdot \left(\frac{c}{\hbar} \nabla - i\epsilon \right) \psi + c \exp \psi \\ &+ \frac{e}{2mc} \left(\frac{c}{\hbar} \nabla + i\epsilon \right) \cdot \left(\frac{\hbar^2}{c} \nabla + i\epsilon \right) \psi - c \exp \psi \end{aligned} \quad (21.5)$$

The potential ψ contains a lot of other stuff around in it. And it is the cut that makes it difficult indeed to express it as a perfect convergence. The whole equation is equivalent to

$$\frac{dP}{dt} = -\nabla \cdot \left[\frac{1}{2mc} \psi^* \left(\frac{\hbar^2}{c} \nabla + i\epsilon \right) \psi + e \left(-\frac{\hbar^2}{c} \nabla - i\epsilon \right) \psi^* \right]. \quad (21.6)$$

It is really not so cumbersome as it seems. It is a symmetrized combination of ψ^* times a certain operator on ψ , plus ψ^* times the complex conjugate operator on ψ . It is some quantity J as its own complex conjugate. In other words it reduces to a spin to be. The upper term has no importance this way. It is just the momentum operator \hbar measured. I could write the current in Eq. (21.6) as

$$J = \frac{1}{2} \left[\left\langle \hat{p} - \frac{e\mathbf{A}}{c} \right\rangle^* \psi + \psi^* \left\langle \hat{p} - \frac{e\mathbf{A}}{c} \right\rangle \psi \right] \quad (21.7)$$

There is then a current J whose components Eq. (21.5).

Equation (21.10) shows that the probability is conserved locally. If a particle disappears from one region it cannot appear in another without conserving probability between. Imagine that our disc region is surrounded by a closed surface for enough cut. But then is zero probability to find the electron at the surface. The total probability to find the electron somewhere within the surface is the volume integral of it. By applying the volume integral theorem the volume integral of the derivative of J is equal to the surface integral of J . If J is zero at the surface Eq. (21.10) says that J is zero, so the probability to find the particle incident along. Only if some of the probability approaches the boundary can come out from it. We can say that it only goes out by passing through the surface and that is local conservation.

21-3. Two kinds of momentum

The equation for the current is rather interesting, and sometimes causes a certain amount of worry. You would think the current would be something like the density of particles times the velocity. The density should be something like $\psi\psi^*$, which is not. Another use in Eq. (21.10) calls for the typical formula the average value of the operator

$$\langle \hat{p} \rangle = \frac{1}{it} \langle \psi | \hat{p} | \psi \rangle. \quad (21.8)$$

So maybe we should think of it as an identity of time. In other words, though we don't have a suggestion for addition of velocity to momentum, because we would also think that momenta are divided by mass, \vec{p}_m , should be a velocity. The two probabilities given by the source potential:

1. It happens that these two possibly true were also discovered in classical physics, even if it was found that momentum could be defined in two ways. One of the is called "kinetic momentum," and the other is called "inertial." Well in this case it is the "inertial momentum." This is the momentum obtained by multiplying mass by velocity. The other is a more fundamental, more abstract momentum, which goes called the "dynamical momentum," which I'll call "p-momentum." The two possibilities are:

$$p_{\text{momentum}} = m v \quad (21.14)$$

$$p_{\text{inertial}} = mv - qA. \quad (21.15)$$

In quantum theory, it is again an identity with magnetic field \vec{B} with parameter q , which is connected to the greatest operator \hat{p} , so it follows that (21.14) is the operator of \vec{v} -velocity.

I'd like to make a brief digression. To show you what this is all about, whether there must be something like Eq. (21.15) in the quantum mechanics. The wave function changes with time according to the Schrödinger equation ... Eq. (21.7). If I would suddenly change the vector potential, the wave function wouldn't change in the first instance only because of changing energy. Now think of what would happen in the following circumstance. Suppose I have a long solenoid, of which I can produce a flux of magnetic field parallel to itself, as shown in Fig. 21-2. And then I change \vec{B} by suddenly turning it off. Suppose this flux nearly instantaneously goes from something ... I start with zero and suddenly, and then I turn off a vector potential. This means that I produce suddenly a discontinuity in \vec{A} . You remember that the line integral of \vec{A} around a loop is the same as the flux of \vec{B} through the loop.¹ Now what happens? I will only turn on a vector potential. According to the quantum mechanical equation the sudden change of \vec{A} does not cause a sudden change of \vec{p} , the wave function is still the same. So the packet is always unchanged.

But remember what happens electrically when I suddenly turn off a flux. During the short time that the flux is going, there's an electric field generated whose line integral is the rate of change of the flux with time:

$$\vec{E} = -\frac{\partial \vec{A}}{\partial t}. \quad (21.16)$$

That electric field is continuous. The flux is changing rapidly, and it gives a finite \vec{E} to the particle. The \vec{E} here is the change from the electric field and suddenly we could up of course not permit \vec{E} to be a real variable. Just as \vec{A} changes it might well be \vec{qA} . In other words, if you suddenly turn off a vector potential at a charge, its charge immediately picks up an "old" momentum equal to $-q\vec{A}$, but there is something that isn't changing immediately and that's the difference between \vec{p} and \vec{qA} . And so the only $\vec{p} = m\vec{v} + q\vec{A}$ concerning which is not changed when you make a sudden change in the vector potential. This $q\vec{A}$ is what we have called the "p-momentum" and is of importance in classical mechanics and the theory of dynamics, but it is an important significance in quantum mechanics. It depends on the character of the wave function, and it is the one to be identified with the equation:

$$\phi = \frac{h}{i} \nabla$$

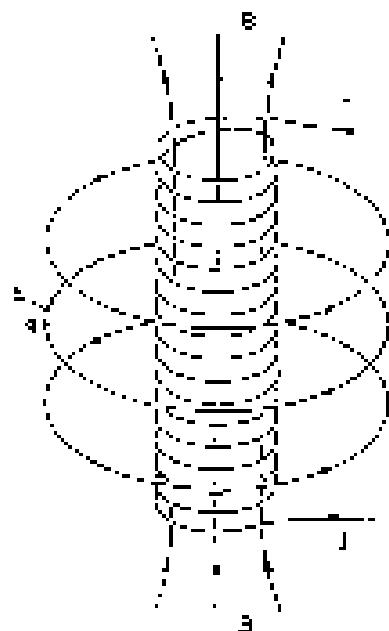


Fig. 21-2. The electric field caused by a sudden cut or increasing current.

¹ See for example, L. D. Landau and E. M. Lifshitz, *Electrodynamics, Part 1*, Pergamon Press, New York (1962), p. 181.

² Volume II, Chapter 11, Sections 11-1.

21-4 The meaning of the wave functions

When Schrödinger first discovered his equation he discovered the conservation law of $\nabla \cdot \mathbf{J}$, or, as a consequence of his equation, that he imagined necessarily that \mathbf{P} was the electric charge density of the electron and that J was the electric current density, so that though \mathbf{J} is an electromagnetic current, electromagnetism felt charges like currents. When he solved his equations for the hydrogen atom and calculated ψ , he didn't calculate the probability of finding the wave in a particular location, but from the interpretation he was completely confident. The charge density was stationary but there were magnetic moments; the charges P and currents J would switch off at electrostatical fields and the atom would radiate light. He even found, making a number of assumptions that, if this work was quite right, it was at this point too, that much more important contributions to our ideas regarding quantum mechanics. It was shown that one could go far as to know interpreted the ψ 's of Schrödinger's equations in terms of a probability amplitude—that very difficult idea is, the square of the amplitude is not the charge density, but is only the probability per unit volume of finding an electron there, and that when you divided the electron over a place the charge density is $\rho = \int \psi^2 d\tau$, where $d\tau$ is due to \mathbf{r} .

The wave function $\psi(r)$ for an electron in an atom does not, can, describe a particle—but what does it? a stationary charge density. The electron is either here or there or somewhere else, but whatever it is, it is a point charge. On the other hand, think of a situation in which there are an enormous number of particles in exactly the same state, a very large number of them will occupy the same wave function. That means that if the ψ is here and one of them is there, and the probability ρ of finding one out of them at a given place is proportional to $|\psi|^2$.

Or since there are so many particles, if it flows in any volume $d\tau$ object, we generally find a number close to zero density. So is a situation in which in the wave function for each of the enormous number of ψ 's there which is called the same state, ρ may be interpreted as the density of particles. If, under these circumstances, each particle carries the same charge, we can, in fact, go further and interpret ρ as the density of electrons. Naturally, ρ is over the dimensions of a probability density, so that it should be multiplied by $d\tau$ to give the intensities of a charge density. For our present purposes we can ignore the normal factor $1/\sqrt{d\tau}$ and take ρ as the electron charge density. With this understanding, J (the current of probability) has calculated becomes exactly the electric current density.

So in the situation in which you can have very many particles in exactly the same state, there is possible a new physical interpretation of the wave functions. The charge density and the electric current can be calculated directly from the wave functions and the wave function can take on a physical meaning which includes classical, macroscopic situations.

Something similar can happen with photons. When we have the wave function of a single photon, it is the amplitude to find a photon somewhere although we haven't ever seen it there, there is an equivalent for the theory wave function corresponding to the Schrödinger equation for the electron. The photon equation is just the same as Maxwell's equations for the electromagnetic field, and the wave function is the same as the scalar potential A . The wave function must be ψ (just the vector potential). The quantum physics in the wave theory is the classical physics because photons are noninteracting mass particles and many of them can be in the same state, as you know. They also can be in the same state. The comment that you have photons in the same state is that in the one electromagnetic theory, you can move in the wave function, which is the scalar potential, directly. Of course, it worked historically the other way. The first calculations were made directly with an amplitude in the same state, in fact associated to it either the current or the force for a single photon by comparing directly with observations on a macroscopic level the square of the wave function.

Now the picture with the electron is that you cannot pull time the ψ in the same state. Therefore, it was long believed that the wave function of the

Schrodinger equation, would never have a true one-particle approach for anything more microscopic than a hydrogen atom. The microscopic representation of the amplitude for atoms. On the other hand, it is now realized that the **standard** of superconductivity refers to a **one-particle** theory.

21-5 Superconductivity

As you know, we ordinary metals become superconducting below a certain temperature, the superconduc^ttive temperature is different for different metals. Some years ago the lowest temperature at which any metal becomes electrically without any resistance. This phenomenon has been observed for a very large number of metals but not for all of them. The reason for its absence has caused a great deal of difficulty. I used to try to understand what was going on in the crystal structure, and I will only describe enough of it for the present purpose. It has a fault due to the irregularities of the arrangement with the vibrations of the atoms in the lattice, here is a small net vibration even between the electrons. They ought to hold the electrons held together, if they speak very curiously and suddenly, bound pairs.

Now you know that a single electron is a Fermi particle. This is the first pair weak, not as a three-particle, because two electrons have paired charge, because it is twice "unpaired," and the mass is only slightly increasing, 4 parts to three per MeV .

The energy of pairing—that is, the attraction energy—is very, very weak. Only at low temperatures is needed to allow the electrons split by thermal agitation, and remain bound back to "unpair" electrons. But when you take the temperature sufficiently low, but they have to be in very regular get into the same very weak state; even may be called in pairs.

I don't wish you to imagine that the pairs are always held together very closely like a point pair etc. As a matter of fact, one of the most difficult problems in understanding the phenomena originally was that there is no one single pair. The two electrons which form the pair can easily spread over a considerable distance, and their relative velocity, which is directly smaller than the size of a μm , several pairs are occupying the same space at the same time. That's why electrons in a metal form pairs and in forming of the energy given up to binding can have been a problem of recent times. This problem, which is the theory of superconductivity was first explained by the class of Bardeen, Cooper and Schrieffer. There is an interesting historical note with every known pair, the idea that the electrons do, in some manner or other, were in place, then we can think of these pairs as being more or less like molecules and that's where we find talk about the wave function for a pair.

Now the Schrodinger equation for the wave function looks like (21.1). There will be one difference in the way the charge of the electron. Also, we don't know the wave—*i.e.* effective mass—for the Co in the solid lattice, so we don't know what number to put in front. We should see that this is going to be very high frequency, for a or something like, to be exactly right. Since the single energy that corresponds to very rapidly varying wave functions that is to say, to break up the pair, at high temperatures, there are always a few pairs which are broken up according to the usual Boltzmann theory. The probability that a pair is broken is proportional to $e^{-E_{\text{Co}}/kT}$. The electrons will be not bound in pairs and other "normal" electrons and will move around in the crystal in the ordinary way. I will, however, want to end my discussion at essentially zero temperature. At any low enough energy of the compass one pair will be broken by these electrons which are not in pairs.

* First discovered by Onnes in 1911. H. K. Onnes, Comm. Roy. Lab., Leiden, Proc. Kon. Akad. Wet. (Amsterdam), Vol. 45, p. 731 (1911). You will find a more detailed discussion of the subject in E. A. Lee, *Superconductivity*, John Wiley and Sons, Inc., New York, 1967.

¹ J. R. Allen, L. M. Cooper, and J. R. Schrieffer, Phys. Rev. 108, 1125 (1957).

Since electrons go in one direction, when there are a lot of them, they will want there is an especially large amplitude for other paths to go in the same direction. So nearly all of the paths will be locked down at the lowest energy, in which the wave function won't be easy to get away from the other paths. That's more angle to go into the same path than it was, quantified also by the Fermi factor, which reflects on the oscillatory nature of the lowest state. So we would expect all the paths to be moving in the same way.

What does this mean for the very last term, $\langle \psi | \hat{p} | \psi \rangle$, the wave function of a particle in the lowest energy state. However, since \hat{p} is going to be proportional to the charge density ρ , I can just as well write $\langle \psi | \hat{p} | \psi \rangle$ as the current density of the superconductor times some physical factor:

$$\langle \psi | \hat{p} | \psi \rangle = \rho \langle \psi | \hat{v} | \psi \rangle^2, \quad (21.17)$$

where ρ and \hat{v} are real functions of x . They complex functions, of course, so we'll do this work. It's clear what we mean when we talk about the charge density, but what is the physical meaning of the product of the wave function? Well, let's see what happens if we substitute $\langle \psi | \hat{v} | \psi \rangle$ Eq. (21.12), and express the current density in terms of basic variables v and A . It's just a lot of numbers and I won't go through it—the algebra, but it's nontrivial.

$$J = -\frac{\hbar}{m} \left(v_3 - \frac{e}{c} A_3 \right) \rho. \quad (21.18)$$

Since both the current density and the charge density have a direct physical meaning, it's the source due to the motion speed both ρ and \hat{v} are real things. The phase is just ignored because it is a phase of the current density J . The situation probably not observable, until the greatest of the phases known everywhere, the phase is known except, for a constant. And can define the phase of the particle, and then its phase contribution is definitely.

Incidentally, the expression for the current has an analytical form, since, when you do it, the current density J is in fact the charge density times the velocity. Current of the fluid of electrons, in fact. Equation (21.17) is then equivalent to

$$J\rho = \rho \hat{v} \hat{v} - \hat{v} \rho. \quad (21.19)$$

Notice that there are two pieces in the current density, one from \hat{v} and one from the vector potential, and the other is coming from from the behavior of the wave function. In other words, the quantity \hat{v} is just what we need to take into account.

21-6 The Meissner effect

Now we can state the sense of the phenomena of superconductivity. First, there is no electrical resistance. There's no resistance because all the electrons are collectively in the same state. In the ordinary flow of current you break one electron or the other out of the regular system, gradually deteriorating the general momentum. But how to get out electrical energy from what is the other is stable is very hard because of the tendency of all these particles to stay in the same state. A current, unfortunately, you always going forward.

It's also easy to understand that if you have a piece of metal in the superconducting state and then put a magnetic field which isn't too strong (We won't go into the details of how strong), the magnetic field can't penetrate the metal. If as you built up the magnetic field, any of it were to build up inside the metal, there would be a flow of charge of flux which would produce an electric field, and unless the field would immediately generate a current itself, by Faraday's law, would oppose the flux. Since all the electrons will move together, and this current also the field will generate enough current to oppose completely any applied magnetic field. So if you turn the field on after you've cooled a metal to the superconducting state, it will be excluded.

Electromagnetic screening is a related phenomenon discussed experimentally by Meissner.⁸ If you heat a piece of the metal to a high temperature so that it is a normal conductor and establish a magnetic field through it, and then you lower the temperature below the critical temperature (below the metal's transition temperature), the field is expelled. In other words, it stores up its own energy—and if you heat it up again until it just reaches the transition, the field exits.

We can see the reason for this in the equations and I'd like to explain how. Suppose that we take a piece of superconducting material which is at zero temperature. Then in a steady situation there would be divergence of the current must be zero because there's no place for it to go. It is easier to do this in terms of the divergence of the electric current. I should explain why choosing the current over charge density makes any loss of generality but I don't want to take the time. Taking the divergence of Eq. (21.28), then gives that the divergence of \mathbf{J} is equal to zero. This is correct. What about the variation of μ^2 ? If you consider a superconductor, there is a regular limit of zero for charge density due to the cosmic law of the force. If the charge density ρ is uniform there is no net electric field in the metal. If there would be an accumulation of charge in one region, the charge wouldn't be free to move and there would have to be negative polarization charge density $\pm \delta$. So in reality it turns out the charge density of the electrons in the superconductor is almost perfectly uniform. I can take $\rho = \text{constant}$. Now the \mathbf{J} is very likely to be zero everywhere. So if one lump of metal is far from another, then \mathbf{J} is zero there. And that means that there is no contribution to J from μ non-uniform. Equation (21.28) then says that no current is generated at point A . So everywhere in a lump of superconducting material the current is necessarily zero; that's in the same potential.

$$J = -\rho \frac{\partial}{\partial r} A. \quad (21.29)$$

Since ρ and dA/dr (in some meaningful sense) are constant, then J is zero again. — (from another) that

$$J = \text{some constant}. \quad (21.30)$$

This equation was originally proposed by London and London⁹ to explain the experimental observations of superconductivity. (See before the quantum-mechanical origin of the effect, see introduction.)

Now we can use Eq. (21.29) in the equations of electromagnetism to solve for the fields. The vector potential is related to the current density by

$$\nabla^2 A + \frac{1}{c^2 \epsilon_0} J = 0. \quad (21.29)$$

If I use Eq. (21.29) for J , then

$$\nabla^2 A = \frac{1}{c^2 \epsilon_0} J. \quad (21.29)$$

where J^2 is just a new constant,

$$J^2 = \rho \frac{q^2}{m_e c^2}. \quad (21.29)$$

We can try to solve this equation for A and see what happens. I do. For example, in the situation Eq. (21.29) has constant, but $c^2 \epsilon_0$ times, of the form $e^{-r/a}$ and $a \ll \lambda$. These solutions mean that the vector potential must decrease exponentially as you go from the surface into the material. (It can't increase

⁸ W. Meissner and R. Ochsenfeld, Naturwiss. 20, 261 (1933).

⁹ H. London and F. London, Proc. Roy. Soc. (London) A149, 1 (1934); Physica 2, 341 (1935).

Actually if the electric field were too strong, plus would be broken up and the "fermion" electron is created to reduce it to help stabilize any excess of positive charge still in there trying to make the system more stable. So the plus point is that a nearly uniform density is rapidly turned non-uniformly.

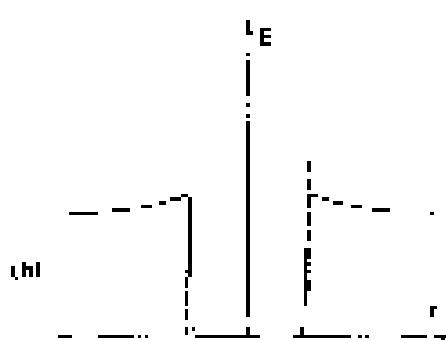
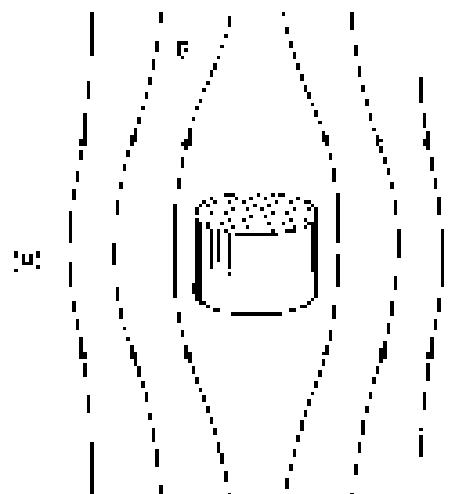


Fig. 21-2(a) (left) A superconducting cylindrical ring (noted in Fig. 21-1b), the resistance $R = 0$. (Right) A sketch of it.

resistance there would be a flow and if the Alice current is very large or equal to 1.5, the field only penetrates about 1/3 of the surface, leaving about 2/3 of the ring free of field, as sketched in Fig. 21-3. This is the explanation of the Meissner effect.

How big is the distance a^2 ? Well, remember that $\mu_0 I$, the total magnetic current in the ring, $1.3 \times 10^{-13} \text{ A}$, is given by

$$\mu_0 I = \frac{q^2}{4\pi R^2},$$

Also, remember that q in Eq. (21-24) is twice the charge on an electron, so

$$\frac{q}{2eR^2} = \frac{\mu_0 I}{4\pi},$$

which gives qR^2 , where R is the radius of the ring, in cm. So we have

$$R^2 = \frac{q^2}{8\pi\mu_0 I}, \quad (21-25)$$

For a small radius like 1 cm about $3 \times 10^{22} \text{ e}$, a current of 1.3 A and a ring with only one wound-on turn, R^2 would be about $7 \times 10^{-10} \text{ cm}^2$. This gives you the order of magnitude.

21-7 Flux quantization

The Faraday equation (21-9) is not enough to account for the anomalies of superconductivity including the Meissner effect. In recent years, however, there have been some extremely accurate predictions. One prediction made by London was so peculiar that nobody paid much attention to it until recently with new experiments. This time instead of taking a single turn, imagine we take a few turns of thickness Δz centered to the ring. Let's start with the simplest of all cases with a magnetic field B and the ring then turned so the external magnetic field is parallel to the axis of the ring. A sketch of this configuration of space is sketched in Fig. 21-4. Once again I say there will be a field in the body of the ring as indicated in part (a) of the figure. When the ring is made superconducting, the field is forced outside of the ring, as we have just seen. There will then be some flux through the body of the ring as shown in part (b). If the external field is now removed, the lines of field going through the ring are "trapped" as shown in part (c). The flux Φ through the center can't decrease because $\partial\Phi/\partial t$ has to equal to the $\partial\phi/\partial t$ integral of E around the ring which is zero in a superconductor. As the external field is removed the permanent flux is flowing through the ring to keep the flux through the ring a constant. This the old south-pole idea, it is with a twist! These currents will, however, all flow near the surface in concentric layers caused by the result of Ohm's law that I made for the static field: these currents can trap the magnetic field out of the body of the ring and part of the permanently trapped magnetic field as well.

Now, however, the ϕ is not constant in B since, and the equations involve a surprising effect. The argument is based on the fact that $\mu_0 I$ is constant in a disk with a current going around it you can see from the following argument.

Well inside the body of the ring the current density J is zero. As Eq. (21-18) gives

$$\nabla \cdot \mathbf{B} = \mu_0 J. \quad (21-26)$$

Now consider what we get if we take the line integral of \mathbf{B} around a circle C , which goes around the ring but the center of its cross-section so that it does not cross the surface, as shown in Fig. 21-5. From Eq. (21-26),

$$\oint_C \nabla \cdot \mathbf{B} d\ell = \oint_C \mu_0 J d\ell. \quad (21-27)$$

Now you know that the line integral of \mathbf{A} around any loop is equal to the flux of \mathbf{B} through the loop:

$$\oint \mathbf{A} \cdot d\mathbf{l} = \Phi.$$

Equation (21.27) now becomes

$$\oint \nabla \psi \cdot d\mathbf{l} = \oint \mathbf{A} \cdot d\mathbf{l}. \quad (21.28)$$

The line integral of a gradient from one point to another goes from point 1 to point 2 is the difference of the values of the function at the two points. Namely,

$$\int_{l_1}^{l_2} \nabla \psi \cdot dl = \psi_2 - \psi_1.$$

If we let the two end points l_1 and l_2 come together to make a closed loop you might first think that Φ would equal $\psi_1 - \psi_1$, so that the integral in Eq. (21.28) would be zero. That would be true for a closed loop in a single-connected piece of superconductor, but it is *not* necessarily true for a ring-shaped piece. The only physical measure that we can make is Φ . There can be only one value of the flux function for each point. What has happened is you have closed the ring when you just have at the starting point the ψ you get must give the same value for the new function

$$\psi = \sqrt{\Phi^2}.$$

This will happen if Φ changes by $2\pi n$, where n is any integer. So if we make one complete turn around the ring the left hand side of Eq. (21.27) must be $n \cdot 2\pi n$. Using Eq. (21.28) this gives that

$$2\pi n = \Phi. \quad (21.29)$$

The magnetic flux always has to change like $2\pi n/\psi$. If you would think of the ring as a classical object with an identity persist (that is, "lived") consistently, you would think that whatever flux was actually found through it would just stay there—any amount of flux at all could be trapped. But the quantum-mechanical theory of superconductivity says that the flux can be zero, or $2\pi/4\pi\hbar/e$, or $4\pi/4\pi\hbar/e$, or $6\pi/4\pi\hbar/e$, and so on, but no value in between. It must be a multiple of a basic quantum-mechanical unit.

London¹² postulated that the flux through a superconductor during a gap would be quantized and said that the possible values of the flux would be given by Eq. (21.29) with Φ equal to the electronic charge. According to London the basic unit of flux should be $2\pi/4\pi\hbar/e$, which is about 1.8×10^{-16} gauss \cdot cm². It should be just a flux, think of a tiny cylinder with a миллиметров diameter, the magnetic field inside it when it carries the amount of flux without any penetration of the coil's magnetic field. It should be possible to observe such a flux by a sensitive magnetometer measurement.

In 1931 such a specimen flux was looked for and found by Deaver and Fairbank¹³ at Stanford University and at about the same time by Dau and Schäfer¹⁴ in Germany.

In the experiment of Deaver and Fairbank, a tiny cylinder of superconductor was made by glistering a thin layer of tin on a one-centimeter length of No. 50 (1.3 \times 10⁻³ cm diameter) copper wire—the tin remains superconducting below 3.87K, while the copper remains a normal metal. The wire was bent in a small cylindrical magnetic field, and the temperature reduced until the tin became superconducting. Then the external source of field was removed. You would

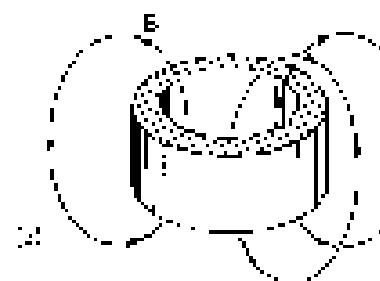
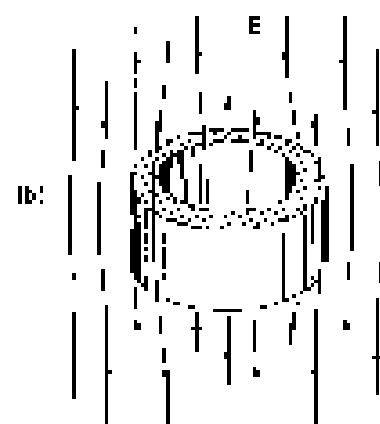
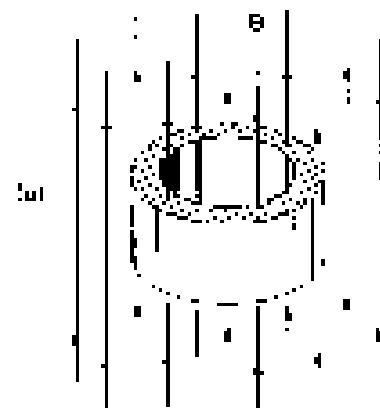


Fig. 21.4. A ring in a magnetic field. (a) in the normal state; (b) in the superconducting state; (c) after the value of field is reversed.

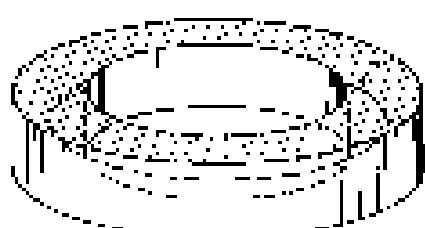


Fig. 21.5. The current in a ring in a superconducting ring.

¹² F. London, *Superfluidity*, Interscience Inc., New York, 1950, Vol. I, p. 152.

¹³ B. S. Deaver, Jr., and W. M. Fairbank, *Nature*, **189**, 740 (1931).

¹⁴ S. Dau and M. Schäfer, *Naturwissenschaften*, **31**, 11 (1940).

expect this to generate a current by Lenz's law so that the flux value would not change. The solenoid would now have magnetic moment proportional to the flux value. The magnetic moment was measured by jiggling the wire up and down like the handle of a screwing machine, but at the rate of 100 cycles per second (about a turn of cable coils at the ends of the air cylinder). The induced voltage in the coils was used a measure of the magnetic moment.

When the experiment was done by Debye and Polderman, they found that the flux was quantized, but this did not seem very remarkable since it was known that magnetism was quantized. Debye and Polderman got the same result. At first this was quite mysterious, but we now understand why it should be so. According to the Bardeev, Debye, and Semenoff theory of superconductivity, the q which appears in Eq. (21.28) is the charge of a pair of electrons and is equal to e_0 . The result is

$$q_0 = \frac{e_0 h}{c} = 1.8 \times 10^{-17} \text{ coulombs} \quad (21.30)$$

which is 11 the amount predicted by London. Everything now fits together, and the measurements show the existence of superconductivity purely quantum-mechanical effect on a large scale.

21-8 The dynamics of superconductivity

The Meissner effect and the flux quantization are two manifestations of our general theory, so far the task of completeness I would like to leave you with. The complete equations of a superconducting fluid would be from this point of view—it is rather interesting. Up to this point I have only put the expression for ψ into equations for charge density and current. If I put ψ into the complete Schrödinger equation, I get equations for ψ and ψ^* . It should be interesting to see what happens, because here we have a "fluid" of electron pairs with a charge density $\rho = e_0 n$ and a current $\mathbf{j} = \mathbf{e}_0 \nabla \psi^* \times \mathbf{B}$. We can try to see what kind of equations we get for such a "fluid". By substituting the wave function of Eq. (21.17) into the Schrödinger equation (21.3) and remembering that ϕ and ψ depend functions of x , y , and z . If we ignore the self and inductive parts we obtain these two equations. To write them in a shorter form I will be leaving Eq. (21.18) with

$$\frac{\partial}{\partial t} \nabla \psi = \frac{i}{\hbar} \mathbf{A} - \sigma \quad (21.31)$$

The first one you might get is then

$$\frac{\partial \psi}{\partial t} = \mathbf{T} \cdot \nabla \psi \quad (21.32)$$

Since ψ is first A , this is just the continuity equation again. The other equation I obtain tells now ψ how it is

$$i \frac{\partial \psi}{\partial t} = -\frac{e_0}{2} \nabla^2 \psi + \sigma \psi - \frac{e^2}{2m} \left\{ \frac{1}{\sqrt{\rho}} \nabla^2 (\sqrt{\rho}) \right\} \quad (21.33)$$

These are thoroughly familiar ψ -hydrodynamic, of which the only difference you may well recognize that is the equation of motion for an electrically charged fluid (we identify σ with the "velocity potential") except that the σ term, which is denoted by the energy of compression of the fluid, has a linear dependence on the density ρ . In any case, one expects to say that sum of charge of the quantity $\psi^* \psi$ given by ψ , A , the energy total, σ , ψ , plus a potential energy total, ϕ_0 , that is addition of ψ , containing the factor A , which we could call a "quantum mechanical energy," we've seen that inside a superconductor $\rho \rightarrow 0$, so

¹¹ It has only been suggested by Debye that this might happen (see B. London, Ref. 1), I don't know whether it has ever been tested.

uniform by the electrostatic forces, so this term can't be zero, only, or neglected in every potential approximation provided we have only one superconducting region. If we have a boundary between two superconductors for other circumstances in which the value of ϕ may change rapidly, this term can become important.

For today, when we put the function $\phi = \phi$ in the equations of hydrodynamics, I will rewrite Eq. (21.23) in a form that makes the physics more transparent by using Eq. (21.30) to express ϕ in terms of v . Taking the gradient of the whole of Eq. (21.30) and expressing $\nabla^2\phi$ in terms of A and v using (21.11), I get

$$\frac{\partial}{\partial t} \left(\frac{q}{m} \left(-\nabla \phi - \frac{\partial A}{\partial t} \right) - v \times (\nabla \times v) - (v \times \nabla) \phi \right) + \frac{q^2}{m^2} \left(\frac{1}{\sqrt{\rho}} \nabla^2 \phi / \rho \right) = 0 \quad (21.34)$$

What does this equation mean? First, consider the

$$-\nabla \phi = \frac{qA}{m} = E \quad (21.35)$$

Next, consider the first line of Eq. (21.10). I get:

$$\nabla \times E = -\frac{i}{e} \nabla \times J_i \quad (21.36)$$

Since the curl of a gradient is always zero (for $\nabla \times A$ is the magnetic field B , all the last two terms can be written as

$$\frac{q}{m} (E - v \times B).$$

Finally, you should understand that $d\phi/dt$ stands for the time derivative of the velocity of the fluid at a point. If you concentrate on a particular point, its *gradient* is the total derivative of ϕ (in, as it is + sometimes called in "fluid dynamics, the "nowhere vanishing gradient"), which is related to v by my¹³

$$\frac{dv}{dt}_{\text{along } \nabla \phi} = \frac{\partial \phi}{\partial t} - (\nabla \cdot \nabla) \phi. \quad (21.37)$$

The extra term also appears as the third term on the right side of Eq. (21.35). Taking it to the left side, I can write Eq. (21.35) in the following form:

$$m \frac{d\phi}{dt}_{\text{along } \nabla \phi} - q(E + v \times B) = \nabla \cdot \left(\frac{1}{m\rho} \nabla^2 \phi / \rho \right). \quad (21.38)$$

We also have from Eq. (21.30) that

$$\nabla \times v = -\frac{q}{m} B. \quad (21.39)$$

These two equations are the equations of motion of an *electron* in a superconductor moving in an electromagnetic field. It says that the acceleration of each particle of the fluid whose charge q comes from the ordinary source term of $E + v \times B$ plus an additional force, which is the gradient of some physical quantity, magnetopotential, a force which is not very big except at the junction between two superconductors. The second equation says that the fluid is "ideal". The word "ideal" here means that the energy of $E + v \times B$ is conserved. That means that the velocity can be expressed in terms of scalar potential. Similarly one writes that $\nabla \times v = 0$ for a conductor, but for an ideal charged fluid in a magnetic field, this is contained in Eq. (21.30).

So, for an *electron's* motion in the first two parts of a superconductor (just as the equations of motion of an electron in a dielectric charged liquid fluid, superconductivity is the same as the problem of the hydrodynamics of a charged liquid). If you want

¹³ See Volume II, Section 10-2

in order say problem about superconductor you take these equations for the field for the superconductor, Eqs. (21.32) and (21.33), and combine them with Maxwell's equations to get the fields. (The charges and currents you use depend on the finite nature of course, include the ones from the superconductor as well as from the external sources.)

The difficulty I believe that Eq. (21.34) is not quite correct but ought to have an additional term involving the density. This new term does not depend on temperature or distance from the ordinary energy associated with temperature gradient. Just as in an ordinary fluid there would be a potential energy density proportional to the square of the variation of a function, ρ_{var} , the electron density (which is, here, also proportional to the charge density of the crystal lattice). Since there will be forces proportional to the gradient of this energy, there should be another term in Eq. (21.34) of the form, $(\nabla \rho_{\text{var}} - \rho_{\text{var}})^2$. This term did not appear from my initial calculations because from the interaction between particles which I neglected by using an independent-particle approximation. If it is, however, just the force I referred to when I made the qualitative statement the electrostatic forces would tend to bring a nearly constant insulating layer around me.

21.9 The Josephson junction

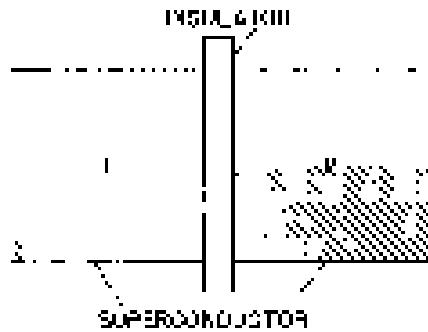


Fig. 21-6. Two superconductors connected by a thin insulator.

I would like to discuss next a very interesting situation that one method of manufacture¹⁴ while making wires might happen to join two superconductors which are separated by a thin layer of insulating material as in Fig. 21-6. Such an arrangement is now called a "Josephson junction." If the insulating layer is thick, the electrons can't go through; but if the layer is thin enough, there can be an exchange of electrons between the two superconductors. This is just another example of the quantum mechanical penetration of a barrier. Josephson analyzed this situation and discovered that a number of strange phenomena can occur.

In order to analyze such a junction I'll let the amplitude to find an electron on one side, ψ_1 , and the amplitude to find it on the other, ψ_2 . In the superconducting state the wave function ψ_1 is the same as wave function of all the electrons on one side, and ψ_2 is the corresponding function on the other side. I would do this problem for t -Fermi bands in a superconductor, but let us make a very simple situation as follows. The current is the same on both sides so that the junction is symmetrical and simple. Also, let μ_1 and μ_2 be the magnetic field. Then the two amplitudes should be related in the following way:

$$\begin{aligned} i \frac{d\psi_1}{dt} &= U_1 \psi_1 + K \psi_2, \\ i \frac{d\psi_2}{dt} &= U_2 \psi_2 + K \psi_1. \end{aligned}$$

The constant K is a characteristic of the junction. If K were zero, these two equations would just describe the lowest energy state—with energy E_0 —of an hydrogen atom—but there is coupling between the two sides by the amplitude K and there may be leakage from one side to the other. Also just the "flip-flop" amplitude of a two-state system. If the two sides are identical, U_1 would equal U_2 and I could just multiply them out. But now suppose that we connect the two superconducting regions to the two terminals of a battery so that there is a potential difference V across the junction. Then $U_1 = U_2 = qV$. Now, for convenience, let us take zero to energy to be halfway between, then the two equations are

$$\begin{aligned} i \frac{d\psi_1}{dt} - \frac{qV}{\hbar} \psi_1 &= K \psi_2, \\ i \frac{d\psi_2}{dt} - - \frac{qV}{\hbar} \psi_2 &= K \psi_1. \end{aligned} \tag{21.41}$$

¹⁴ R. D. Josephson, *Nature* Letters 2, 80 (1962).

Now we have two differential equations for two quantum mechanical states coupled together. This time, let's analyze these equations in another way. Let's make the substitutions

$$\begin{aligned} \psi_1 &= \sqrt{\rho_1} e^{i\phi_1}, \\ \psi_2 &= \sqrt{\rho_2} e^{i\phi_2}, \end{aligned} \quad (21.41)$$

where ρ_1 and ρ_2 are the masses of the two sites or the junctions and ϕ_1 and ϕ_2 are the identities of electrons at those two sites. Remember, i.e., in actual practice ρ_1 and ρ_2 are chosen exactly the same and same for ϕ_1 , the total identity of electrons in the superconducting material. Now if you substitute these equations for ψ_1 and ψ_2 into (21.39), you get two equations of equating the real and imaginary parts in reference. Letting $\theta_1 = \phi_1 - \phi_2$ for short, the result is

$$\begin{aligned} \rho_1 &= -\frac{2}{\hbar} E_V \rho_2 \rho_1 \sin \theta \\ \rho_2 &= -\frac{2}{\hbar} E_V \rho_1 \rho_2 \sin \theta \\ \theta &= +\frac{g}{\pi} \sqrt{\frac{\rho_1}{\rho_2}} \cos \theta - \frac{d\tilde{\rho}}{2\hbar} \\ \theta &= +\frac{g}{\pi} \sqrt{\frac{\rho_1}{\rho_2}} \cos \theta + \frac{d\tilde{\rho}}{2\hbar} \end{aligned} \quad (21.42)$$

The first two equations say that $\rho_1 = -\rho_2$. "But," you say, "they may both be positive, since ρ_1 and ρ_2 are both negative and equal to zero." Not quite. These equations is not the whole story. They say what ϕ_1 and ϕ_2 would be if there were no vector forces from due to the imbalance between the electron fluid and the background of positive ions. That will allow the currents to start to change, and therefore current I_1 and current I_2 would begin to flow. The current from site 1 to site 2 would be just $i_2(x - \delta_2, \omega)$.

$$I_2 = \frac{e}{\mu} \sqrt{\rho_2} i_2 \sin \theta. \quad (21.43)$$

Such a current would soon change up site 1, except that we have forgotten that the two sites are connected by wires to the rest of the circuit. The current that flows will not change \rightarrow again 2 for discrete periodicity because all sites will have the periodic component. These currents form the extra I_2 that has been included in our expression. When they are included, ρ_1 and ρ_2 are not longer charge, but the current across the junction is still given by (21.43).

Since ρ_1 and ρ_2 remain constant and equal to $\rho = kT$ set $E_V \rho_1 \Delta \theta = I_2$, and we get

$$\theta = I_2 \sin \theta. \quad (21.45)$$

$I_2 \cos \theta$ is then a constant which is characteristic of the particular junction.

The other pair of equations (21.41) is also consistent since $\theta_1 = \theta_2$. We are interested in the difference $\tilde{\theta} = \theta_2 - \theta_1$ so use Eq. (21.45) we get is

$$\tilde{\theta} = \theta_2 - \theta_1 = \frac{g^2}{\pi} \cdot \frac{\tilde{\rho}}{\rho}. \quad (21.46)$$

(21.46) means that we can write

$$\tilde{\theta}(x) = \theta_0 + \frac{g}{\pi} \int V(x) dx, \quad (21.47)$$

where θ_0 is the value of $\tilde{\theta}$ at $x = 0$. However, depending on the choice of θ_0 , we get $\tilde{\theta} = 0$. In both (21.46) and (21.47) we have the important result, the general theory of the Josephson Junction.

Now what are the consequences? First, potential voltage. If you put some dc voltage, V_0 , the current of the two junctions ($I_1 = -I_2 = V_0/R$) is a small number compared to ordinary voltage since it's passing through a diode steadily and the net current is nothing. The problem, since the bias voltage is not zero, you could say it's not due to the conduction by "normal" electrons. On the other hand if you pass over voltage, the junction you change its current! With no voltage the current can be any amount between $+I_2$ and $-I_2$ depending on the value of E_{ph} . But when you put a voltage across it and the current goes to zero, this change in behavior has been measured experimentally.¹⁸

That is another way to generate a current—by applying a voltage at a very low frequency in addition to a dc voltage, etc.

$$V = V_0 + V_{AC} \cos \omega t$$

$V_{AC} \rightarrow 0$ for $\omega \rightarrow 0$

$$I_2 = \frac{d}{dt} V_{AC} = \frac{d}{dt} V_{AC} \cos \omega t$$

Now for the small

$$\sin \omega t = \omega t \Rightarrow \sin \omega t = \omega t \cos \omega t$$

Using this approximation for $\sin \omega t$ I get

$$I = I_0 \cos \left(\phi_0 + \frac{1}{\hbar} E_{ph} \right) + \frac{V_0}{\hbar \omega} \sin \omega t \cos \left(\phi_0 + \frac{1}{\hbar} E_{ph} \right).$$

Let's first take $\omega \rightarrow 0$ at the average. In the second term I get

$$x = \frac{1}{\hbar} E_{ph}$$

The x looks like a current if we change this just a little because Chapter 3 seems to have current with a minus sign in there.

If you look up x in any subject you will find that they often write two formula for the current as

$$J = J_0 \sin \left(\phi_0 - \frac{E_{ph}}{\hbar} + \frac{1}{\hbar} \omega t - \phi \right), \quad (2.45)$$

where the integral $\rightarrow 0$ to take it across the junction. The reason for this is that there's a vector potential across the junction. The further distance is modified in such a way that we expand it out. If you do that this phase through, it comes out as given above.

Finally, I would like to describe a very dramatic and interesting experiment which has recently been made on the relationships of the variables from such a two junctions. In quantum mechanics we're used to the interaction between amplitude from two different states. Here we're going to see the interaction between two processes caused by the difference in the phase of the current of the same is through two different paths. In Fig. 21.7, I show two different junctions, "A" and "B", connected in parallel. Then the Point C is connected to a differential ammeter which measures anti-current. That is, if normal current, I_A , were to be the sum of the currents through the two junctions, $I_A = I_1 + I_2$ or the currents through the two junctions, and let their phases be ϕ_1 and ϕ_2 . Now, if a phase difference of the wave function between ϕ_1 and ϕ_2 is the same whatever you give one more to the other, along current path through junction "B", the phase difference between ϕ_1 and ϕ_2 is π , plus the last integral of the total potential along the upper diode.

$$\Delta \Phi_{B(A)}(\phi_2) = \phi_2 + \frac{e}{\hbar} \int_{\text{upper}}^{\text{lower}} A \cdot d\mathbf{r} \quad (21.46)$$

¹⁸ S. S. Antropov and J. M. Bozler, Phys. Rev. Letters 11, 1011 (1963).

¹⁹ R. Stoenescu, Rev. Roum. Phys. 10, 391 (1965).

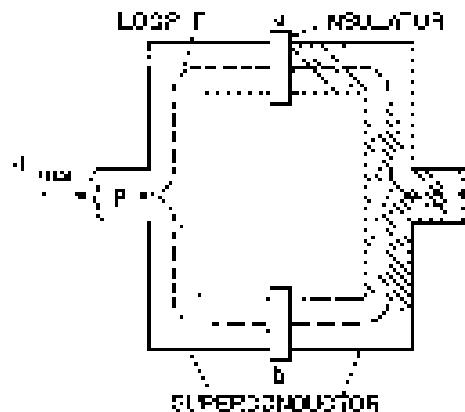


Fig. 21-7. Two-probe junctions in parallel.

Why? Because the phase is related to ϕ by Eq. (21-6). If you integrate this equation along some path, the left-hand side gives the phase change, which is then just proportional to the line integral of A , as we have written here. The phase change along each loop can then be written similarly:

$$\Delta\phi_{\text{loop}C} = \phi_C + \frac{ie}{\hbar} \int_{\text{loop}C} d\ell \cdot A_\ell. \quad (21-20)$$

There are two to be equal, and I estimate them to get the difference of the loops first at the 20 integrals, or a few π 's of the period:

$$\phi_1 - \phi_2 = \frac{2e}{\hbar} \int_{\text{loop}C} d\ell \cdot A_\ell.$$

Now the integral is around the closed loop C in Fig. 21-7. Most curves through both branches. The integral over A is the magnetic flux Φ through the loop. So the two ϕ 's are going to differ by $2\pi/\hbar$ times. In magnetic flux Φ when placed between the two branches of the circuit:

$$\phi_1 - \phi_2 = \frac{2\pi}{\hbar} \Phi. \quad (21-21)$$

I can control this phase difference by changing the magnetic field B on the circuit, and I can adjust the difference in phase and see whether it makes total current and I_c flow through the two junctions above one another of not. Let's do it. The resistance R is proportional to ϕ_1 and ϕ_2 . For convenience, I will write

$$I_1 = J_1 - \frac{2\pi}{\hbar} \Phi, \quad I_2 = J_2 - \frac{2\pi}{\hbar} \Phi.$$

Then,

$$\begin{aligned} I_{\text{total}} &= J_1 [\sin(\phi_1 - \frac{2\pi}{\hbar} \Phi) + \cos(\phi_1 - \frac{2\pi}{\hbar} \Phi)] \\ &= J_1 \sin \phi_1 \cos \frac{2\pi}{\hbar} \Phi. \end{aligned} \quad (21-22)$$

Now we don't know anything about ϕ_1 , we have to adjust that anyway she wants depending on the circumstances. In particular, it will depend on the forward voltage applied to the junction. No matter what we do, however, I_{total} is definitely not going to be zero. So the current I_{total} for any given Φ is given by

$$I_{\text{total}} = J_1 \sin \left| \frac{2\pi}{\hbar} \Phi \right|$$

This maximum current will vary with the total voltage V because whenever

$$\Phi = \frac{q}{\hbar} V$$

with q some integer. This is to say, the current takes on its maximum value whenever the magnetic flux just reaches quantum values $(21-22)$: 1.7×10^{-19} Wb.

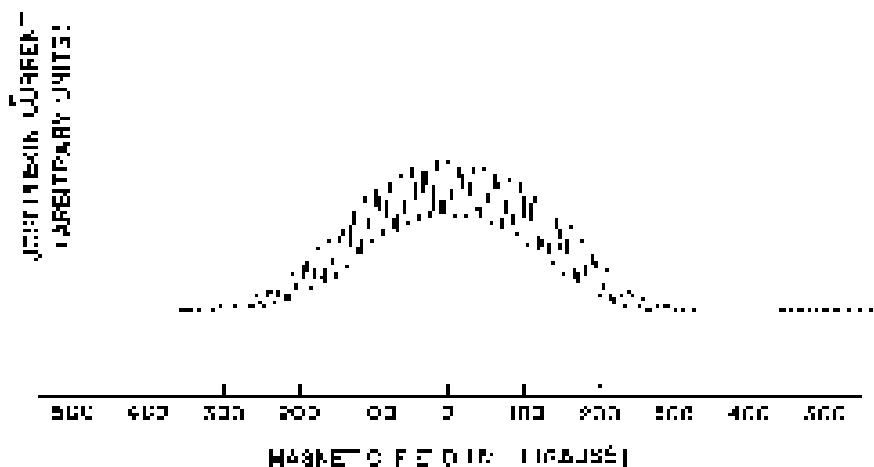


Fig. 21-5. A recording of the current through pairs of 10-ohm resistors in a portion of the magnetic field in the region between the two junctions, see Fig. 21-7. [The recording was made by R. C. Johnson, J. Lamb, A. H. Silver, and J. E. Goss, of the Bell Telephone Laboratories, New Haven, Connecticut.]

The Josephson current can be simply given by the relation of the magnetic field in the area between the two junctions. The results are shown in Fig. 21-6. There is a series of blocks, and of course, from various effects we have neglected, will be a great number of steps in the current with changes in the magnetic field due to the interference term as given by Eq. (21-1).

One of the big questions in solid quantum mechanics is the question of whether the wave potential exists in a phase wave function or not.¹¹ This aspect of wave theory just described has also been done with a very similar between the two junctions so that the only significant magnetic field is inside the junctions and a negligibly small it is outside superconducting rings themselves. Yet it is remarkable that the amount of current depends nonlinearly on the flux of magnetic field outside the solenoid even though that field never crosses the wire—an important demonstration of the "potential reality" of the vector potential.¹²

I don't know what will come next. But look what can be done. First, notice that the interference between two junctions can be used to make a sensitive magnetometer. If a set of junctions is made with an encoder gear set, say, with the max in the cause of Fig. 21-5 which is separated by 2×10^{-4} cm/cm², at room temperature to be 10° K, then you are 1000° K away from zero peaks, and it should be possible to use such a gear to measure magnetic field as small as 2×10^{-7} gauss, or to me is no longer nothing to such a precision. They should be able to go even farther. Suppose for example we put a set of 10 or 20 junctions close together and equally spaced. Then we will have the difference between the currents due to something like a magnetic field as well as temperature and distance. Instead of a field measurement we compare a 30, or perhaps even 100-bit instead of the 10, measuring the magnetic field. Perhaps we can make the measurement of a synapse. This will be a great technique of communications equipment—especially assuming more as power as the measurement of wavelength of light.

These then are some illustrations of some of the things that are happening in modern electronics. The last and new three months, where 20 more junctions appear, have not yet shown. The quantum mechanics which was the norm 10 years ago and nearly 10 years of development along rather systematically has begun to be exploited in many practical and cool ways. We are truly getting control of nature on a very delicate and basic fundamental.

I am sorry to say, gentlemen, that as participants in this seminar you will probably not participate and you will not be granted a Nobel Prize if possible. If you do have time in the future we would like a way to make an application so you at the earliest possible moment, the inventor of this part of physics.

¹¹ Johnson, Lamb, Silver, and Meissner, Phys. Rev. Letters 12, 274 (1959).

¹² Johnson, Lamb, Silver, and Meissner, Phys. Rev. Letters 14, 274 (1960).

Feymann's Epilogue

Well, I've been talking to you for two years and now I'm going to quit. In some ways I would like to apologize and in other ways not. I chose an feet. I know—
that 1990 is three years off your own birth date so follow everything were given
exclusion, and last had a good time with it. But I also know that "the powers of
intuition are of very little effect except in those tragic circumstances in which
they are powerfully exercised." So far, however, he unknown who has
reduced exclusion, only I say I have done nothing but show you the camp. For
the others, if I have made you like the subject I'm sorry. I have taught the history
payoffs before, and I apologize. I just hope that I haven't caused any trouble
to you... and that you do not leave this exciting process. I hope that someone else
can teach it to you in a way that doesn't give you information, and that you will
find something that, while all is not as horrific as I believe.

Finally, since I said that the total purpose of my teaching has not been to
prepare you for war or exclusion. It was not the. To perhaps you to serve in
it, stay in the military. I have chosen to give you some appreciation of the understand-
ing work and the payment way of working at it which, I believe, is a major part
of the true culture of modern times. (There are probably processes of other sub-
jects where one might feel that there is no. They are still, I think, working.)

Perhaps you will not only have some appreciation of this as well; it is even
possible that you may want to join in the greatest adventure that the human mind
has ever begun.