The Feynman LECTURES ON PHYSICS

MAINLY ELECTROMAGNETISM AND MATTER

RICHARD P. FEYNMAN

Richard Chace Tolman Professor of Theoretical Physics
California Institute of Technology

ROBERT B. LEIGHTON

Professor of Physics
California Institute of Technology

MATTHEW SANDS

Professor
Stanford University

ADDISON-WESLEY PUBLISHING COMPANY, INC.
READING, MASSACHUSETTS · PALO ALTO · LONDON

Printed in the United States of America

Library of Congress Catalog Card No. 63-20717

Second printing—November, 1964

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 " (Gibbons)

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

For some forty years Richard P. Feynman focussed his curiosity on the mysterious workings of the physical world, and bent his intellect to searching out the order in its chaos. Now, he has given two years of his ability and his energy to his Lectures on Physics for beginning students. For them he has distilled the essence of his knowledge, and has created in terms they can hope to grasp a picture of the physicist's universe. To his lectures he has brought the brilliance and clarity of his thought, the originality and vitality of his approach, and the contagious 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 tried in this the second volume to make some kind of a record of a part of the second year's lectures—which were given to the sophomore class during the 1962–1963 academic year. The rest of the second year's lectures will make up Volume III.

Of the second year of lectures, the first two-thirds were devoted to a fairly complete treatment of the physics of electricity and magnetism. Its presentation was intended to serve a dual purpose. We hoped, first, to give the students a complete view of one of the great chapters of physics—from the early gropings of Franklin, through the great synthesis of Maxwell, on to the Lorentz electron theory of material properties, and ending with the still unsolved dilemmas of the electromagnetic self-energy. And we hoped, second, by introducing at the outset the calculus of vector fields, to give a solid introduction to the mathematics of field theories To emphasize the general utility of the mathematical methods, related subjects from other parts of physics were sometimes analyzed together with their electric counterparts. We continually tried to drive home the generality of the mathematics. ("The same equations have the same solutions.") And we emphasized this point by the kinds of exercises and examinations we gave with the course.

Following the electromagnetism there are two chapters each on elasticity and fluid flow. In the first chapter of each pair, the elementary and practical aspects are treated. The second chapter on each subject attempts to give an overview of the whole complex range of phenomena which the subject can lead to. These four chapters can well be omitted without serious loss, since they are not at all a necessary preparation for Volume III.

The last quarter, approximately, of the second year was dedicated to an introduction to quantum mechanics. This material has been put into the third volume.

In this record of the Feynman Lectures we wished to do more than provide a transcription of what was said. We hoped to make the written version as clear an exposition as possible of the ideas on which the original lectures were based For some of the lectures this could be done by making only minor adjustments of the wording in the original transcript. For others of the lectures a major reworking and rearrangement of the material was required. Sometimes we felt we should add some new material to improve the clarity or balance of the presentation. Throughout the process we benefitted from the continual help and advice of Professor Feynman

The translation of over 1,000,000 spoken words into a coherent text on a tight schedule is a formidable task, particularly when it is accompanied by the

other onerous burdens which come with the introduction of a new course—preparing for recitation sections, and meeting students, designing exercises and examinations, and grading them, and so on. Many hands—and heads—were involved. In some instances we have, I believe, been able to render a faithful image—or a tenderly retouched portrait—of the original Feynman. In other instances we have fallen far short of this ideal. Our successes are owed to all those who helped. The failures, we regret.

As explained in detail in the Foreword to Volume I, these lectures were but one aspect of a program initiated and supervised by the Physics Course Revision Committee (R. B. Leighton, Chairman, H. V. Neher, and M. Sands) at the California Institute of Technology, and supported financially by the Ford Foundation. In addition, the following people helped with one aspect or another of the preparation of textual material for this second volume: T. K. Caughey, M. L. Clayton, J. B. Curcio, J. B. Hartle, T. W. H. Harvey, M. H. Israel, W. J. Karzas, R. W. Kavanagh, R. B. Leighton, J. Mathews, M. S. Plesset, F. L. Warren, W. Whaling, C. H. Wilts, and B. Zimmerman. Others contributed indirectly through their work on the course: J. Blue, G. F. Chapline, M. J. Clauser, R. Dolen, H. H. Hill, and A. M. Title. Professor Gerry Neugebauer contributed in all aspects of our task with a diligence and devotion far beyond the dictates of duty.

The story of physics you find here would, however, not have been, except for the extraordinary ability and industry of Richard P. Feynman.

MATTHEW SANDS

March, 1964

Contents

CHAPTER 1. ELECTROMAGNETISM

1–1 Electrical forces 1–1
1–2 Electric and magnetic fields 1–3
1–3 Characteristics of vector fields 1–4
1–4 The laws of electromagnetism 1–5
1–5 What are the fields? 1–9
1–6 Electromagnetism 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—T and h 2–2
2–3 Derivatives of fields—the gradient 2–4
2–4 The operator ∇ 2–6
2–5 Operations with ∇ 2–7
2–6 The differential equation of heat flow 2–8
2–7 Second derivatives of vector fields 2–9
2–8 Pitfalls 2–11

CHAPTER 3. VECTOR INTEGRAL CALCULUS

3–1 Vector integrals; the line integral of $\nabla\Psi$ 3–1
3–2 The flux of a vector field 3–2
3–3 The flux from a cube; Gauss' theorem 3–4
3–4 Heat conduction; the diffusion equation 3–6
3–5 The circulation of a vector field 3–8
3–6 The circulation around a square; Stokes' theorem 3–9
3–7 Curl-free and divergence-free fields 3–10
3–8 Summary 3–11

CHAPTER 4. ELECTROSTATICS

4–1 Statics 4–1
4–2 Coulomb's law; superposition 4–2
4–3 Electric potential 4–4
4–4 $E = -\nabla\phi$ 4–6
4–5 The flux of E 4–7
4–6 Gauss' law; divergence of E 4–9
4–7 Field of a sphere of charge 4–10
4–8 Field lines; equipotential surfaces 4–11

CHAPTER 5. APPLICATION OF GAUSS' LAW

5–1 Electrostatics is Gauss's law plus . . . 5–1
5–2 Equilibrium in an electrostatic field 5–1
5–3 Equilibrium with conductors 5–2
5–4 Stability of atoms 5–3
5–5 The field of a line charge 5–3
5–6 A sheet of charge; two sheets 5–4
5–7 A sphere of charge; a spherical shell 5–4
5–8 Is the field of a point charge exactly $1/r^2$? 5–5
5–9 The fields of a conductor 5–7
5–10 The field in a cavity of a conductor 5–8

CHAPTER 6. THE ELECTRIC FIELD IN VARIOUS CIRCUMSTANCES

6–1 Equations of the electrostatic potential 6–1
6–2 The electric dipole 6–2
6–3 Remarks on vector equations 6–4
6–4 The dipole potential as a gradient 6–4
6–5 The dipole approximation for an arbitrary distribution 6–6
6–6 The fields of charged conductors 6–8
6–7 The method of images 6–8
6–8 A point charge near a conducting plane 6–9
6–9 A point charge near a conducting sphere 6–10
6–10 Condensers; parallel plates 6–11
6–11 High-voltage breakdown 6–13
6–12 The field-emission microscope 6–14

CHAPTER 7. THE ELECTRIC FIELD IN VARIOUS CIRCUMSTANCES (Continued)

7–1 Methods for finding the electrostatic field 7–1
7–2 Two-dimensional fields; functions of the complex variable 7–2
7–3 Plasma oscillations 7–5
7–4 Colloidal particles in an electrolyte 7–8
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. Forces on charged conductors 8–2
8–3 The electrostatic energy of an ionic crystal 8–4
8–4 Electrostatic energy in nuclei 8–6
8–5 Energy in the electrostatic field 8–9
8–6 The energy of a point charge 8–12

CHAPTER 9. ELECTRICITY 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 Thunderstorms 9–5
9–5 The mechanism of charge separation 9–7
9–6 Lightning 9–10

CHAPTER 10. DIELECTRICS

10–1 The dielectric constant 10–1
10–2 The polarization vector $\boldsymbol{P}$ 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. INSIDE DIELECTRICS

11–1 Molecular dipoles 11–1
11–2 Electronic polarization 11–1
11–3 Polar molecules; orientation polarization 11–3
11–4 Electric fields in cavities of a dielectric 11–5
11–5 The dielectric constant of liquids; the Clausius-Mossotti equation 11–6
11–6 Solid dielectrics 11–8
11–7 Ferroelectricity; $BaTiO_3$ 11–8

CHAPTER 12. ELECTROSTATIC ANALOGS

12–1 The same equations have the same solutions 12–1
12–2 The flow of heat; a point source near an infinite plane boundary 12–2
12–3 The stretched membrane 12–5
12–4 The diffusion of neutrons; a uniform spherical source in a homogeneous medium 12–6
12–5 Irrotational fluid flow; the flow past a sphere 12–8
12–6 Illumination; the uniform lighting of a plane 12–10
12–7 The "underlying unity" of nature 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; Ampere's law 13–3
13–5 The magnetic field of a straight wire and of a solenoid; atomic currents 13–5
13–6 The relativity of magnetic and electric fields 13–6
13–7 The transformation of currents and charges 13–11
13–8 Superposition; the right-hand rule 13–11

CHAPTER 14. THE MAGNETIC FIELD IN VARIOUS SITUATIONS

14–1 The vector potential 14–1
14–2 The vector potential of known currents 14–3
14–3 A straight wire 14–4
14–4 A long solenoid 14–5
14–5 The field of a small 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 VECTOR POTENTIAL

15–1 The forces on a current loop; energy of a dipole 15–1
15–2 Mechanical and electrical energies 15–3
15–3 The energy of steady currents 15–6
15–4 $\boldsymbol{B}$ versus $\boldsymbol{A}$ 15–7
15–5 The vector potential and quantum mechanics 15–8
15–6 What is true for statics is false for dynamics 15–14

CHAPTER 16. INDUCED CURRENTS

16–1 Motors and generators 16–1
16–2 Transformers and inductances 16–4
16–3 Forces on induced currents 16–5
16–4 Electrical technology 16–8

CHAPTER 17. THE LAWS OF INDUCTION

17–1 The physics of induction 17–1
17–2 Exceptions to the "flux rule" 17–2
17–3 Particle acceleration by an induced electric field; the betatron 17–3
17–4 A paradox 17–5
17–5 Alternating-current generator 17–6
17–6 Mutual inductance 17–9
17–7 Self-inductance 17–11
17–8 Inductance and magnetic energy 17–12

CHAPTER 18. THE MAXWELL EQUATIONS

18–1 Maxwell's equations 18–1
18–2 How the new term works 18–3
18–3 All of classical physics 18–5
18–4 A travelling field 18–5
18–5 The speed of light 18–8
18–6 Solving Maxwell's equations; the potentials and the wave equation 18–9

CHAPTER 19. THE PRINCIPLE OF LEAST ACTION

A special lecture—almost verbatim 19–1
A note added after the lecture 19–14

CHAPTER 20. SOLUTIONS OF MAXWELL'S EQUATIONS IN FREE SPACE

20–1 Waves in free space; plane waves 20–1
20–2 Three-dimensional waves 20–8
20–3 Scientific imagination 20–9
20–4 Spherical waves 20–12

CHAPTER 21. SOLUTIONS OF MAXWELL'S EQUATIONS WITH CURRENTS AND CHARGES

21–1 Light and 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 fields of an oscillating dipole 21–5
21–5 The potentials of a moving charge; the general solution of Liénard and Wiechert 21–9
21–6 The potentials for a charge moving with constant velocity; the Lorentz formula 21–12

CHAPTER 22. AC CIRCUITS

22–1 Impedances 22–1
22–2 Generators 22–5
22–3 Networks of ideal elements; Kirchhoff's rules 22–7
22–4 Equivalent circuits 22–10
22–5 Energy 22–11
22–6 A ladder network 22–12
22–7 Filters 22–14
22–8 Other circuit elements 22–16

CHAPTER 23. CAVITY RESONATORS

23–1 Real circuit elements 23–1
23–2 A capacitor at high frequencies 23–2
23–3 A resonant cavity 23–6
23–4 Cavity modes 23–9
23–5 Cavities and resonant circuits 23–10

CHAPTER 24. WAVEGUIDES

24–1 The transmission line 24–1
24–2 The rectangular waveguide 24–4
24–3 The cutoff frequency 24–6
24–4 The speed of the guided waves 24–7
24–5 Observing guided waves 24–7
24–6 Waveguide plumbing 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 NOTATION

25–1 Four-vectors 25–1
25–2 The scalar product 25–3
25–3 The four-dimensional gradient 25–6
25–4 Electrodynamics in four-dimensional notation 25–8
25–5 The four-potential of a moving charge 25–9
25–6 The invariance of the equations of electrodynamics 25–10

CHAPTER 26. LORENTZ TRANSFORMATIONS OF THE FIELDS

26–1 The four-potential of a moving charge 26–1
26–2 The fields of a point charge with a constant velocity 26–2
26–3 Relativistic transformation of the fields 26–5
26–4 The equations of motion in relativistic notation 26–11

CHAPTER 27. FIELD ENERGY AND FIELD MOMENTUM

27–1 Local conservation 27–1
27–2 Energy conservation and electromagnetism 27–2
27–3 Energy density and energy flow in the electromagnetic field 27–3
27–4 The ambiguity of the field energy 27–6
27–5 Examples of energy flow 27–6
27–6 Field momentum 27–9

CHAPTER 28. ELECTROMAGNETIC MASS

28–1 The field energy of a point charge 28–1
28–2 The field momentum of a moving charge 28–2
28–3 Electromagnetic mass 28–3
28–4 The force of an electron on itself 28–4
28–5 Attempts to modify the Maxwell theory 28–6
28–6 The nuclear force field 28–12

CHAPTER 29. THE MOTION OF CHARGES IN ELECTRIC AND MAGNETIC FIELDS

29–1 Motion in a uniform electric or magnetic field 29–1
29–2 Momentum analysis 29–1
29–3 An electrostatic lens 29–2
29–4 A magnetic lens 29–3
29–5 The electron microscope 29–3
29–6 Accelerator guide fields 29–4
29–7 Alternating-gradient focusing 29–6
29–8 Motion in crossed electric and magnetic fields 29–8

CHAPTER 30. THE INTERNAL GEOMETRY OF CRYSTALS

30–1 The internal geometry of crystals 30–1
30–2 Chemical bonds in crystals 30–2
30–3 The growth of crystals 30–3
30–4 Crystal lattices 30–3
30–5 Symmetries in two dimensions 30–4
30–6 Symmetries in three dimensions 30–7
30–7 The strength of metals 30–8
30–8 Dislocations and crystal growth 30–9
30–9 The Bragg-Nye crystal model 30–10

CHAPTER 31. TENSORS

31–1 The tensor of polarizability 31–1
31–2 Transforming the tensor components 31–3
31–3 The energy ellipsoid 31–3
31–4 Other tensors; the tensor of inertia 31–6
31–5 The cross product 31–8
31–6 The tensor of stress 31–9
31–7 Tensors of higher rank 31–11
31–8 The four-tensor of electromagnetic momentum 31–12

CHAPTER 32. REFRACTIVE INDEX OF DENSE MATERIALS

32–1 Polarization of matter 32–1
32–2 Maxwell's equations in a dielectric 32–3
32–3 Waves in a dielectric 32–5
32–4 The complex index of refraction 32–8
32–5 The index of a mixture 32–8
32–6 Waves in metals 32–10
32–7 Low-frequency and high-frequency approximations; the skin depth and the plasma frequency 32–11

CHAPTER 33. REFLECTION FROM SURFACES

33–1 Reflection and refraction of light 33–1
33–2 Waves in dense materials 33–2
33–3 The boundary conditions 33–4
33–4 The reflected and transmitted waves 33–7
33–5 Reflection from metals 33–11
33–6 Total internal reflection 33–12

CHAPTER 34. THE MAGNETISM OF MATTER

34–1 Diamagnetism and paramagnetism 34–1
34–2 Magnetic moments and angular momentum 34–3
34–3 The precession of atomic magnets 34–4
34–4 Diamagnetism 34–5
34–5 Larmor's theorem 34–6
34–6 Classical physics gives neither diamagnetism nor paramagnetism 34–8
34–7 Angular momentum in quantum mechanics 34–8
34–8 The magnetic energy of atoms 34–11

CHAPTER 35. PARAMAGNETISM AND MAGNETIC RESONANCE

35–1 Quantized magnetic states 35–1
35–2 The Stern-Gerlach experiment 35–3
35–3 The Rabi molecular-beam method 35–4
35–4 The paramagnetism of bulk materials 35–6
35–5 Cooling by adiabatic demagnetization 35–9
35–6 Nuclear magnetic resonance 35–10

CHAPTER 36. FERROMAGNETISM

36–1 Magnetization currents 36–1
36–2 The field H 36–5
36–3 The magnetization curve 36–6
36–4 Iron-core inductances 36–8
36–5 Electromagnets 36–9
36–6 Spontaneous magnetization 36–11

CHAPTER 37. MAGNETIC MATERIALS

37–1 Understanding ferromagnetism 37–1
37–2 Thermodynamic properties 37–4
37–3 The hysteresis curve 37–5
37–4 Ferromagnetic materials 37–10
37–5 Extraordinary magnetic materials 37–11

CHAPTER 38. ELASTICITY

38–1 Hooke's law 38–1
38–2 Uniform strains 38–2
38–3 The torsion bar; shear waves 38–5
38–4 The bent beam 38–9
38–5 Buckling 38–11

CHAPTER 39. ELASTIC MATERIALS

39–1 The tensor of strain 39–1
39–2 The tensor of elasticity 39–4
39–3 The motions in an elastic body 39–6
39–4 Nonelastic behavior 39–8
39–5 Calculating the elastic constants 39–10

CHAPTER 40. THE FLOW OF DRY WATER

40–1 Hydrostatics 40–1
40–2 The equations of motion 40–2
40–3 Steady flow—Bernoulli's theorem 40–6
40–4 Circulation 40–9
40–5 Vortex lines 40–10

CHAPTER 41. THE FLOW OF WET WATER

41–1 Viscosity 41–1
41–2 Viscous flow 41–4
41–3 The Reynolds number 41–5
41–4 Flow past a circular cylinder 41–7
41–5 The limit of zero viscosity 41–9
41–6 Couette flow 41–10

INDEX

I

Electromagnetism

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

1–1 Electrical forces

Consider a force like gravitation which varies predominantly inversely as the square of the distance, but which is about a *billion-billion-billion-billion* times stronger. And with another difference. There are two kinds of "matter," which we can call positive and negative. Like kinds repel and unlike kinds attract—unlike gravity where 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 attractions. The net result would be that the terrific forces would balance themselves out almost perfectly, by forming tight, fine mixtures of the positive and the negative, and between two separate bunches of such mixtures there would be practically no attraction or repulsion at all.

Review: Chapter 12, Vol. I, *Characteristics of Force*

There is such a force: the electrical force. And all matter is a mixture of positive protons and negative electrons which are attracting and repelling with this great force. So perfect is the balance, however, that when you stand near someone else you don't feel any force at all. If there were even a little bit of unbalance you would know it. If you were standing at arm's length from someone and each of you had *one percent* more electrons than protons, the repelling force would be incredible. How great? Enough to lift the Empire State Building? No! To lift Mount Everest? No! 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 is not hard to understand that matter, trying to keep its positive and negative charges in the finest balance, can have a great stiffness and strength. The Empire State Building, for example, swings only eight feet in the wind because the electrical forces hold every electron and proton more or less in its proper place. On the other hand, if we look at matter on 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 residual electrical 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 as 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 charges 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 force that holds the atoms together, and the chemical forces 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 atoms are made with positive protons in the nucleus and with electrons outside. You may ask: "If this electrical force is so terrific, why don't the protons and electrons just get on top of each other? If they want to be in an intimate mixture, why isn't it still more intimate?" The answer has to do with the quantum effects. If we try to confine our electrons in a region that is very close to the protons, then according to the uncertainty principle they must have some mean square momentum which is larger the more we try to confine them. It is this motion, required by the laws of quantum mechanics, that keeps the electrical attraction from bringing the charges any closer together.

Lower case Greek letters and commonly used capitals

α		alpha
β		beta
γ	Γ	gamma
δ	Δ	delta
ϵ		epsilon
ζ		zeta
η		eta
θ	Θ	theta
ι		iota
κ		kappa
λ	Λ	lambda
μ		mu
ν		nu
ξ	Ξ	xi (ksi)
o		omicron
π	Π	pi
ρ		rho
σ	Σ	sigma
τ		tau
υ	Υ	upsilon
ϕ	Φ	phi
χ		chi (khi)
ψ	Ψ	psi
ω	Ω	omega

There is another question: "What holds the nucleus together"? In a nucleus there are several protons, all of which are positive. Why don't they push themselves apart? It turns out that in nuclei there are, in addition to electrical forces, nonelectrical forces, called nuclear forces, which are greater than the electrical forces and which are able to hold the protons together in spite of the electrical repulsion. The nuclear forces, however, have a short range—their force falls off much more rapidly than $1/r^2$. 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. An example is uranium, with 92 protons. The nuclear forces act mainly between each proton (or neutron) and its nearest neighbor, while the electrical forces act over larger distances, giving a repulsion between each proton and all of the others in the nucleus. The more protons in a nucleus, the stronger is the electrical repulsion, until, as in the case of uranium, the balance is so delicate that the nucleus is almost ready to fly apart from the repulsive electrical force. If such a nucleus is just "tapped" lightly (as can be done by sending in a slow neutron), it breaks into two pieces, each with positive charge, and these pieces fly apart by electrical repulsion. The energy which is liberated is the energy of the atomic bomb. This energy is usually called "nuclear" energy, but it is really "electrical" energy released when electrical forces have overcome the attractive nuclear forces.

We may ask, finally, what holds a negatively charged electron 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. Why, then, doesn't it fly apart? 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 that the electron does not act upon itself. Perhaps. All we can say is that the question of what holds the electron together has produced many difficulties in the attempts to form a complete theory of electromagnetism. The question has never been answered. We will entertain overselves by discussing this subject some more in later chapters.

As we have seen, we should expect that it is a combination of electrical forces and quantum-mechanical effects that will determine the detailed structure of materials in bulk, and, therefore, their properties. Some materials are hard, some are soft. Some are electrical "conductors"—because their electrons are free to move about; others are "insulators"—because their electrons are held tightly to individual atoms. We shall consider later how some of these properties come about, but that is a very complicated subject, so we will begin by looking at the electrical forces only in simple situations. We begin by treating only the laws of electricity—including magnetism, which is really a part of the same subject.

We have said that the electrical force, like a gravitational force, decreases 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 motions 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 electrical effect. That is why we call the subject "electromagnetism."

There is an important general principle that makes it possible to treat electromagnetic forces in a relatively simple way. We find, from experiment, that the force that acts on a particular charge—no matter how many other charges there are or how they are moving—depends only on the position of that particular charge, on the velocity of the charge, and on the amount of charge. We can write the force $\boldsymbol{F}$ on a charge q moving with a velocity $\boldsymbol{v}$ as

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{1.1}$$

We call $\boldsymbol{E}$ the *electric field* and $\boldsymbol{B}$ 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 vectors. Their values will depend on *where* the charge is, and may change with *time*. Furthermore, if we replace that charge with another charge, the force on the new charge will be just in proportion to the amount of charge so long as all the rest of the charges in the

world do not 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 fields *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

$$\frac{d}{dt}\left[\frac{m\mathbf{v}}{(1-v^2/c^2)^{1/2}}\right] = \mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \tag{1.2}$$

So if $\mathbf{E}$ and $\mathbf{B}$ are given, we can find the motions. Now we need to know how the $\mathbf{E}$'s and $\mathbf{B}$'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 $\mathbf{E}_1$, and another set of charges would produce $\mathbf{E}_2$. 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

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2. \tag{1.3}$$

This fact 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 produced 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 $\mathbf{E}$ and $\mathbf{B}$ produced by each of the charges B, C, D, etc., and then add the $\mathbf{E}$'s and $\mathbf{B}$'s from all the charges to find the fields, and from them the forces acting on charge A. If it had only turned out that the field produced by a single charge was simple, this would be the neatest way to describe the laws of electrodynamics. We have already given a description of this law (Chapter 28, Vol. I) and it is, unfortunately, rather complicated.

It turns out that the form in which the laws of electrodynamics are simplest are not what you might expect. It is *not* simplest to give a formula for the force that one charge produces on another. It is true that when charges are standing still the Coulomb force law is simple, but when charges are moving about the relations are complicated by delays in time and by the effects of acceleration, among others. As a result, we do not wish to present electrodynamics only through the force laws between charges; we find it more convenient to consider 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, $\mathbf{E}$ and $\mathbf{B}$. We have defined them in terms of the forces that are felt by a charge. We wish now to speak of electric and magnetic fields *at a point* even when there is no charge present. We are saying, in effect, that since there are forces "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 $\mathbf{F}$ given by Eq. (1.1) we associate the vectors $\mathbf{E}$ and $\mathbf{B}$ with *the point* in space (x, y, z). We may think of $\mathbf{E}(x, y, z, t)$ and $\mathbf{B}(x, y, z, t)$ as giving the forces that *would be* experienced at the time t by a charge located at (x, y, z), *with the condition* that placing the charge there *did not disturb* the positions or motions of all the other charges responsible for the fields.

Following this idea, we associate with *every* point (x, y, z) in space two vectors $\mathbf{E}$ and $\mathbf{B}$, 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 $\mathbf{E}(x, y, z, t)$ and $\mathbf{B}(x, y, z, t)$ represent three mathematical functions of x, y, z, and t.

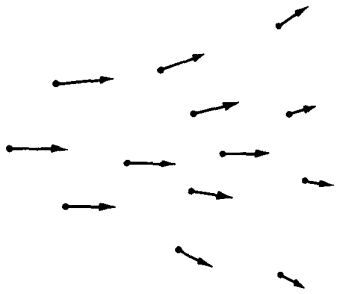


Fig. 1–1. A vector field may be represented by drawing a set of arrows whose magnitudes and directions indicate the values of the vector field at the points from which the arrows are drawn.

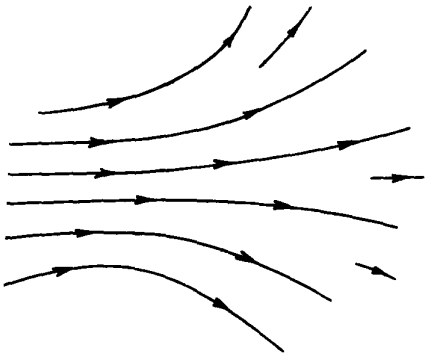


Fig. 1–2. A vector field can be represented by drawing lines which are tangent to the direction of the field vector at each point, and by drawing the density of lines proportional to the magnitude of the field vector.

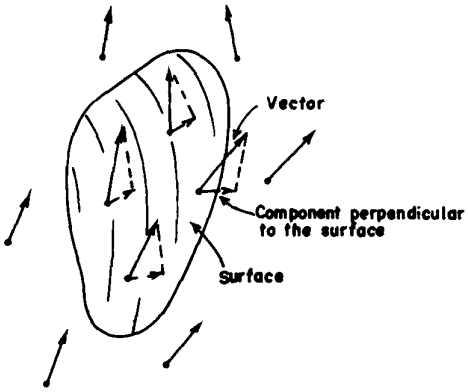


Fig. 1–3. The flux of a vector field through a surface is defined as the average value of the normal component of the vector times the area of the surface.

It is precisely because E (or 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 $v(x, y, z, t)$ for the velocity of the liquid at each point in space at the time t. It is a vector field.

Returning to the electromagnetic fields—although they are produced by charges according to complicated formulas, they have the following important characteristic: the relationships between the values of the fields at *one point* and the values at a *nearby point* are very simple. With only a few such relationships in the form of differential equations we can describe the fields completely. It is in terms of such equations that the laws of electrodynamics are most simply written.

There have been various inventions to help the mind visualize 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 mental picture of the field by drawing vectors at many points in space, each of which gives the field 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, so to speak, follow the arrows and keep track of the direction of the field. When we do this we lose track of the *lengths* of the vectors, but we can keep track of the strength of the field by drawing the lines far apart when the field is weak and close together when 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 new lines sometimes 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.

1–3 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 electricity from the field point of view. Suppose we imagine a closed surface of some kind and ask whether we are losing "something" from the inside; that is, does the field have a quality of "outflow"? For instance, for a velocity field we might ask whether the velocity is always outward on the surface or, more generally, whether more fluid flows out (per unit time) than comes in. We call the net amount of fluid going out through the surface per unit time the "flux of velocity" through the surface. The flow through an element of a surface is just equal to the component of the velocity perpendicular to the surface times the area of the surface. For an arbitrary closed surface, the *net outward flow*—or *flux*—is the average outward normal component of the velocity, times the area of the surface:

$$\text{Flux} = (\text{average normal component})\cdot(\text{surface area}). \tag{1.4}$$

In the case of an electric field, we can mathematically define something analogous to an outflow, and we again call it the flux, but of course 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 still has a useful significance. We speak, then, of the *electric flux*—also defined by Eq. (1.4). 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 line, rather than 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?

By that we mean: Is there a net rotational motion around some loop? Suppose that we instantaneously 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 in the trapped liquid—that is, if there is more momentum heading one way around the tube than the other. We define a quantity called the *circulation* as the resulting speed of the liquid in the tube times its circumference. We can again extend our ideas and define the "circulation" for any vector field (even when there isn't anything moving). For any vector field the *circulation around any imagined 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).

$$\text{Circulation} = (\text{average tangential component})\cdot(\text{distance around}). \quad (1.5)$$

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 laws right away, but they will give you some idea of the way the physics of electromagnetism will be ultimately described.

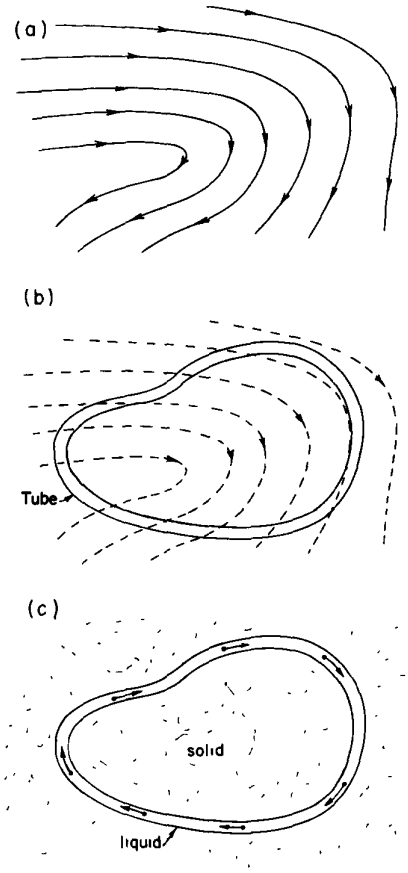


Fig. 1–4. (a) The velocity field in a liquid. Imagine a tube of uniform cross section that follows an arbitrary closed curve as in (b). If the liquid were suddenly frozen everywhere except inside the tube, the liquid in the tube would circulate as shown in (c).

1–4 The laws of electromagnetism

The first law of electromagnetism describes the flux of the electric field:

$$\text{The flux of } \boldsymbol{E} \text{ through any closed surface} = \frac{\text{the net charge inside}}{\epsilon_0}, \quad (1.6)$$

where ϵ_0 is a convenient constant. (The constant ϵ_0 is usually read as "epsilon-zero" or "epsilon-naught".) If there are no charges inside the surface, even though there are charges nearby outside the surface, the *average* normal component of $\boldsymbol{E}$ is zero, so there is no net flux through the surface. To show the power of this type of statement, we can show that Eq. (1.6) is the same as Coulomb's law, provided only that we also add the idea 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 $\boldsymbol{E}$ at any point, since the field must be directed 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 still 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 not, in general, zero (although it is for the Coulomb field). Rather, for electricity there is a second law that states: for any surface S (not closed) whose edge is the curve C,

$$\text{Circulation of } \boldsymbol{E} \text{ around } C = \frac{d}{dt}(\text{flux of } \boldsymbol{B} \text{ through } S). \quad (1.7)$$

We can complete the laws of the electromagnetic field by writing two corresponding equations for the magnetic field $\boldsymbol{B}$.

$$\text{Flux of } \boldsymbol{B} \text{ through any closed surface} = 0. \quad (1.8)$$

For a surface S bounded by the curve C,

$$c^2(\text{circulation of } \boldsymbol{B} \text{ around } C) = \frac{d}{dt}(\text{flux of } \boldsymbol{E} \text{ through } S) + \frac{\text{flux of electric current through } S}{\epsilon_0}. \quad (1.9)$$

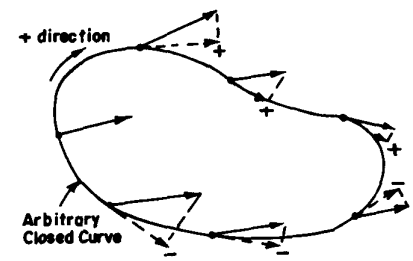


Fig. 1–5. The circulation of a vector field is the average tangential component of the vector (in a consistent sense) times the circumference of the loop.

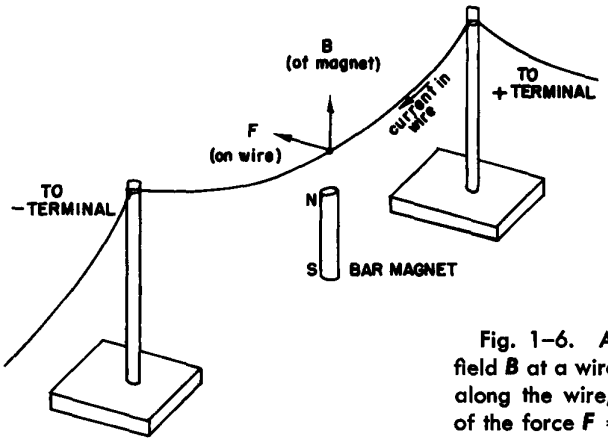


Fig. 1–6. A bar magnet gives a field $\boldsymbol{B}$ at a wire. When there is a current along the wire, the wire moves because of the force $\boldsymbol{F} = q\boldsymbol{v} \times \boldsymbol{B}$.

The constant c^2 that appears in Eq. (1.9) is the square of the velocity of light. It appears because magnetism is in reality a relativistic effect of electricity. The constant ϵ_0 has been stuck in to make the units of electric current come out in a convenient way.

Equations (1.6) through (1.9), 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 all 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. You have experienced 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 $\boldsymbol{F} = q\boldsymbol{v} \times \boldsymbol{B}$. When a current exists, the charges inside the wire are moving, so they have a velocity $\boldsymbol{v}$, and the magnetic field from the magnet exerts a force on them, which results in pushing 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 propulsion 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 push 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

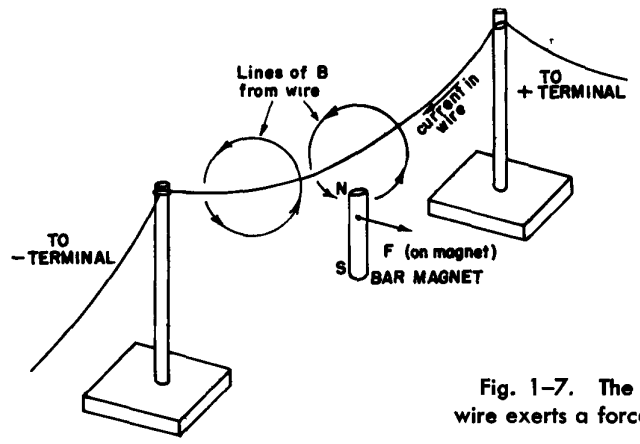


Fig. 1–7. The magnetic field of the wire exerts a force on the magnet.

* We need only to add a remark about some conventions for the *sign* of the circulation.

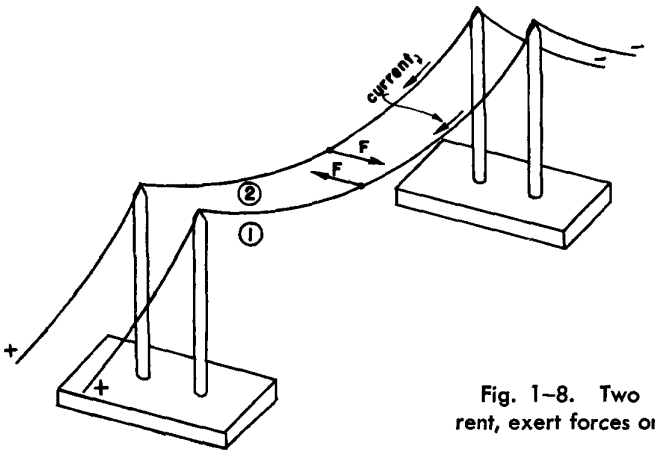


Fig. 1–8. Two wires, carrying current, exert forces on each other.

term in Eq. (1.9), a current must have a *circulation* of $\boldsymbol{B}$—in this case, the lines of $\boldsymbol{B}$ are loops around the wire, as shown in Fig. 1–7. This $\boldsymbol{B}$-field is responsible for the force on the magnet.

Equation (1.9) tells us that for a fixed current through the wire the circulation of $\boldsymbol{B}$ 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 $\boldsymbol{B}$ must decrease. You can see that we would, in fact, expect $\boldsymbol{B}$ to decrease linearly with the distance from a long straight wire.

Now, we have said that a current through a wire produces a magnetic field, and that when there is a magnetic field present there is a force on a wire carrying a current. Then we should also expect that if we make a magnetic field 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 currents are in the same direction, the two wires attract, but when the currents are opposite, they repel.

In short, electrical currents, as well as magnets, 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 imitates a magnet. It appears, then, that a piece of iron acts as though it contains a perpetual 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).

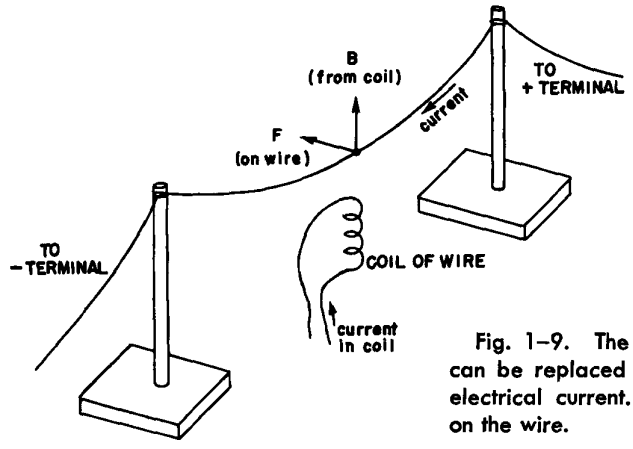


Fig. 1–9. The bar magnet of Fig. 1–6 can be replaced by a coil carrying an electrical current. A similar force acts on the wire.

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 moving around in an atom, an electron also spins about on its own axis—something like the spin of the earth—and it is the current from this spin that gives the magnetic field in iron. (We say "something like the spin of the earth" because the question is so deep in quantum mechanics that the classical ideas do not really describe things too well.) In most substances, some electrons spin one way and 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 magnets. 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 found.

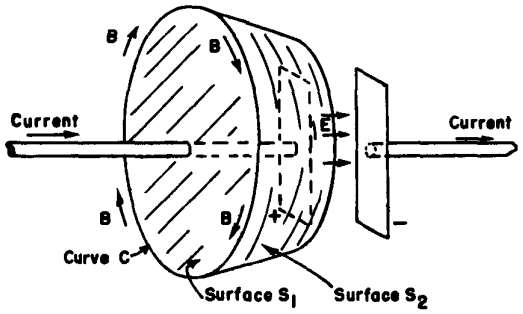


Fig. 1–10. The circulation of $\boldsymbol{B}$ around the curve C is given either by the current passing through the surface S_1, or by the rate of change of the flux of $\boldsymbol{E}$ through the surface S_2.

The first term on the right-hand side of Eq. (1.9) 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 could 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 toward one plate and away from the other, as shown in Fig. 1–10. We draw a curve C around one of the wires and fill it in with a surface which crosses the wire, as shown by the surface S_1 in the figure. According to Eq. (1.9), the circulation of $\boldsymbol{B}$ around C is given by the current in the wire (times c^2). But what if we fill 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. But, surely, just changing the location of an imaginary surface is not going to change a real magnetic field! The circulation of $\boldsymbol{B}$ must be what it was before. The first term on the right-hand side of Eq. (1.9) 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 $\boldsymbol{B}$ is given in terms of the rate of change of the flux of $\boldsymbol{E}$ between the plates of the capacitor. And it works out that the changing $\boldsymbol{E}$ is related to the current in just the way required for Eq. (1.9) to be correct. Maxwell saw that it was needed, and he was the first to write the complete equation.

With the setup shown in Fig. 1–6 we can demonstrate another of the laws of electromagnetism. We disconnect the ends of the hanging 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 $\boldsymbol{F} = q\boldsymbol{v} \times \boldsymbol{B}$. The electrons have a sidewise velocity because they move with the wire. This $\boldsymbol{v}$ with a vertical $\boldsymbol{B}$ from the magnet results in a force on the electrons directed *along* the wire, which starts the electrons moving toward the galvanometer.

Suppose, however, that we leave the wire 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. How does the magnetic field produce forces on charges at rest? According to Eq. (1.1) there must be an electric field. A moving magnet must make an electric field. How that happens is said quantitatively by Eq. (1.7). This equation describes many phenomena of great practical interest, such as those that occur in electric generators and transformers.

The most remarkable consequence of our equations is that the combination of Eq. (1.7) and Eq. (1.9) contains the explanation of the radiation of electromagnetic effects over large distances. The reason is roughly something like this: suppose that somewhere we have a magnetic field which is increasing 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 work their way through space without the need 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 fields.

1–5 What are the fields?

We now make a few remarks on our way of looking at this subject. You may be saying: "All this business of fluxes and circulations is pretty abstract. There are electric fields at every point in space; then there are these 'laws.' But what is *actually* happening? 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 inconceivable. (How could they find an idea inconceivable when it had already been conceived?) They would say: "Look, the only forces we know are the direct action of one piece of matter on another. It is impossible that there can be a force with nothing to transmit it." But what really happens when we study the "direct action" of one piece of matter right against another? We discover that it is not one piece right against the other; they are slightly separated, and there are electrical forces acting on a tiny scale. Thus we find that we are going to explain so-called direct-contact action in terms of the picture for electrical forces. It is certainly not sensible to try to insist that an electrical force has to look like the old, familiar, muscular push or pull, when it will turn out that the muscular 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 effects. Some people prefer to represent them as the interaction at a distance of charges, and to use a complicated law. Others love the field lines. They draw field lines all the time, and feel that writing $\boldsymbol{E}$'s and $\boldsymbol{B}$'s is too abstract. The field lines, however, are only a crude way of describing a field, and it is very difficult to give the correct, quantitative laws directly in terms of field lines. Also, the ideas 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 line patterns will look like when both sets are present together. From the mathematical standpoint, on the other hand, superposition is easy—we simply add the two vectors. The field lines have some advantage in giving a vivid picture, but they also have some disadvantages. The direct interaction way of thinking has great advantages when thinking of electrical charges at rest, but has great disadvantages when dealing with charges in rapid motion.

The best way is to use the abstract field idea. That it is abstract is unfortunate, but necessary. The attempts to try to represent the electric field as the motion of some kind of gear wheels, or in terms of lines, or of stresses in some kind of material have used up more effort of physicists than it would have taken simply to get the right answers about electrodynamics. It is interesting that the correct equations for the behavior of light in crystals were worked out by McCullough in 1843. But

people said to him: "Yes, but there is no real material whose mechanical properties could possibly satisfy those equations, and since light is an oscillation that must vibrate in *something*, we cannot believe this abstract equation business." If people had been more open-minded, they might have believed in the right equations for the behavior of light a lot earlier than they did.

In the case of the magnetic field we can make the following point: Suppose that 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 at the same speed and parallel to each other. Because they are moving, they will behave like two currents and will have a magnetic field associated with them (like the currents in the wires of Fig. 1–8). 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 draw field lines are in a similar difficulty. Not only is it not possible to say whether the field lines move or do not move with charges—they may disappear completely in certain coordinate frames

What we are saying, then, 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 must correspond to the magnetic force. But what about the force between the two wires in our experiment (Fig. 1–8). 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^2/c^2 is about 10^{-25}. Surely a negligible "correction." But no! Although the magnetic force is, in this case, 10^{-25} of the "normal" electrical force between the moving electrons, remember that the "normal" electrical forces have disappeared because of the almost perfect balancing out—because the wires have the same number of protons as electrons. The balance is much more precise than one part in 10^{25}, and the small relativistic term which we call the magnetic force is the only term left. It becomes the dominant term.

It is the near-perfect cancellation of electrical effects which allowed relativity effects (that is, magnetism) to be studied and the correct equations—to order v^2/c^2—to be discovered, even though physicists didn't *know* that's what was happening. And that is why, when relativity was discovered, the electromagnetic laws didn't need to be changed. They—unlike mechanics—were already 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 papyrus, and that there was a strange rock from the island of Magnesia which attracted iron. It is amazing 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 fantastic precision of the balancing of charges that we mentioned earlier. Study by scientists who came after the Greeks uncovered one new phenomena after another that were really some aspect 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, technical possibilities that defied the imagination of the people that came before were appearing: it became possible to signal by telegraph over long distances, and to talk to another person miles away without any connections between, and to run huge power systems—a great water wheel, connected by

filaments over hundreds of miles to another engine that turns in response to the master wheel—many thousands of branching filaments—ten thousand engines in ten thousand places running the machines of industries and homes—all turning because of the 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 man is doing by the way he affects the electrons in a thin metal rod hundreds of miles away. All we need to do is to use 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 19th century will be judged as Maxwell's discovery of the laws of electrodynamics. The American Civil War will pale into provincial insignificance in comparison with this important scientific event of the same decade.

2

Differential Calculus of Vector Fields

2–1 Understanding physics

2–2 Scalar and vector fields—T and h

2–3 Derivatives of fields—the gradient

2–4 The operator ∇

2–5 Operations with ∇

2–6 The differential equation of heat flow

2–7 Second derivatives of vector fields

2–8 Pitfalls

Review: Chapter 11, Vol. I, *Vectors*

2–1 Understanding physics

The physicist needs a facility in looking at problems from several points of view. The exact analysis of real physical problems is usually 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. Ideas such as the field lines, capacitance, 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 feel as to what should happen in different electromagnetic situations. On the other hand, none of the heuristic models, such as field lines, is really adequate and accurate for all situations. There is only one precise way of presenting the laws, and that is by means of differential equations. They have the advantage of being fundamental and, so far as we know, precise. If you have learned the differential equations you can always go back to them. There is nothing to unlearn.

It will take you some time 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 study 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 very mathematical minds, are often led astray when "studying" physics because they lose sight of the physics. They say: "Look, these differential equations—the Maxwell equations—are all there is to electrodynamics; it is admitted by the physicists that there is nothing which is not contained in the equations. The equations are complicated, but after all they are only mathematical equations and if I understand them mathematically inside out, I will understand the physics inside out." Only it doesn't work that way. Mathematicians who study physics with that point of view—and there have been many of them—usually make little contribution to physics and, in fact, little to mathematics. They fail because the actual 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 unmathematical, imprecise, and inexact 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 take first the *complete* laws, and then to step back and apply them to simple situations, developing the physical ideas as we go along. And that is what we are going to do.

Our approach is completely opposite to the historical approach in which one develops 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 ingenious people, and as we have only a limited time to acquire 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 laboratory some of this lack can be corrected. You can also fill in what we must leave out by reading the Encyclopedia Brittanica, 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.

2–2 Scalar and vector fields—T and h

We begin now with the abstract, mathematical view of the theory of electricity and magnetism. The ultimate idea is to explain the meaning of the laws given in Chapter 1. But to do this we must first explain a new and peculiar notation that we want to use. So let us forget electromagnetism for the moment and discuss the mathematics of vector fields. It is of very great importance, 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.

Listed below are a few facts from the algebra of vectors. It is assumed that you already know them.

$$A \cdot B = \text{scalar} = A_xB_x + A_yB_y + A_zB_z \tag{2.1}$$

$$A \times B = \text{vector} \tag{2.2}$$
$$(A \times B)_z = A_xB_y - A_yB_x$$
$$(A \times B)_x = A_yB_z - A_zB_y$$
$$(A \times B)_y = A_zB_x - A_xB_z$$

$$A \times A = 0 \tag{2.3}$$

$$A \cdot (A \times B) = 0 \tag{2.4}$$

$$A \cdot (B \times C) = (A \times B) \cdot C \tag{2.5}$$

$$A \times (B \times C) = B(A \cdot C) - C(A \cdot B) \tag{2.6}$$

Also we will want to use the two following equalities from the calculus:

$$\Delta f(x, y, z) = \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial z}\Delta z, \tag{2.7}$$

$$\frac{\partial^2 f}{\partial x\,\partial y} = \frac{\partial^2 f}{\partial y\,\partial x}. \tag{2.8}$$

The first equation (2.7) is, of course, true only in the limit that Δx, Δy, and Δz go toward zero.

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 in time, 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 of a scalar field, consider a solid block of material which has been heated at some places and cooled at others, so that the temperature of the body 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 rectangular coordinate system. Temperature is a scalar field.

Writing vectors by hand.

Some people use
$\vec{E}$ or $\bar{E}$ or just $\bar{E}$.

Others prefer
$\underline{E}$.

We like the following way:

A B C D E F G
H I J K L M N
O P Q R S T U
V W X Y Z

Small letters are harder:

a b c d e f g
h i j k l m n
o p q r s t u
v w x y z

You can invent your own.

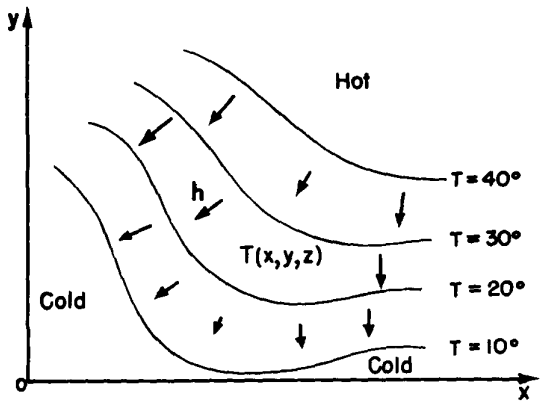


Fig. 2–1. Temperature T is an example of a scalar field. With each point (x, y, z) in space there is associated a number $T(x, y, z)$. All points on the surface marked $T = 20°$ (shown as a curve at $z = 0$) are at the same temperature. The arrows are samples of the heat flow vector $\boldsymbol{h}$.

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 called "isothermal surfaces" or 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 high at one place and low at another, there will be a flow of heat from the hotter places to the colder. The heat will be flowing in different directions in different parts of the block. The heat flow is a directional quantity which we call $\boldsymbol{h}$. Its magnitude is a measure of how much heat is flowing. Examples of the heat flow vector are also shown in Fig. 2–1.

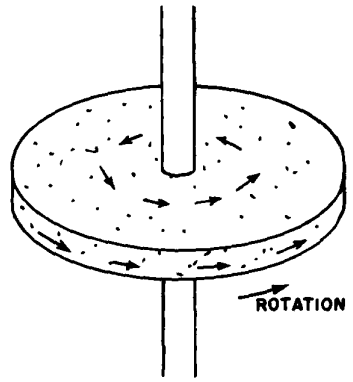


Fig. 2–2. The velocity of the atoms in a rotating object is an example of a vector field.

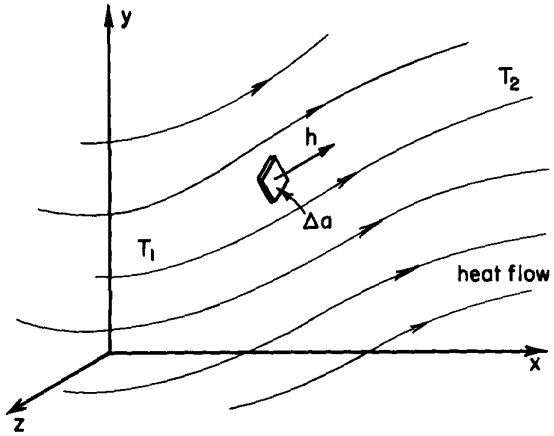


Fig. 2–3. Heat flow is a vector field. The vector $\boldsymbol{h}$ points along the direction of the flow. Its magnitude is the energy transported per unit time across a surface element oriented perpendicular to the flow, divided by the area of the surface element.

Let's make a more precise definition of $\boldsymbol{h}$: The magnitude of the vector heat flow 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 ΔJ is the thermal energy that passes per unit time through the surface element Δa, then

$$\boldsymbol{h} = \frac{\Delta J}{\Delta a}\, \boldsymbol{e}_f, \tag{2.9}$$

where $\boldsymbol{e}_f$ is a *unit vector* in the direction of flow.

The vector $\boldsymbol{h}$ can be defined in another way—in terms of its components. We ask how much heat flows through a small surface at *any* angle with respect to the flow. In Fig. 2–4 we show a small surface Δa_2 inclined with respect to Δa_1, which is perpendicular to the flow. The *unit vector* $\boldsymbol{n}$ is normal to the surface Δa_2. The

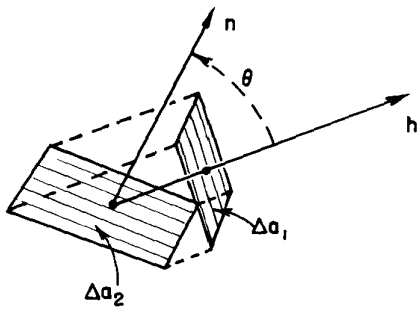


Fig. 2–4. The heat flow through Δa_2 is the same as through Δa_1.

angle θ between $\boldsymbol{n}$ and $\boldsymbol{h}$ is the same as the angle between the surfaces (since $\boldsymbol{h}$ is normal to Δa_1). Now what is the heat flow *per unit area* through Δa_2? The flow through Δa_2 is the same as through Δa_1; only the areas are different. In fact, $\Delta a_1 = \Delta a_2 \cos\theta$. The heat flow through Δa_2 is

$$\frac{\Delta J}{\Delta a_2} = \frac{\Delta J}{\Delta a_1}\cos\theta = \boldsymbol{h}\cdot\boldsymbol{n}. \tag{2.10}$$

We interpret this equation: the heat flow (per unit time and per unit area) through *any* surface element whose unit normal is $\boldsymbol{n}$, is given by $\boldsymbol{h}\cdot\boldsymbol{n}$. Equally, we could say: the component of the heat flow perpendicular to the surface element Δa_2 is $\boldsymbol{h}\cdot\boldsymbol{n}$. We can, if we wish, consider that these statements *define* $\boldsymbol{h}$. We will be applying the same ideas to other vector fields.

2–3 Derivatives of fields—the gradient

When fields vary in time, we can describe the variation by giving their derivatives with respect to t. We want to describe the variations with position in a similar way, because we are interested in the relationship between, say, the temperature in one place and the temperature at a nearby place. How shall we take the derivative of the temperature with respect to position? Do we differentiate the temperature with respect to x? Or with respect to y, or z?

Useful physical laws do not depend upon the orientation of the coordinate system. They should, therefore, be written in a form in which either both sides are scalars or both sides are vectors. What is the derivative of a scalar field, say $\partial T/\partial x$? Is it a scalar, or a vector, or what? It is neither a scalar nor a vector, as you can easily appreciate, because if we took a different x-axis, $\partial T/\partial x$ would certainly be different. But notice: We have three possible derivatives: $\partial T/\partial x$, $\partial T/\partial y$, and $\partial T/\partial z$. Since there 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:

$$\left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z}\right) \stackrel{?}{=} \text{a vector.} \tag{2.11}$$

Of course it is not generally true that *any* three numbers form a vector. It is true only if, when we rotate the coordinate system, the components of the vector transform among themselves in the correct way. So it is necessary to analyze how these derivatives are changed 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 see this in several ways. One way is to ask a question whose answer is independent of the coordinate system, and try to express the answer in an "invariant" form. For instance, if $S = \boldsymbol{A}\cdot\boldsymbol{B}$, and if $\boldsymbol{A}$ and $\boldsymbol{B}$ are vectors, we know—because we proved it in Chapter 11 of Vol. I—that S is a scalar. We *know* that S is a scalar without investigating whether it changes with changes in coordinate systems. It *can't*, because it's a dot product of two vectors. Similarly, if we *know* that $\boldsymbol{A}$ is a vector, and we have three numbers B_1, B_2, and B_3, and we find out that

$$A_xB_1 + A_yB_2 + A_zB_3 = S, \tag{2.12}$$

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 $\boldsymbol{B}$.

Now let's think of the temperature field. Suppose we take two points P_1 and P_2, separated by the small interval $\Delta\boldsymbol{R}$. The temperature at P_1 is T_1 and at P_2 is T_2, and the difference $\Delta T = T_2 - T_1$. The temperatures at these real, physical points certainly do not depend on what axis 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 $\Delta \boldsymbol{R}$ (Fig. 2–5). Remembering Eq. (2.7), we can write

$$\Delta T = \frac{\partial T}{\partial x}\Delta x + \frac{\partial T}{\partial y}\Delta y + \frac{\partial T}{\partial z}\Delta z. \tag{2.13}$$

The left side of Eq. (2.13) is a scalar. The right side is the sum of three products with 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 an upside-down Δ, and is supposed to remind us of differentiation. People read ∇T in various ways: "del-T," or "gradient of T," or "grad T;"

$$\text{grad } T = \nabla T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z}\right). * \tag{2.14}$$

Using this notation, we can rewrite Eq. (2.13) in the more compact form

$$\Delta T = \nabla T \cdot \Delta \boldsymbol{R}. \tag{2.15}$$

In words, 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 proof above that ∇T is indeed a vector.

Perhaps you are still not convinced? Let's prove it in a different way. (Although if you look carefully, you may be able to see that it's really the same proof in a longer-winded form!) We shall show that the components of ∇T transform in just the same way that components of $\boldsymbol{R}$ do. If they do, ∇T is a vector according to our original definition of a vector in Chapter 11 of Vol. I. We take 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 simpler, we let $z = z'$, so that we can forget about the z-coordinate. (You can check out the more general case for yourself.)

We take an $x'y'$-system rotated an angle θ with respect to the xy-system, as in Fig. 2–6(a). For a point (x, y) the coordinates in the prime system are

$$x' = x\cos\theta + y\sin\theta, \tag{2.16}$$

$$y' = -x\sin\theta + y\cos\theta. \tag{2.17}$$

Or, solving for x and y,

$$x = x'\cos\theta - y'\sin\theta, \tag{2.18}$$

$$y = x'\sin\theta + y'\cos\theta. \tag{2.19}$$

If any pair of numbers transforms with 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, chosen as in Fig. 2–6(b). If we calculate with the x- and y-coordinates, we would write

$$\Delta T = \frac{\partial T}{\partial x}\Delta x \tag{2.20}$$

—since Δy is zero.

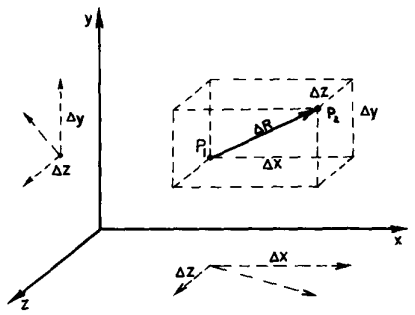


Fig. 2–5. The vector $\Delta \boldsymbol{R}$, whose components are Δx, Δy, and Δz.

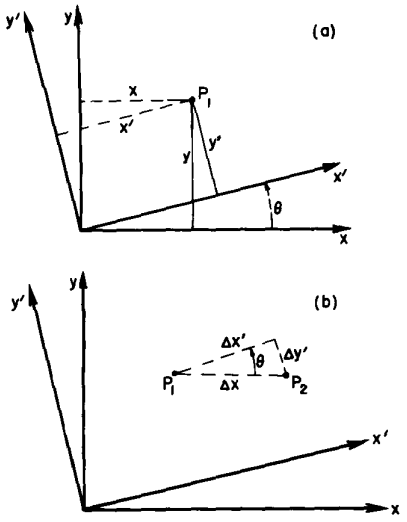


Fig. 2–6. (a) Transformation to a rotated coordinate system. (b) Special case of an interval $\Delta \boldsymbol{R}$ parallel to the x-axis.

* In our notation, the expression (a, b, c) represents a vector with components a, b, and c. If you like to use the unit vectors $\boldsymbol{i}, \boldsymbol{j}$, and $\boldsymbol{k}$, you may write

$$\nabla T = \boldsymbol{i}\frac{\partial T}{\partial x} + \boldsymbol{j}\frac{\partial T}{\partial y} + \boldsymbol{k}\frac{\partial T}{\partial z}.$$

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 $\Delta \boldsymbol{R}$ (Fig. 2–5). Remembering Eq. (2.7), we can write

$$\Delta T = \frac{\partial T}{\partial x}\Delta x + \frac{\partial T}{\partial y}\Delta y + \frac{\partial T}{\partial z}\Delta z. \tag{2.13}$$

The left side of Eq. (2.13) is a scalar. The right side is the sum of three products with 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 an upside-down Δ, and is supposed to remind us of differentiation. People read ∇T in various ways: "del-T," or "gradient of T," or "grad T;"

$$\text{grad } T = \nabla T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \frac{\partial T}{\partial z}\right).^{*} \tag{2.14}$$

Using this notation, we can rewrite Eq. (2.13) in the more compact form

$$\Delta T = \nabla T \cdot \Delta \boldsymbol{R}. \tag{2.15}$$

In words, 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 proof above that ∇T is indeed a vector.

Perhaps you are still not convinced? Let's prove it in a different way. (Although if you look carefully, you may be able to see that it's really the same proof in a longer-winded form!) We shall show that the components of ∇T transform in just the same way that components of $\boldsymbol{R}$ do. If they do, ∇T is a vector according to our original definition of a vector in Chapter 11 of Vol. I. We take 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 simpler, we let $z = z'$, so that we can forget about the z-coordinate. (You can check out the more general case for yourself.)

We take an $x'y'$-system rotated an angle θ with respect to the xy-system, as in Fig. 2–6(a). For a point (x, y) the coordinates in the prime system are

$$x' = x\cos\theta + y\sin\theta, \tag{2.16}$$

$$y' = -x\sin\theta + y\cos\theta. \tag{2.17}$$

Or, solving for x and y,

$$x = x'\cos\theta - y'\sin\theta, \tag{2.18}$$

$$y = x'\sin\theta + y'\cos\theta. \tag{2.19}$$

If any pair of numbers transforms with 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, chosen as in Fig. 2–6(b). If we calculate with the x- and y-coordinates, we would write

$$\Delta T = \frac{\partial T}{\partial x}\Delta x \tag{2.20}$$

—since Δy is zero.

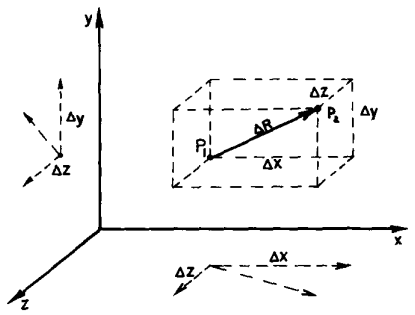
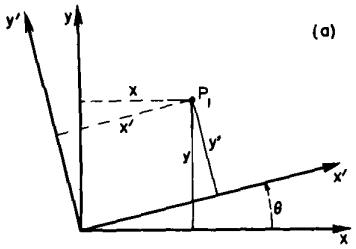


Fig. 2–5. The vector $\Delta \boldsymbol{R}$, whose components are Δx, Δy, and Δz.



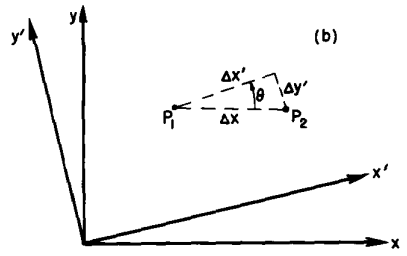


Fig. 2–6. (a) Transformation to a rotated coordinate system. (b) Special case of an interval $\Delta \boldsymbol{R}$ parallel to the x-axis.

* In our notation, the expression (a, b, c) represents a vector with components a, b, and c. If you like to use the unit vectors $\boldsymbol{i}, \boldsymbol{j}$, and $\boldsymbol{k}$, you may write

$$\nabla T = \boldsymbol{i}\frac{\partial T}{\partial x} + \boldsymbol{j}\frac{\partial T}{\partial y} + \boldsymbol{k}\frac{\partial T}{\partial z}.$$

What would a computation in the prime system give? We would have written

$$\Delta T = \frac{\partial T}{\partial x'}\,\Delta x' + \frac{\partial T}{\partial y'}\,\Delta y'. \tag{2.21}$$

Looking at Fig. 2–6(b), we see that

$$\Delta x' = \Delta x \cos\theta \tag{2.22}$$

and

$$\Delta y' = -\Delta x \sin\theta, \tag{2.23}$$

since Δy is negative when Δx is positive. Substituting these in Eq. (2.21), we find that

$$\Delta T = \frac{\partial T}{\partial x'}\,\Delta x \cos\theta - \frac{\partial T}{\partial y'}\,\Delta x \sin\theta \tag{2.24}$$

$$= \left(\frac{\partial T}{\partial x'}\cos\theta - \frac{\partial T}{\partial y'}\sin\theta\right)\Delta x. \tag{2.25}$$

Comparing Eq. (2.25) with (2.20), we see that

$$\frac{\partial T}{\partial x} = \frac{\partial T}{\partial x'}\cos\theta - \frac{\partial T}{\partial y'}\sin\theta. \tag{2.26}$$

This equation says that $\partial T/\partial x$ 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). So $\partial T/\partial x$ 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 amusing and ingenious—and characteristic of the things that make mathematics beautiful. The argument that grad T, or ∇T, is a vector did not depend upon *what* scalar field we were differentiating. All the arguments would go the same if T were replaced by *any scalar field.* Since the transformation equations are the same no matter what we differentiate, we could just as well omit the T and replace Eq. (2.26) by the operator equation

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial x'}\cos\theta - \frac{\partial}{\partial y'}\sin\theta. \tag{2.27}$$

We leave the operators, as Jeans said, "hungry for something to differentiate."

Since the differential operators themselves transform as the components of a vector should, we can call them components of a *vector operator.* We can write

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right), \tag{2.28}$$

which means, of course,

$$\nabla_x = \frac{\partial}{\partial x}, \qquad \nabla_y = \frac{\partial}{\partial y}, \qquad \nabla_z = \frac{\partial}{\partial z}. \tag{2.29}$$

We have abstracted the gradient away from the T—that is the wonderful idea.

You must always remember, of course, that ∇ is an operator. Alone, it means nothing. If ∇ 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 does not mean anything. Its x-component is

$$T\frac{\partial}{\partial x}, \tag{2.30}$$

which is not a number, but is still some kind of operator. However, according to the algebra of vectors we would still call $T\nabla$ a vector.

Now let's multiply ∇ by a scalar on the other side, so that we have the product (∇T). In ordinary algebra

$$TA = AT, \tag{2.31}$$

but we have to remember that operator algebra is a little different from ordinary vector algebra. With operators we must always keep the sequence right, so that the operations make proper sense. You will have no difficulty if you just remember that the operator ∇ obeys the same convention as the derivative notation. What is to be differentiated must be placed on the right of the ∇. The order is important.

Keeping in mind this problem of order, we understand that $T\nabla$ is an operator, but the product ∇T is no longer a hungry operator; the operator is completely satisfied. It is indeed a physical vector having a meaning. It represents the spatial rate of change of T. The x-component of ∇T is how fast T changes in the x-direction. What is the direction of the vector ∇T? We know that the rate of change of T in any direction is the component of ∇T in that direction (see Eq. 2.15). It follows that the direction of ∇T is that in which it has the largest possible component—in other words, the direction in which T changes the fastest. The gradient of T has the direction of the steepest uphill slope (in T).

2–5 Operations with ∇

Can we do any other algebra with the vector operator ∇? Let us try combining it with a vector. We can combine two vectors by making a dot product. We could make the products

$$(\text{a vector}) \cdot \nabla, \qquad \text{or} \qquad \nabla \cdot (\text{a vector}).$$

The first one doesn't mean anything yet, because it is still an operator. What it might ultimately mean would depend on what it is made to operate on. The second product is some scalar field. ($\boldsymbol{A} \cdot \boldsymbol{B}$ is always a scalar.)

Let's try the dot product of ∇ with a vector field we know, say $\boldsymbol{h}$. We write out the components:

$$\nabla \cdot \boldsymbol{h} = \nabla_x h_x + \nabla_y h_y + \nabla_z h_z \tag{2.32}$$

or

$$\nabla \cdot \boldsymbol{h} = \frac{\partial h_x}{\partial x} + \frac{\partial h_y}{\partial y} + \frac{\partial h_z}{\partial z}. \tag{2.33}$$

The sum is invariant under a coordinate transformation. If we were to choose a different system (indicated by primes), we would have*

$$\nabla' \cdot \boldsymbol{h} = \frac{\partial h_{x'}}{\partial x'} + \frac{\partial h_{y'}}{\partial y'} + \frac{\partial h_{z'}}{\partial z'}, \tag{2.34}$$

which is the *same* number as would be gotten from Eq. (2.33), even though it looks different. That is,

$$\nabla' \cdot \boldsymbol{h} = \nabla \cdot \boldsymbol{h} \tag{2.35}$$

for every point in space. So $\nabla \cdot \boldsymbol{h}$ is a scalar field, which must represent some physical quantity. You should realize that the combination of derivatives in $\nabla \cdot \boldsymbol{h}$ is rather special. There are all sorts of other combinations like $\partial h_y/\partial x$, which are neither scalars nor components of vectors.

The scalar quantity $\nabla \cdot$ (a vector) is extremely useful in physics. It has been given the name the *divergence*. For example,

$$\nabla \cdot \boldsymbol{h} = \text{div}\, \boldsymbol{h} = \text{"divergence of } \boldsymbol{h}\text{."} \tag{2.36}$$

As we did for ∇T, we can ascribe a physical significance to $\nabla \cdot \boldsymbol{h}$. We shall, however, postpone that until later.

---

* We think of $\boldsymbol{h}$ as a *physical* quantity that depends on position in space, and not strictly as a mathematical function of three variables. When $\boldsymbol{h}$ is "differentiated" with respect to x, y, and z, or with respect to x', y', and z', the mathematical expression for $\boldsymbol{h}$ must first be expressed as a function of the appropriate variables.

First, we wish to see what else we can cook up with the vector operator ∇. What about a cross product? We must expect that

$$\nabla \times \boldsymbol{h} = \text{a vector.} \tag{2.37}$$

It is a vector whose components we can write by the usual rule for cross products (see Eq. 2.2):

$$(\nabla \times \boldsymbol{h})_z = \nabla_x h_y - \nabla_y h_x = \frac{\partial h_y}{\partial x} - \frac{\partial h_x}{\partial y}. \tag{2.38}$$

Similarly,

$$(\nabla \times \boldsymbol{h})_x = \nabla_y h_z - \nabla_z h_y = \frac{\partial h_z}{\partial y} - \frac{\partial h_y}{\partial z} \tag{2.39}$$

and

$$(\nabla \times \boldsymbol{h})_y = \nabla_z h_x - \nabla_x h_z = \frac{\partial h_x}{\partial z} - \frac{\partial h_z}{\partial x}. \tag{2.40}$$

The combination $\nabla \times \boldsymbol{h}$ is called "the *curl* of $\boldsymbol{h}$." The reason for the name and the physical meaning of the combination will be discussed later.

Summarizing, we have three kinds of combinations with ∇:

$$\begin{aligned} \nabla T &= \text{grad } T = \text{a vector,} \\ \nabla \cdot \boldsymbol{h} &= \text{div } \boldsymbol{h} = \text{a scalar,} \\ \nabla \times \boldsymbol{h} &= \text{curl } \boldsymbol{h} = \text{a vector.} \end{aligned}$$

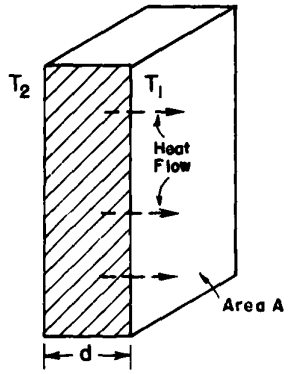
Using these combinations, we can write about the spatial variations of fields in a convenient way—in a way that is general, in that it doesn't depend on any particular set of axes.

As an example of the use of our vector differential operator ∇, we write a set of vector equations which contain the same laws of electromagnetism that we gave in words in Chapter 1. They are called Maxwell's equations.

Maxwell's Equations

$$\begin{aligned} &(1) \quad \nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0} \\ &(2) \quad \nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \\ &(3) \quad \nabla \cdot \boldsymbol{B} = 0 \\ &(4) \quad c^2 \nabla \times \boldsymbol{B} = \frac{\partial \boldsymbol{E}}{\partial t} + \frac{\boldsymbol{j}}{\epsilon_0} \end{aligned} \tag{2.41}$$

where ρ (rho), the "electric charge density," is the amount of charge per unit volume, and $\boldsymbol{j}$, the "electric current density," is the rate at which charge flows through a unit area per second. These four equations contain the complete classical theory of the electromagnetic field. You see what an elegantly simple form we can get with our new notation!



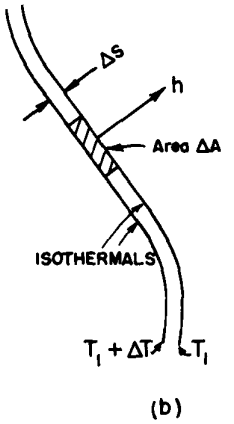


Fig. 2–7. (a) Heat flow through a slab. (b) An infinitesimal slab parallel to an isothermal surface in a large block.

2–6 The differential equation of heat flow

Let us give another example of a law of physics written in vector notation. The law is not a precise one, but 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_2 and cool the other to a different temperature T_1, the heat will flow through the material from T_2 to T_1 [Fig. 2–7(a)]. The heat flow is proportional to the area A of the faces, and to the temperature difference. It is also inversely proportional to d, the distance between the plates. (For a given temperature difference, the thinner the slab the greater the heat flow.) Letting J be the thermal energy that passes per unit time through the slab, we write

$$J = \kappa(T_2 - T_1)\frac{A}{d}. \tag{2.42}$$

The constant of proportionality κ (kappa) is called the *thermal conductivity*.

What will happen in 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 miniature scale. We orient the faces parallel to the isothermal surfaces, as in Fig. 2–7(b), so that Eq. (2.42) is correct for the small slab.

If the area of the small slab is ΔA, the heat flow per unit time is

$$\Delta J = \kappa \, \Delta T \, \frac{\Delta A}{\Delta s}, \tag{2.43}$$

where Δs is the thickness of the slab. Now $\Delta J/\Delta A$ we have defined earlier as the magnitude of $\boldsymbol{h}$, whose direction is the heat flow. The heat flow will be from $T_1 + \Delta T$ toward T_1, and so it will be perpendicular to the isotherms, as drawn in Fig. 2–7(b). Also, $\Delta T/\Delta s$ is just the rate of change of T with position. And since the position change is perpendicular to the isotherms, our $\Delta T/\Delta s$ is the maximum rate of change. It is, therefore, just the magnitude of ∇T. Now since the direction of ∇T is opposite to that of $\boldsymbol{h}$, we can write (2.43) as a vector equation:

$$\boldsymbol{h} = -\kappa \, \nabla T. \tag{2.44}$$

(The minus sign is necessary because heat flows "downhill" in temperature.) Equation (2.44) is the differential equation of heat conduction in bulk materials. You see that it is a proper vector equation. Each side is a vector if κ is just a number. It is the generalization to arbitrary cases of the special relation (2.42) for rectangular slabs. Later we should learn to write all sorts of elementary physics relations like (2.42) in the more sophisticated vector notation. This notation is useful not only because it makes the equations *look* simpler. It also shows most clearly the *physical content* of the equations without reference to any arbitrarily chosen coordinate system.

2–7 Second derivatives of vector fields

So far we have had only first derivatives. Why not second derivatives? We could have several combinations:

$$\begin{aligned} &\text{(a)} \quad \nabla \cdot (\nabla T) \\ &\text{(b)} \quad \nabla \times (\nabla T) \\ &\text{(c)} \quad \nabla(\nabla \cdot \boldsymbol{h}) \\ &\text{(d)} \quad \nabla \cdot (\nabla \times \boldsymbol{h}) \\ &\text{(e)} \quad \nabla \times (\nabla \times \boldsymbol{h}) \end{aligned} \tag{2.45}$$

You can check that these are all the possible combinations.

Let's look first at the second one, (b). It has the same form as

$$\boldsymbol{A} \times (\boldsymbol{A}T) = (\boldsymbol{A} \times \boldsymbol{A})T = 0,$$

since $\boldsymbol{A} \times \boldsymbol{A}$ is always zero. So we should have

$$\text{curl (grad } T) = \nabla \times (\nabla T) = 0. \tag{2.46}$$

We can see how this equation comes about if we go through once with the components:

$$\begin{aligned} [\nabla \times (\nabla T)]_z &= \nabla_x(\nabla T)_y - \nabla_y(\nabla T)_x \\ &= \frac{\partial}{\partial x}\left(\frac{\partial T}{\partial y}\right) - \frac{\partial}{\partial y}\left(\frac{\partial T}{\partial x}\right), \end{aligned} \tag{2.47}$$

which is zero (by Eq. 2.8). It goes the same for the other components. So $\nabla \times (\nabla T) = 0$, for any temperature distribution—in fact, for *any* 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:

$$\boldsymbol{A} \cdot (\boldsymbol{A} \times \boldsymbol{B}) = 0. \tag{2.48}$$

because $\boldsymbol{A} \times \boldsymbol{B}$ is perpendicular to $\boldsymbol{A}$, and so has no components in the direction $\boldsymbol{A}$. The same combination appears in (d) of (2.45), so we have

$$\nabla \cdot (\nabla \times \boldsymbol{h}) = \text{div (curl } \boldsymbol{h}) = 0. \tag{2.49}$$

Again, it is easy to show that it is zero by carrying through the operations with components.

Now we are going to state two mathematical theorems that 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—say of the vector field $\boldsymbol{A}$—is zero. Now we have seen (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, then, that $\boldsymbol{A}$ is the gradient of some quantity, because then its curl would necessarily be zero. The interesting theorem is that if the curl $\boldsymbol{A}$ is zero, then $\boldsymbol{A}$ is *always* the gradient of *something*—there is some scalar field ψ (psi) such that $\boldsymbol{A}$ is equal to grad ψ. In other words, we have the

THEOREM:

$$\begin{aligned} &\text{If} && \nabla \times \boldsymbol{A} = 0 \\ &\text{there is a} && \psi \\ &\text{such that} && \boldsymbol{A} = \nabla\psi. \end{aligned} \tag{2.50}$$

There is a similar theorem if the divergence of $\boldsymbol{A}$ is zero. We have seen in Eq. (2.49) that the divergence of a curl of something is always zero. If you come across a vector field $\boldsymbol{D}$ for which div $\boldsymbol{D}$ is zero, then you can conclude that $\boldsymbol{D}$ is the curl of some vector field $\boldsymbol{C}$.

THEOREM:

$$\begin{aligned} &\text{If} && \nabla \cdot \boldsymbol{D} = 0 \\ &\text{there is a} && \boldsymbol{C} \\ &\text{such that} && \boldsymbol{D} = \nabla \times \boldsymbol{C}. \end{aligned} \tag{2.51}$$

In looking at the possible combinations of two ∇ operators, we have found that two of them always give zero. Now we look at the ones that are *not* zero. Take the combination $\nabla \cdot (\nabla T)$, which was first on 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

$$\begin{aligned} \nabla \cdot (\nabla T) &= \nabla_x(\nabla_x T) + \nabla_y(\nabla_y T) + \nabla_z(\nabla_z T) \\ &= \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}, \end{aligned} \tag{2.52}$$

which would, in general, come out to be some number. It is a scalar field.

You see that we do not need to keep the parentheses, but can write, without any chance of confusion,

$$\nabla \cdot (\nabla T) = \nabla \cdot \nabla T = (\nabla \cdot \nabla) T = \nabla^2 T. \tag{2.53}$$

We look at ∇^2 as a new operator. It is a scalar operator. Because it appears often in physics, it has been given a special name—the *Laplacian*.

$$\text{Laplacian} = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{2.54}$$

Since the Laplacian is a scalar operator, we may operate with it on a vector—by which we mean the same operation on each component in rectangular coordinates:

$$\nabla^2 \boldsymbol{h} = (\nabla^2 h_x, \nabla^2 h_y, \nabla^2 h_z).$$

Let's look at one more possibility: $\nabla \times (\nabla \times \boldsymbol{h})$, which was (e) in the list (2.45). Now the curl of the curl can be written differently if we use the vector equality (2.6):

$$\boldsymbol{A} \times (\boldsymbol{B} \times \boldsymbol{C}) = \boldsymbol{B}(\boldsymbol{A} \cdot \boldsymbol{C}) - \boldsymbol{C}(\boldsymbol{A} \cdot \boldsymbol{B}). \tag{2.55}$$

In order to use this formula, we should replace $\boldsymbol{A}$ and $\boldsymbol{B}$ by the operator ∇ and put $\boldsymbol{C} = \boldsymbol{h}$. If we do that, we get

$$\nabla \times (\nabla \times \boldsymbol{h}) = \nabla(\nabla \cdot \boldsymbol{h}) - \boldsymbol{h}(\nabla \cdot \nabla) \ldots ???$$

Wait a minute! Something is wrong. The first two terms are vectors all right (the operators are satisfied), 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

$$\boldsymbol{A} \times (\boldsymbol{B} \times \boldsymbol{C}) = \boldsymbol{B}(\boldsymbol{A} \cdot \boldsymbol{C}) - (\boldsymbol{A} \cdot \boldsymbol{B})\boldsymbol{C}. \tag{2.56}$$

The order of terms looks better. Now let's make our substitution in (2.56). We get

$$\nabla \times (\nabla \times \boldsymbol{h}) = \nabla(\nabla \cdot \boldsymbol{h}) - (\nabla \cdot \nabla)\boldsymbol{h}. \tag{2.57}$$

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

$$\nabla \times (\nabla \times \boldsymbol{h}) = \nabla(\nabla \cdot \boldsymbol{h}) - \nabla^2 \boldsymbol{h}. \tag{2.58}$$

We have had something to say about all of the combinations in our list of double ∇'s, except for (c), $\nabla(\nabla \cdot \boldsymbol{h})$. It is a possible vector field, but there 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 our conclusions:

$$\begin{aligned}
&\text{(a)} \quad \nabla \cdot (\nabla T) = \nabla^2 T = \text{a scalar field} \\
&\text{(b)} \quad \nabla \times (\nabla T) = 0 \\
&\text{(c)} \quad \nabla(\nabla \cdot \boldsymbol{h}) = \text{a vector field} \\
&\text{(d)} \quad \nabla \cdot (\nabla \times \boldsymbol{h}) = 0 \\
&\text{(e)} \quad \nabla \times (\nabla \times \boldsymbol{h}) = \nabla(\nabla \cdot \boldsymbol{h}) - \nabla^2 \boldsymbol{h} \\
&\text{(f)} \quad (\nabla \cdot \nabla)\boldsymbol{h} = \nabla^2 \boldsymbol{h} = \text{a vector field}
\end{aligned} \tag{2.59}$$

You may notice that we haven't tried to invent 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 have to be careful, though, because it is possible to go astray. There are two pitfalls which we will mention, although they will not come up in this course. What would you say about the following expression, that involves the two scalar functions ψ and ϕ (phi):

$$(\nabla\psi) \times (\nabla\phi)?$$

You might want to say: it must be zero because it's just like

$$(Aa) \times (Ab),$$

which is zero because the cross product of two *equal* vectors $A \times A$ is always zero. But in our example the two operators ∇ are not equal! The first one operates on one function, ψ; the other operates on a different function, ϕ. So although we represent them by the same symbol ∇, they must be considered as different operators. Clearly, the direction of $\nabla\psi$ depends on the function ψ, so it is not likely to be parallel to $\nabla\phi$.

$$(\nabla\psi) \times (\nabla\phi) \neq 0 \quad \text{(generally)}.$$

Fortunately, we won't have to use such expressions. (What we have said doesn't change the fact that $\nabla \times \nabla\psi = 0$ for any scalar field, because here both ∇'s operate on the same function.)

Pitfall number two (which, again, we need not get into in our course) is the following: The rules that we have outlined here are simple and nice when we use rectangular coordinates. For example, if we have $\nabla^2\boldsymbol{h}$ and we want the x-component, it is

$$(\nabla^2\boldsymbol{h})_x = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) h_x = \nabla^2 h_x. \tag{2.60}$$

The same expression would *not* work if we were to ask for the *radial* component of $\nabla^2\boldsymbol{h}$. The radial component of $\nabla^2\boldsymbol{h}$ is not equal to $\nabla^2 h_r$. 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 in, say, polar coordinates, what we call the "radial" direction varies from point to point. So we can get into a lot of trouble when we start 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 worth mentioning: Since the Laplacian ∇^2, is a scalar, we can write it in any coordinate system we want to (for example, in polar coordinates). But 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 shall express all of our vector fields in terms of their x-, y-, and z-components when we write our vector differential equations out in components.

3

Vector Integral Calculus

3–1 Vector integrals; the line integral of $\nabla\Psi$

3–2 The flux 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 divergence-free fields

3–8 Summary

3–1 Vector integrals; the line integral of $\nabla\Psi$

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 operators $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ are the three components of a vector operator ∇. We would now 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 equation means.

We have already discussed the meaning of the gradient operation (∇ on a scalar). Now we turn to the meanings of the divergence and curl operations. The interpretation of these quantities is best done in terms of certain vector integrals and equations relating such integrals. These equations cannot, unfortunately, be obtained from vector algebra by some easy substitution, so you will just have to learn them as something new. Of these integral formulas, one is practically 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 content 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. General theorems like 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 delightful, however, that in the beginning of our subject there will be many simple problems which can be solved with the three integral formulas we are going to treat. We will see, however, as the problems get harder, that we can no 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 $\psi(x, y, z)$. At any two points (1) and (2), the function ψ will have the values $\psi(1)$ and $\psi(2)$, respectively. [We use a convenient notation, in which (2) represents the point (x_2, y_2, z_2) and $\psi(2)$ means the same thing as $\psi(x_2, y_2, z_2)$.] If Γ(gamma) is any curve joining (1) and (2), as in Fig. 3–1, the following relation is true:

THEOREM 1.
$$\psi(2) - \psi(1) = \int_{\substack{(1)\\ \text{along } \Gamma}}^{(2)} (\nabla\psi)\cdot d\boldsymbol{s}. \tag{3.1}$$

The integral is a *line integral*, from (1) to (2) along the curve Γ, of the dot product of $\nabla\psi$—a vector—with $d\boldsymbol{s}$—another vector which is an infinitesimal line element of the curve Γ (directed 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 these points 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

$$\int_{\substack{(1)\\ \text{along } \Gamma}}^{(2)} f\,ds$$

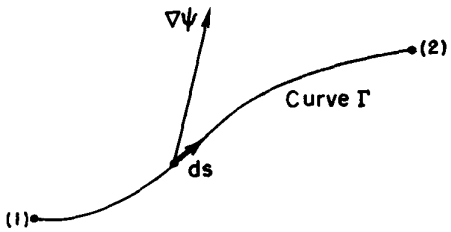


Fig. 3–1. The terms used in Eq. (3.1). The vector $\nabla\psi$ is evaluated at the line element $d\boldsymbol{s}$.

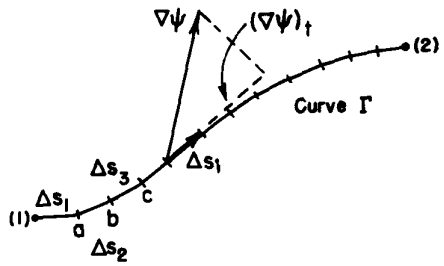


Fig. 3–2. The line integral is the limit of a sum.

we mean the limit of the sum

$$\sum f_i \, \Delta s_i,$$

where f_i is the value of the function at the ith segment. The limiting value is what the sum approaches as we add more and more segments (in a sensible way, so that the largest $\Delta s_i \to 0$).

The integral in our theorem, Eq. (3.1), means the same thing, although it looks a little different. Instead of f, we have another scalar—the component of $\nabla\psi$ in the direction of $\Delta \boldsymbol{s}$. If we write $(\nabla\psi)_t$ for this tangential component, it is clear that

$$(\nabla\psi)_t \, \Delta s = (\nabla\psi) \cdot \Delta \boldsymbol{s}. \tag{3.2}$$

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 $\Delta \boldsymbol{R}$ was the rate of change of ψ in the direction of ΔR. Consider the line segment $\Delta \boldsymbol{s}$ from (1) to point a in Fig. 3–2. According to our definition,

$$\Delta\psi_1 = \psi(a) - \psi(1) = (\nabla\psi)_1 \cdot \Delta \boldsymbol{s}_1. \tag{3.3}$$

Also, we have

$$\psi(b) - \psi(a) = (\nabla\psi)_2 \cdot \Delta \boldsymbol{s}_2, \tag{3.4}$$

where, of course, $(\nabla\psi)_1$ means the gradient evaluated at the segment $\Delta \boldsymbol{s}_1$, and $(\nabla\psi)_2$, the gradient evaluated at $\Delta \boldsymbol{s}_2$. If we add Eqs. (3.3) and (3.4), we get

$$\psi(b) - \psi(1) = (\nabla\psi)_1 \cdot \Delta \boldsymbol{s}_1 + (\nabla\psi)_2 \cdot \Delta \boldsymbol{s}_2. \tag{3.5}$$

You can see that if we keep adding such terms, we get the result

$$\psi(2) - \psi(1) = \sum (\nabla\psi)_i \cdot \Delta \boldsymbol{s}_i. \tag{3.6}$$

The left-hand side doesn't depend on how we choose our intervals—if (1) and (2) are kept always the same—so we can take the limit of the right-hand side. We have therefore proved Eq. (3.1).

You can see from our proof that just as the equality doesn't depend on how the points $a, b, c, \ldots$, are chosen, similarly it doesn't depend on what we choose for the curve Γ to join (1) and (2). Our theorem is correct for *any* curve from (1) to (2).

One remark on notation: You will see that there is no confusion if we write, for convenience,

$$(\nabla\psi) \cdot d\boldsymbol{s} = \nabla\psi \cdot d\boldsymbol{s}. \tag{3.7}$$

With this notation, our theorem is

THEOREM 1.

$$\psi(2) - \psi(1) = \int_{\substack{(1) \\ \text{any curve from} \\ (1) \text{ to } (2)}}^{(2)} \nabla\psi \cdot d\boldsymbol{s}. \tag{3.8}$$

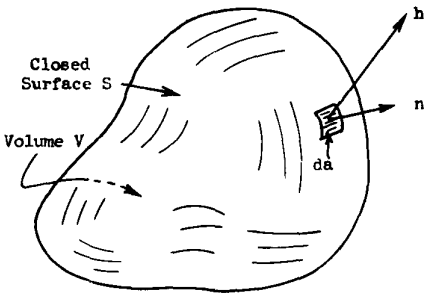


Fig. 3–3. The closed surface S defines the volume V. The unit vector $\boldsymbol{n}$ is the outward normal to the surface element da, and $\boldsymbol{h}$ is the heat-flow vector at the surface element.

3–2 The flux of a vector field

Before we consider our next integral theorem—a theorem about the divergence —we would like to study a certain idea which has an easily understood physical significance in the case of heat flow. We have defined the vector $\boldsymbol{h}$, which represents the heat that flows through a unit area in a unit time. Suppose that inside a block of material we have some closed surface S which encloses the volume V (Fig. 3–3). We would like to find out how much heat is flowing out of this *volume*. We can, of course, find it 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 area happened to be in the xy-plane we would have

$$da = dx \, dy.$$

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

$$dV = dx\,dy\,dz.$$

Some people like to write d^2a instead of da to remind themselves that it is kind of a second-order quantity. They would also write d^3V instead of dV. We will use the simpler notation, and assume that you can remember that an area has two dimensions and a volume has three.

The heat flow out through the surface element da is the area times the component of $\boldsymbol{h}$ perpendicular to da. We have already defined $\boldsymbol{n}$ as a unit vector pointing outward at right angles to the surface (Fig. 3–3). The component of $\boldsymbol{h}$ that we want is

$$h_n = \boldsymbol{h} \cdot \boldsymbol{n}. \tag{3.9}$$

The heat flow out through da is then

$$\boldsymbol{h} \cdot \boldsymbol{n}\, da. \tag{3.10}$$

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:

$$\text{Total heat flow outward through } S = \int_S \boldsymbol{h} \cdot \boldsymbol{n}\, da. \tag{3.11}$$

We are also going to call this surface integral "the flux of $\boldsymbol{h}$ through the surface." Originally the word flux meant flow, so that the surface integral just means the flow of $\boldsymbol{h}$ through the surface. We may think: $\boldsymbol{h}$ is the "current density" of heat flow and the surface integral of it is the total heat current directed out of the surface; that is, the thermal energy per unit time (joules per second).

We would like to generalize this idea to the case where the vector does not represent 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 if we wish. Although it is not the flow of anything, we still call it the "flux." We say

$$\text{Flux of } \boldsymbol{E} \text{ through the surface } S = \int_S \boldsymbol{E} \cdot \boldsymbol{n}\, da. \tag{3.12}$$

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 considered is not a closed one, as it is here.

Returning to the special case of heat flow, let us take a situation in which *heat is conserved.* For example, imagine some material in which after an initial heating no further heat energy is generated or absorbed. Then, if there is a net heat flow out of a closed surface, the heat content of the volume inside must decrease. So, in circumstances in which heat would be conserved, we say that

$$\int_S \boldsymbol{h} \cdot \boldsymbol{n}\, da = -\frac{dQ}{dt}, \tag{3.13}$$

where Q is the heat inside the surface. The heat flux out of S is equal to minus the rate of change with respect to time of the total heat Q inside of S. This interpretation is possible because we are speaking of heat flow and also because we supposed that the heat was conserved. We could not, of course, speak of the total heat inside the volume if heat were being generated there.

Now we shall point out an interesting fact about the flux of any vector. You may think of the heat flow vector if you wish, but what we say will be true for any vector field $\boldsymbol{C}$. 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_a and of the surface of the cut, S_{ab}. The volume V_2 is enclosed by S_2, which is made up of the rest of the original surface S_b and closed off by the cut S_{ab}. Now consider the

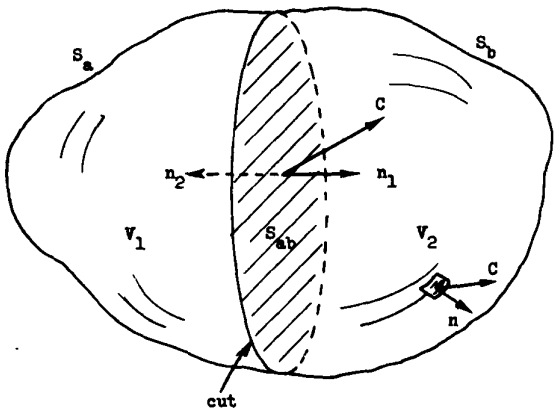


Fig. 3–4. A volume V contained inside the surface S is divided into two pieces by a "cut" at the surface S_{ab}. We now have the volume V_1 enclosed in the surface $S_1 = S_a + S_{ab}$ and the volume V_2 enclosed in the surface $S_2 = S_b + S_{ab}$.

following question: Suppose we calculate the flux out through surface S_1 and add to it the flux through surface S_2. Does the sum equal the flux through the whole surface that we started with? The answer is yes. The flux through the part of the surfaces S_{ab} common to both S_1 and S_2 just exactly cancels out. For the flux of the vector $\boldsymbol{C}$ out of V_1, we can write

$$\text{Flux through } S_1 = \int_{S_a} \boldsymbol{C} \cdot \boldsymbol{n}\, da + \int_{S_{ab}} \boldsymbol{C} \cdot \boldsymbol{n}_1\, da, \tag{3.14}$$

and for the flux out of V_2,

$$\text{Flux through } S_2 = \int_{S_b} \boldsymbol{C} \cdot \boldsymbol{n}\, da + \int_{S_{ab}} \boldsymbol{C} \cdot \boldsymbol{n}_2\, da. \tag{3.15}$$

Note that in the second integral we have written $\boldsymbol{n}_1$ for the outward normal for S_{ab} when it belongs to S_1, and $\boldsymbol{n}_2$ when it belongs to S_2, as shown in Fig. 3–4. Clearly, $\boldsymbol{n}_1 = -\boldsymbol{n}_2$, so that

$$\int_{S_{ab}} \boldsymbol{C} \cdot \boldsymbol{n}_1\, da = -\int_{S_{ab}} \boldsymbol{C} \cdot \boldsymbol{n}_2\, da. \tag{3.16}$$

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 just the sum of two integrals which, taken together, give the flux through the original surface $S = S_a + S_b$.

We see that the flux through the complete outer surface S can be considered as the sum of the fluxes from the two pieces into which the volume was broken. We can similarly subdivide again—say by cutting V_1 into two pieces. You see that the same arguments apply. So for *any* way of dividing the original volume, it must be 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 little interior pieces.

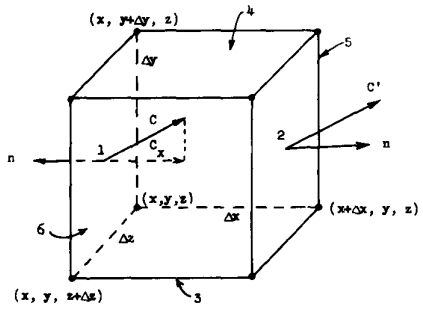


Fig. 3–5. Computation of the flux of C out of a small cube.

3–3 The flux from a cube; Gauss' theorem

We now take the special case of a small cube* and find an interesting formula for the flux out of it. Consider a cube whose edges are lined up with the axes as in Fig. 3–5. Let us suppose that the coordinates of the corner nearest the origin are x, y, z. Let Δx be the length of the cube in the x-direction, Δy be the length in the y-direction, and Δz be the length in the z-direction. We wish to find the flux of a vector field $\boldsymbol{C}$ through the surface of the cube. We shall do this by making a sum of the fluxes through each of the six faces. First, consider the face marked 1 in the figure. The flux *outward* on this face is the negative of the x-component of $\boldsymbol{C}$, integrated over the area of the face. This flux is

$$-\int C_x\, dy\, dz.$$

Since we are considering a *small* cube, we can approximate this integral by the

* The following development applies equally well to any rectangular parallelepiped.

value of C_x at the center of the face—which we call the point (1)—multiplied by the area of the face, $\Delta y\,\Delta z$:

$$\text{Flux out of 1} = -C_x(1)\,\Delta y\,\Delta z.$$

Similarly, for the flux out of face 2, we write

$$\text{Flux out of 2} = C_x(2)\,\Delta y\,\Delta z.$$

Now $C_x(1)$ and $C_x(2)$ are, in general, slightly different. If Δx is small enough, we can write

$$C_x(2) = C_x(1) + \frac{\partial C_x}{\partial x}\,\Delta x.$$

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

$$\text{Flux out of 2} = \left[C_x(1) + \frac{\partial C_x}{\partial x}\,\Delta x\right]\Delta y\,\Delta z.$$

Summing the fluxes for faces 1 and 2, we get

$$\text{Flux out of 1 and 2} = \frac{\partial C_x}{\partial x}\,\Delta x\,\Delta y\,\Delta z.$$

The derivative should really be evaluated at the center of face 1; that is, at $[x, y + (\Delta y/2), z + (\Delta z/2)]$. But in the limit of an infinitesimal cube, we make a negligible error if we evaluate it at the corner (x, y, z).

Applying the same reasoning to each of the other pairs of faces, we have

$$\text{Flux out of 3 and 4} = \frac{\partial C_y}{\partial y}\,\Delta x\,\Delta y\,\Delta z$$

and

$$\text{Flux out of 5 and 6} = \frac{\partial C_z}{\partial z}\,\Delta x\,\Delta y\,\Delta z.$$

The total flux through all the faces is the sum of these terms. We find that

$$\int_{\text{cube}} \boldsymbol{C}\cdot\boldsymbol{n}\,da = \left(\frac{\partial C_x}{\partial x} + \frac{\partial C_y}{\partial y} + \frac{\partial C_z}{\partial z}\right)\Delta x\,\Delta y\,\Delta z,$$

and the sum of the derivatives is just $\nabla\cdot\boldsymbol{C}$. Also, $\Delta x\,\Delta y\,\Delta z = \Delta V$, the volume of the cube. So we can say that *for an infinitesimal cube*

$$\int_{\text{surface}} \boldsymbol{C}\cdot\boldsymbol{n}\,da = (\nabla\cdot\boldsymbol{C})\,\Delta V. \tag{3.17}$$

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—the outgoing "flow" of $\boldsymbol{C}$—*per unit volume*, in the neighborhood of P.

We have connected the divergence of $\boldsymbol{C}$ to the flux of $\boldsymbol{C}$ out of each infinitesimal volume. For any finite volume we can use the fact we proved above—that the total flux from a volume is the sum of the fluxes out of each part. We can, that is, 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 theorem is named after Gauss.

GAUSS' THEOREM.

$$\int_S \boldsymbol{C}\cdot\boldsymbol{n}\,da = \int_V \nabla\cdot\boldsymbol{C}\,dV, \tag{3.18}$$

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 little cube, we would write, following Eq. (3.17),

$$\text{Heat out} = \int_{\text{cube}} \boldsymbol{h} \cdot \boldsymbol{n}\, da = \nabla \cdot \boldsymbol{h}\, \Delta V. \tag{3.19}$$

But this must equal the rate of loss of the heat inside the cube. If q is the heat per unit volume, the heat in the cube is $q\,\Delta V$, and the rate of *loss* is

$$-\frac{d}{dt}(q\,\Delta V) = -\frac{dq}{dt}\,\Delta V. \tag{3.20}$$

Comparing (3.19) and (3.20), we see that

$$-\frac{dq}{dt} = \nabla \cdot \boldsymbol{h}. \tag{3.21}$$

Take careful note of the form of this equation; the form appears often in physics. It expresses a conservation law—here the conservation of heat. We have expressed the same physical fact in another way in Eq. (3.13). Here 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. (3.13) to an infinitesimal cube. We can also go the other way. For a big volume V bounded by S, Gauss' law says that

$$\int_S \boldsymbol{h} \cdot \boldsymbol{n}\, da = \int \nabla \cdot \boldsymbol{h}\, dV. \tag{3.22}$$

Using (3.21), the integral on the right-hand side is found to be just $-dQ/dt$, and again we have Eq. (3.13).

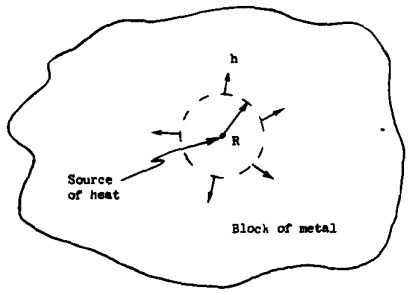


Fig. 3–6. In the region near a point source of heat, the heat flow is radially outward.

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 that there are some wires running into a tiny resistor that is being heated by an electric current. We shall suppose that the heat is generated practically at a point, and let W 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 vector $\boldsymbol{h}$ look like at various places in the metal? How much heat flow is there at each point?

We know that if we integrate the normal component of $\boldsymbol{h}$ over a closed surface that encloses the source, we will always get W. All the heat that is being generated at the point source must flow out through the surface, 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 W. 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 $\boldsymbol{h}$ should be radial if the block of material is large and we don't get too close to the edges, and it should also have the same magnitude at all points on the sphere. You see that we are adding a certain amount of guesswork—usually called "physical intuition"—to our mathematics in order to find the answer.

When $\boldsymbol{h}$ is radial and spherically symmetric, the integral of the normal component of $\boldsymbol{h}$ over the area is very simple, because the normal component is just

the magnitude of $\boldsymbol{h}$ and is constant. The area over which we integrate is $4\pi R^2$. We have then that

$$\int_S \boldsymbol{h} \cdot \boldsymbol{n}\, da = h \cdot 4\pi R^2 \tag{3.23}$$

(where h is the magnitude of $\boldsymbol{h}$). This integral should equal W, the rate at which heat is produced at the source. We get

$$h = \frac{W}{4\pi R^2},$$

or

$$\boldsymbol{h} = \frac{W}{4\pi R^2}\,\boldsymbol{e}_r, \tag{3.24}$$

where, as usual, $\boldsymbol{e}_r$ represents a unit vector in the radial direction. Our result says that $\boldsymbol{h}$ is proportional to W and varies inversely as the square of the distance from the source.

The result we have just obtained applies to 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, keeping only the condition that heat is conserved. We will be dealing only with what happens at places outside of any sources or absorbers of heat.

The differential equation for the conduction of heat was derived in Chapter 2. According to Eq. (2.44),

$$\boldsymbol{h} = -\kappa\, \boldsymbol{\nabla} T. \tag{3.25}$$

(Remember that this relationship is an approximate one, 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 (3.25), we get

$$-\frac{dq}{dt} = \boldsymbol{\nabla} \cdot \boldsymbol{h} = -\,\boldsymbol{\nabla} \cdot (\kappa\, \boldsymbol{\nabla} T),$$

or

$$\frac{dq}{dt} = \kappa\, \boldsymbol{\nabla} \cdot \boldsymbol{\nabla} T = \kappa\, \nabla^2 T, \tag{3.26}$$

if κ is a constant. You remember that q is the amount of heat in a unit volume and $\boldsymbol{\nabla} \cdot \boldsymbol{\nabla} = \nabla^2$ 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 one more 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

$$\Delta q = c_v\, \Delta T$$

or

$$\frac{dq}{dt} = c_v \frac{dT}{dt}. \tag{3.27}$$

The rate of change of heat is proportional to the rate of change of temperature. The constant or proportionality c_v is, here, the specific heat per unit *volume* of the material. Using Eq. (3.27) with (3.26), we get

$$\frac{dT}{dt} = \frac{\kappa}{c_v}\, \nabla^2 T. \tag{3.28}$$

We find that the *time* rate 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 t—for the temperature T.

The differential equation (3.28) is called the *heat diffusion equation.* It is often written as

$$\frac{dT}{dt} = D\nabla^2 T, \tag{3.29}$$

where D is called the *diffusion* constant, and is here equal to κ/c_v.

The diffusion equation appears in many physical problems—in the diffusion of gases, in the diffusion of neutrons, and in 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 the diffusion equation to find how the temperature varies in particular cases. We turn back now to consider other theorems about vector fields.

3–5 The circulation of a vector field

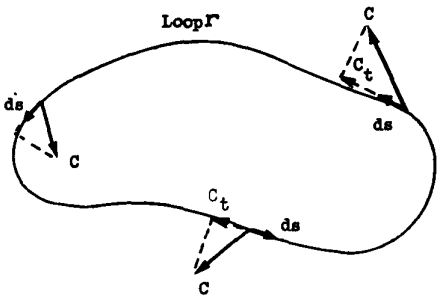


Fig. 3–7. The circulation of C around the curve Γ is the line integral of C_t, the tangential component of C.

We wish now to look at the curl in somewhat the same way we looked at the divergence. We obtained Gauss' 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. How did we 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 apparent equal lack of justification, we shall calculate something else about a vector and show that it is related to the curl. This time we calculate what is called the circulation of a vector field. If $\boldsymbol{C}$ is any vector field, we take its component along a curved line 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 considered a line integral of $\nabla\psi$ earlier in this chapter. Now we do the same kind of thing for *any* vector field $\boldsymbol{C}$.

Let Γ be any closed loop in space—imaginary, of course. An example is given in Fig. 3–7. The line integral of the tangential component of $\boldsymbol{C}$ around the loop is written as

$$\oint_\Gamma C_t\, ds = \oint_\Gamma \boldsymbol{C}\cdot d\boldsymbol{s}. \tag{3.30}$$

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 Γ. The name came originally from considering the circulation of a liquid. But the name—like flux—has been extended to apply to any field even when there is no material "circulating."

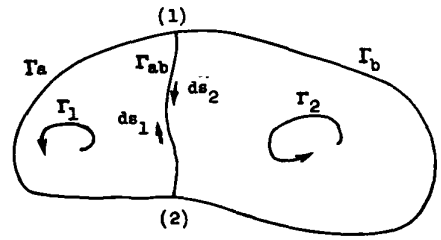


Fig. 3–8. The circulation around the whole loop is the sum of the circulations around the two loops: $\Gamma_1 = \Gamma_a + \Gamma_{ab}$ and $\Gamma_2 = \Gamma_b + \Gamma_{ab}$.

Playing the same kind of game we did with the flux, we can show that the circulation around a loop is the sum of the circulations around two partial 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 some line that cuts across as shown in Fig. 3–8. There are now two loops, Γ_1 and Γ_2. Γ_1 is made up of Γ_a, which is that part of the original curve to the left of (1) and (2), plus Γ_{ab}, the "short cut." Γ_2 is made up of the rest of the original curve plus the short cut.

The circulation around Γ_1 is the sum of an integral along Γ_a and along Γ_{ab}. Similarly, the circulation around Γ_2 is the sum of two parts, one along Γ_b and the other along Γ_{ab}. The integral along Γ_{ab} will have, for the curve Γ_2, the opposite sign from what it has for Γ_1, because the direction of travel is opposite—we must take both our line integrals with the same "sense" of rotation.

Following the same kind of argument we used before, you can see that the sum of the two circulations will give just the line integral around the original curve Γ. The parts due to Γ_{ab} cancel. The circulation around the one part plus the circulation around the second part equals the circulation about the outer 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 on their adjacent portions, so that the sum is equivalent to the circulation around the original single loop.

Now let us suppose that the original loop is the boundary of some surface. There are, of course, an infinite number of surfaces which all have the original loops as the boundary. Our results will not, however, depend on which surface we choose. First, we break our original loop into a number of small loops that 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 Γ by finding the circulations around all of the little squares and then taking their sum.

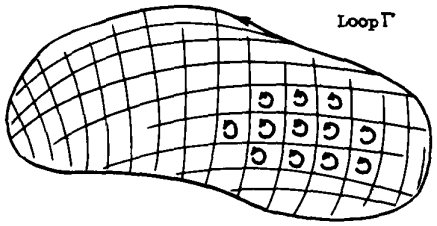


Fig. 3–9. Some surface bounded by the loop Γ is chosen. The surface is divided into a number of small areas, each approximately a square. The circulation around Γ is the sum of the circulations around the little loops.

3–6 The circulation around a square; Stokes' theorem

How shall we find the circulation for each little square? One question is, how is the square oriented in space? We could easily make the calculation if it had a special orientation. For example, if it were in one of the coordinate planes. Since we have not assumed anything as yet about the orientation of the coordinate axes, we can just as well choose the axes so that the one little square we are concentrating on at the moment lies in the xy-plane, as in Fig. 3–10. If our result is expressed in vector notation, 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 $\boldsymbol{C}$ around our little square. It will be easy to do the line integral if we make the square small enough that the vector $\boldsymbol{C}$ doesn't change much along any one side of the square. (The assumption is better the smaller the square, so we are really talking about infinitesimal squares.) Starting at the point (x, y)—the lower left corner of the figure—we go around in 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 $C_y(2)\,\Delta y$. Along the third, we get $-C_x(3)\,\Delta x$, and along the fourth, $-C_y(4)\,\Delta y$. The minus signs are required because we want the tangential component in the direction of travel. The whole line integral is then

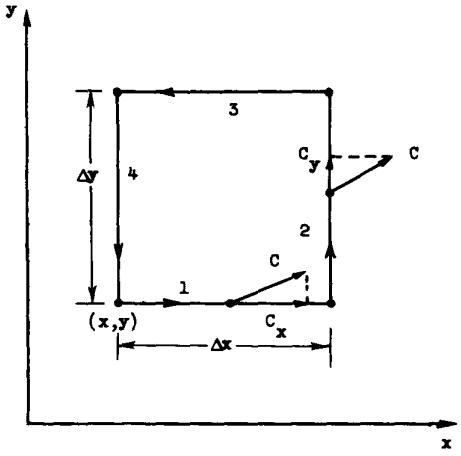


Fig. 3–10. Computing the circulation of C around a small square.

$$\oint \boldsymbol{C} \cdot d\boldsymbol{s} = C_x(1)\,\Delta x + C_y(2)\,\Delta y - C_x(3)\,\Delta x - C_y(4)\,\Delta y. \tag{3.31}$$

Now let's look at the first and third pieces. Together they are

$$[C_x(1) - C_x(3)]\,\Delta x. \tag{3.32}$$

You might think that to 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 of C_x. If we do, we may write

$$C_x(3) = C_x(1) + \frac{\partial C_x}{\partial y}\,\Delta y. \tag{3.33}$$

If we included the next approximation, it would involve terms in $(\Delta y)^2$, but since we will ultimately think of the limit as $\Delta y \to 0$, such terms can be neglected. Putting (3.33) together with (3.32), we find that

$$[C_x(1) - C_x(3)]\,\Delta y = -\frac{\partial C_x}{\partial y}\,\Delta x\,\Delta y. \tag{3.34}$$

The derivative can, to our approximation, be evaluated at (x, y).

Similarly, for the other two terms in the circulation, we may write

$$C_y(2)\,\Delta y - C_y(4)\,\Delta y = \frac{\partial C_y}{\partial x}\,\Delta x\,\Delta y. \tag{3.35}$$

The circulation around our square is then

$$\left(\frac{\partial C_y}{\partial x} - \frac{\partial C_x}{\partial y}\right)\Delta x\,\Delta y, \tag{3.36}$$

which is interesting, because the two terms in the parentheses are just the z-component of the curl. Also, we note that $\Delta x\,\Delta y$ is the area of our square. So we can write our circulation (3.36) as

$$(\nabla \times C)_z\,\Delta a.$$

But the z-component really means the component *normal* to the surface element. We can, therefore, write the circulation around a differential square in an invariant vector form:

$$\oint C \cdot ds = (\nabla \times C)_n\,\Delta a = (\nabla \times C) \cdot n\,\Delta a. \tag{3.37}$$

Our result is: 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 Γ can now be easily related to the curl of the vector field. We fill in the loop with any convenient surface S, as in Fig. 3–11, and add the circulations around 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 Mr. Stokes).

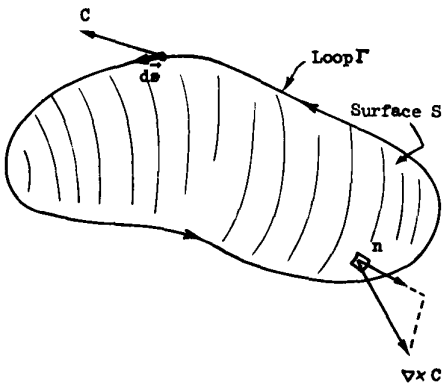


Fig. 3–11. The circulation of C around Γ is the surface integral of the normal component of $\nabla \times C$.

STOKES' THEOREM.

$$\oint_\Gamma C \cdot ds = \int_S (\nabla \times C)_n\,da, \tag{3.38}$$

where S is any surface bounded by Γ.

We must now speak about a convention of signs. In Fig. 3–10 the z-axis would point *toward* you in a "usual"—that is, "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 the other way, we would have gotten the opposite sign. Now how 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. 3–10. 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 Γ, with the fingertips pointed in the direction of the positive sense of ds, then your thumb points in the direction of the *positive* normal to the surface S.

3–7 Curl-free and divergence-free fields

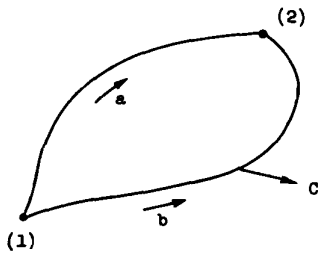


Fig. 3–12. If $\nabla \times C$ is zero, the circulation around the closed curve Γ is zero. The line integral of $C \cdot ds$ from (1) to (2) along a must be the same as the line integral along b.

We would like, now, to consider some consequences of our new theorems. 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 points (1) and (2) on a closed curve (Fig. 3–12), it follows that the line integral of the tangential component from (1) to (2) is independent of which of the two possible paths is taken. We can conclude that the integral from (1) to (2) can depend only on the location of these points—that is to say, it is some function of position only. The same logic was used in Chapter 14 of Vol. I, where we proved that if the integral around a closed loop of some quantity is always zero, then that integral can be represented as the difference of a function of the position of the two ends. This fact allowed us to invent the idea of a potential. We proved, furthermore, that the vector field was the gradient of this potential function (see Eq. 14.13 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 some ψ (psi) for which $C = \nabla\psi$—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 *any* scalar field ϕ (phi). If we take its gradient, $\nabla\phi$, the integral of this vector around any closed loop must be zero. Its line integral from point (1) to point (2) is $[\phi(2) - \phi(1)]$. If (1) and (2)

are the same points, our Theorem 1, Eq. (3.8), tells us that the line integral is zero:

$$\oint_{\text{loop}} \nabla\phi \cdot d\boldsymbol{s} = 0.$$

Using Stokes' theorem, we can conclude that

$$\int \nabla \times (\nabla\phi)\, da = 0$$

over *any* surface. But if the integral is zero over *any* surface, the integrand must be zero. So

$$\nabla \times (\nabla\phi) = 0, \quad \text{always.}$$

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 Γ with a *large* surface S, as indicated in Fig. 3–13. We would like, in fact, to see what happens when the loop shrinks down to a point, so that the surface boundary disappears—the surface becomes closed. Now if the vector $\boldsymbol{C}$ is everywhere finite, the line integral around Γ must go to zero as we shrink the loop—the integral is roughly proportional to the circumference of Γ, which goes to zero. According to Stokes' theorem, the surface integral of $(\nabla \times \boldsymbol{C})_n$ must also vanish. Somehow, as we close the surface we add in contributions that cancel out what was there before. So we have a new theorem:

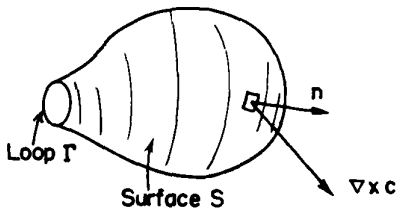


Fig. 3–13. Going to the limit of a closed surface, we find that the surface integral of $(\nabla \times \mathbf{C})_n$ must vanish.

$$\int_{\substack{\text{any closed}\\ \text{surface}}} (\nabla \times \boldsymbol{C})_n\, da = 0. \tag{3.39}$$

Now this is interesting, because we already have a theorem about the surface integral of a vector field. Such a surface integral is equal to the volume integral of the divergence of the vector, according to Gauss' theorem (Eq. 3.18). Gauss' theorem, applied to $\nabla \times \boldsymbol{C}$, says

$$\int_{\substack{\text{closed}\\ \text{surface}}} (\nabla \times \boldsymbol{C})_n\, da = \int_{\substack{\text{volume}\\ \text{inside}}} \nabla \cdot (\nabla \times \boldsymbol{C})\, dV. \tag{3.40}$$

So we conclude that the second integral must also be zero:

$$\int_{\substack{\text{any}\\ \text{volume}}} \nabla \cdot (\nabla \times \boldsymbol{C})\, dV = 0, \tag{3.41}$$

and this is true for any vector field $\boldsymbol{C}$ whatever. Since Eq. (3.41) is true for *any volume*, it must be true that at *every point* in space the integrand is zero. We have

$$\nabla \cdot (\nabla \times \boldsymbol{C}) = 0, \quad \text{always.}$$

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 found about the vector calculus. These are really the salient 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 at all between the first and second points:

$$\psi(2) - \psi(1) = \int_{\substack{(1)\\ \text{any curve}}}^{(2)} \nabla\psi \cdot d\boldsymbol{s}. \tag{3.42}$$

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:

$$\int_{\substack{\text{closed}\\ \text{surface}}} \boldsymbol{C} \cdot \boldsymbol{n}\, da = \int_{\substack{\text{volume}\\ \text{inside}}} \nabla \cdot \boldsymbol{C}\, dV. \tag{3.43}$$

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.

$$\int_{\text{boundary}} \boldsymbol{C} \cdot d\boldsymbol{s} = \int_{\text{surface}} (\nabla \times \boldsymbol{C}) \cdot \boldsymbol{n}\, da. \tag{3.44}$$

4

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:

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{4.1}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{4.2}$$

$$c^2 \nabla \times \boldsymbol{B} = \frac{\partial \boldsymbol{E}}{\partial t} + \frac{\boldsymbol{j}}{\epsilon_0}, \tag{4.3}$$

$$\nabla \cdot \boldsymbol{B} = 0. \tag{4.4}$$

The situations that are described by these equations can be very complicated. We will consider first relatively simple situations, and learn how to handle them before we take up more complicated ones. The easiest circumstance to treat is one in which nothing depends on the time—called the *static* case. All charges are permanently fixed in space, or if they do move, they move as a steady flow in a circuit (so ρ and $\boldsymbol{j}$ are constant in time). In these circumstances, all of the terms in the Maxwell equations which are time derivatives of the field are zero. In this case, the Maxwell equations become:

Electrostatics:

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{4.5}$$

$$\nabla \times \boldsymbol{E} = 0. \tag{4.6}$$

Magnetostatics:

$$\nabla \times \boldsymbol{B} = \frac{\boldsymbol{j}}{\epsilon_0 c^2}, \tag{4.7}$$

$$\nabla \cdot \boldsymbol{B} = 0. \tag{4.8}$$

You will notice an interesting thing about this set of four equations. It can be separated into two pairs. The electric field $\boldsymbol{E}$ appears only in the first two, and the magnetic field $\boldsymbol{B}$ appears only in the second two. The two fields are not interconnected. This means that *electricity and magnetism are distinct phenomena so long as charges and currents are static.* The interdependence of $\boldsymbol{E}$ and $\boldsymbol{B}$ does not appear until there are changes in charges or currents, as when a condensor 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 $\boldsymbol{E}$ and $\boldsymbol{B}$ 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 ideal from the point of view of learning about the mathematical properties of vector fields. Electrostatics is a neat example of a vector field with *zero curl* and a *given divergence.* Magnetostatics is a neat example of a field with *zero divergence* and a *given curl.* The more conventional—and you may be thinking, more satisfactory—way of presenting

4–1 Statics

4–2 Coulomb's law; superposition

4–3 Electric potential

4–4 $\boldsymbol{E} = -\nabla\phi$

4–5 The flux of $\boldsymbol{E}$

4–6 Gauss' law; the divergence of $\boldsymbol{E}$

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*

$$\epsilon_0 c^2 = \frac{10^7}{4\pi}$$

$$\frac{1}{4\pi\epsilon_0} \approx 9 \times 10^9$$

$$[\epsilon_0] = \text{coulomb}^2/\text{newton-meter}^2$$

the theory of electromagnetism is to start first with electrostatics and thus to learn about the divergence. Magnetostatics and the curl are taken up later. Finally, electricity and magnetism are put together. We have chosen to start with the complete theory of vector calculus. Now we shall apply it to the special case of electrostatics, the field of E given by the first pair of equations.

We will begin with the simplest situations—ones in which the positions of all charges are specified. If we had only to study electrostatics at this level (as we shall do in the next two chapters), life would be very simple—in fact, almost trivial. Everything can be obtained 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 that the charges take up depend on the E field, which in turn depends on the positions of the charges. Then things can get quite complicated. If, for instance, a charged body is brought near a conductor or insulator, the electrons and protons in the conductor or insulator will move around. The charge density ρ in Eq. (4.5) may have one part that we know about, from the charge that we brought up; but there will be other parts from charges that have moved around in the conductor. And all of the charges must be taken 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 ones.

4–2 Coulomb's law; superposition

It would be logical to use Eqs. (4.5) and (4.6) as our starting points. It will be easier, however, if we start somewhere else and come back to these equations. The results will be equivalent. We will start with a law that we have talked about before, called 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.

Coulomb's law:

$$\boldsymbol{F}_1 = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r_{12}^2}\boldsymbol{e}_{12} = -\boldsymbol{F}_2. \tag{4.9}$$

$\boldsymbol{F}_1$ is the force *on* charge q_1, $\boldsymbol{e}_{12}$ is the unit vector in the direction *to* q_1 *from* q_2, and r_{12} is the distance between q_1 and q_2. The force $\boldsymbol{F}_2$ on q_2 is equal and opposite to $\boldsymbol{F}_1$.

The constant of proportionality, for historical reasons, is written as $1/4\pi\epsilon_0$. In the system of units which we use—the mks system—it is defined as exactly 10^{-7} 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·meter^2 per coulomb^2 or volt·meter per coulomb.

$$\begin{aligned}\frac{1}{4\pi\epsilon_0} &= 10^{-7}c^2 \quad \text{(by definition)}\\ &= 9.0 \times 10^9 \text{ (by experiment).}\end{aligned} \tag{4.10}$$

Unit: newton·meter^2/coulomb^2,

or volt·meter/coulomb.

When there are more than two charges present—the only really interesting times—we must supplement Coulomb's law with one other fact of nature: the force on any charge is the vector sum of the Coulomb forces from each of the other charges. This fact is called "the principle of superposition." That's all there is to electrostatics. If we combine the Coulomb law and the principle of superposition, there is nothing else. Equations (4.5) and (4.6)—the electrostatic equations—say no more and no less.

When applying Coulomb's law, it is convenient to introduce the idea of an electric field. We say that the field $\boldsymbol{E}(1)$ is the force *per unit charge* on q_1 (due to all other charges). Dividing Eq. (4.9) by q_1, we have, for one other charge besides q_1,

$$\boldsymbol{E}(1) = \frac{1}{4\pi\epsilon_0}\frac{q_2}{r_{12}^2}\boldsymbol{e}_{12}. \tag{4.11}$$

Also, we consider that $\boldsymbol{E}(1)$ describes something about the point (1) even if q_1 were not there—assuming that all other charges keep their same positions. We say: $\boldsymbol{E}(1)$ is the electric field *at* the point (1).

The electric field $\boldsymbol{E}$ is a vector, so by Eq. (4.11) we really mean three equations —one for each component. Writing out explicitly the x-component, Eq. (4.11) means

$$E_x(x_1, y_1, z_1) = \frac{q_2}{4\pi\epsilon_0}\frac{x_1 - x_2}{[(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2]^{3/2}}, \tag{4.12}$$

and similarly for the other components.

If there are many charges present, the field $\boldsymbol{E}$ at any point (1) is a sum of the contributions from each of the other charges. Each term of the sum will look like (4.11) or (4.12). Letting q_j be the magnitude of the jth charge, and $\boldsymbol{r}_{1j}$ the displacement from q_j to the point (1), we write

$$\boldsymbol{E}(1) = \sum_j \frac{1}{4\pi\epsilon_0}\frac{q_j}{r_{1j}^2}\boldsymbol{e}_{1j}. \tag{4.13}$$

Which means, of course,

$$E_x(x_1, y_1, z_1) = \sum_j \frac{1}{4\pi\epsilon_0}\frac{q_j(x_1 - x_j)}{[(x_1 - x_j)^2 + (y_1 - y_j)^2 + (z_1 - z_j)^2]^{3/2}}, \tag{4.14}$$

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 a continuous smear —or in a "distribution," as it is called. This is O.K. so long as we are not interested in what is happening on too small a scale. We describe a charge distribution by the "charge density," $\rho(x, y, z)$. If the amount of charge in a small volume ΔV_2 located at the point (2) is Δq_2, then ρ is defined by

$$\Delta q_2 = \rho(2)\,\Delta V_2. \tag{4.15}$$

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

$$\boldsymbol{E}(1) = \frac{1}{4\pi\epsilon_0}\int_{\substack{\text{all}\\\text{space}}} \frac{\rho(2)\boldsymbol{e}_{12}\,dV_2}{r_{12}^2}. \tag{4.16}$$

Some people prefer to write

$$\boldsymbol{e}_{12} = \frac{\boldsymbol{r}_{12}}{r_{12}},$$

where $\boldsymbol{r}_{12}$ is the vector displacement *to* (1) *from* (2), as shown in Fig. 4–1. The integral for $\boldsymbol{E}$ is then written as

$$\boldsymbol{E}(1) = \frac{1}{4\pi\epsilon_0}\int_{\substack{\text{all}\\\text{space}}} \frac{\rho(2)\boldsymbol{r}_{12}\,dV_2}{r_{12}^3}. \tag{4.17}$$

When we want to calculate something with these integrals, we usually have to write them out in explicit detail. For the x-component of either Eq. (4.16) or (4.17), we would have

$$E_x(x_1, y_1, z_1) = \int_{\substack{\text{all}\\\text{space}}} \frac{(x_1 - x_2)\rho(x_2, y_2, z_2)\,dx_2\,dy_2\,dz_2}{4\pi\epsilon_0[(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2]^{3/2}}. \tag{4.18}$$

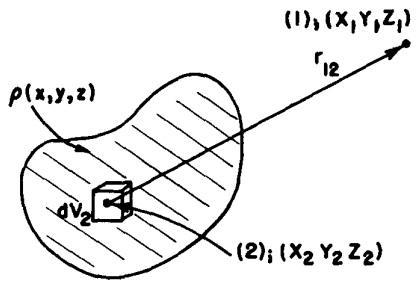


Fig. 4–1. The electric field $\boldsymbol{E}$ at point (1), from a charge distribution, is obtained from an integral over the distribution. Point (1) could also be inside the distribution.

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 electrostatic problems in which we know the locations of all of the charges. Given the charges, what are the fields? *Answer:* Do this 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 fields produced by a sheet of charge, from a line of charge, from a 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 merely a matter of it being sometimes easier to do an integral by some clever guesswork than by actually carrying it out. The guesswork requires learning all kinds of strange things. In practice, it might be easier to forget trying to be clever and always to do the integral directly instead of being so 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 take up 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 about how much work it would take 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,

$$W = -\int_a^b \boldsymbol{F} \cdot d\boldsymbol{s},$$

where $\boldsymbol{F}$ is the electrical force *on* the charge at each point, and $d\boldsymbol{s}$ is the differential vector displacement along the path. (See Fig. 4–2.)

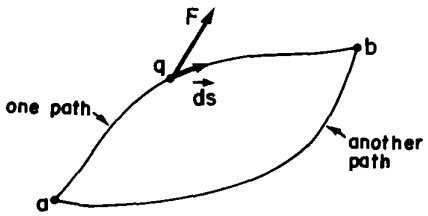


Fig. 4–2. The work done in carrying a charge from a to b is the negative of the integral of $\boldsymbol{F} \cdot d\boldsymbol{s}$ along the path taken.

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 W(unit), we write

$$W(\text{unit}) = -\int_a^b \boldsymbol{E} \cdot d\boldsymbol{s}. \tag{4.19}$$

Now, in general, what we get with this kind of an integral depends on the path we take. But 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 to b along the path for which W is smaller and *back* along the other, getting *out* more work than we put *in*.

There is nothing impossible, in principle, about getting energy out of a field. We shall, in fact, encounter fields where it is possible. It could be that as you move a charge you produce forces on the other part of the "machinery." If the "machinery" moved against the force it would lose energy, thereby keeping the total energy in the world constant. For *electrostatics*, however, there is no such "machinery." We know what the forces back on the sources of the field are. They are the Coulomb forces on the charges responsible for the field. If the other charges are fixed in position—as we assume in *electrostatics* only—these back forces can do no work on them. 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 show that it must follow from Coulomb's law of force.

We consider first what 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

be one unit, from a to b. Let's start with the easiest possible path to calculate. We carry our test charge first along the arc of a circle, then along a radius, as shown in part (a) of Fig. 4–3. Now on that particular path it is child's play 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 a'. The field is radial (from Coulomb's law), so it is at right angles to the direction of motion. Next, on the path from a' to b, the field is in the direction of motion and varies as $1/r^2$. Thus the work done on the test charge in carrying it from a to b would be

$$-\int_a^b \boldsymbol{E}\cdot d\boldsymbol{s} = -\frac{q}{4\pi\epsilon_0}\int_{a'}^{b}\frac{dr}{r^2} = -\frac{q}{4\pi\epsilon_0}\left(\frac{1}{r_a}-\frac{1}{r_b}\right). \tag{4.20}$$

Now let's take another easy path. For instance, the one shown in part (b) of Fig. 4–3. It goes for awhile along an arc of a circle, then radially for awhile, then along an arc again, then radially, and so on. 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_a to $r_{a'}$, then along the next radial stretch from $r_{a'}$ to $r_{a''}$, and so on. The sum of all these integrals is the same as a single integral directly from r_a to r_b. We get the same answer for this path that we did for the first path we tried. It is clear that we would 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? Would 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 unit charge from a to b is independent of the path.

$$\left.\begin{matrix} W(\text{unit}) \\ a \to b \end{matrix}\right\} = -\int_{\substack{a \\ \text{any} \\ \text{path}}}^{b} \boldsymbol{E}\cdot d\boldsymbol{s}.$$

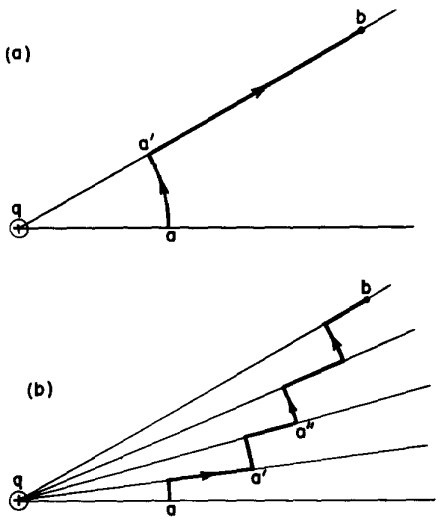


Fig. 4–3. In carrying a test charge from a to b the same work is done along either path.

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 goes *by way of* point P_0. Let $\phi(a)$ stand for the work done against the field in going *from* P_0 to point a, and let $\phi(b)$ be the work done in going *from* P_0 to point b (Fig. 4–4). The work in going *to* P_0 from a (on the way to b) is the negative of $\phi(a)$, so we have that

$$-\int_a^b \boldsymbol{E}\cdot d\boldsymbol{s} = \phi(b) - \phi(a). \tag{4.21}$$

Since only the difference in the function ϕ at two 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.

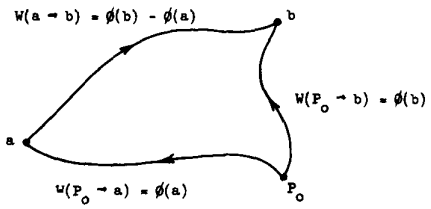


Fig. 4–4. The work done in going along any path from a to b is the negative of the work from some point P_0 to a plus the work from P_0 to b.

Electrostatic potential:

$$\phi(P) = -\int_{P_0}^{P} \boldsymbol{E}\cdot d\boldsymbol{s}. \tag{4.22}$$

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):

$$\phi(x, y, z) = \frac{q}{4\pi\epsilon_0}\frac{1}{r}. \tag{4.23}$$

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

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 charges, we can get the complete formulas for the potential ϕ at a point we call (1):

$$\phi(1) = \sum_j \frac{1}{4\pi\epsilon_0}\frac{q_j}{r_{1j}}, \tag{4.24}$$

$$\phi(1) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(2)\,dV_2}{r_{12}}. \tag{4.25}$$

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 point in space from some reference point.

4–4 $\boldsymbol{E} = -\nabla\phi$

Who cares about ϕ? Forces on charges are given by $\boldsymbol{E}$, the electric field. The point is that $\boldsymbol{E}$ can be obtained easily from ϕ—it is as 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 unit 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:

$$\Delta W = \phi(x + \Delta x, y, z) - \phi(x, y, z) = \frac{\partial\phi}{\partial x}\Delta x.$$

But the work done against the field for the same path is

$$\Delta W = -\int \boldsymbol{E}\cdot d\boldsymbol{s} = -E_x\,\Delta x.$$

We see that

$$E_x = -\frac{\partial\phi}{\partial x}. \tag{4.26}$$

Similarly, $E_y = -\partial\phi/\partial y$, $E_z = -\partial\phi/\partial z$, or, summarizing with the notation of vector analysis,

$$\boldsymbol{E} = -\nabla\phi. \tag{4.27}$$

This equation is the differential form of Eq. (4.22). Any problem with specified 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

$$\int_a^b \nabla\phi\cdot d\boldsymbol{s} = \phi(b) - \phi(a). \tag{4.28}$$

According to Eq. (4.25) the scalar potential ϕ is given by a three-dimensional integral similar to the one we had for $\boldsymbol{E}$. Is there any advantage to computing ϕ rather than $\boldsymbol{E}$? Yes. There is only one integral for ϕ, while there are three integrals for $\boldsymbol{E}$—because it is a vector. Furthermore, $1/r$ is usually a little easier to integrate than x/r^3. 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 $\boldsymbol{E}$. It is merely a practical matter.

There is also a deeper physical significance to the potential ϕ. We have shown that $\boldsymbol{E}$ of Coulomb's law is obtained from $\boldsymbol{E} = -\text{grad}\,\phi$, when ϕ is given by (4.22). But if $\boldsymbol{E}$ is equal to the gradient of a scalar field, then we know from the vector calculus that the curl of $\boldsymbol{E}$ must vanish:

$$\nabla\times\boldsymbol{E} = 0. \tag{4.29}$$

But that is just our second fundamental equation of electrostatics, Eq. (4.6). We have shown that Coulomb's law gives an $\boldsymbol{E}$ field that staisfies that condition. So far, everything is all right.

We had really proved that $\nabla \times \boldsymbol{E}$ was zero before we defined the potential. We had shown that the work done around a closed path is zero. That is, that

$$\oint \boldsymbol{E} \cdot d\boldsymbol{s} = 0$$

for *any* path. We saw in Chapter 3 that for any such field $\nabla \times \boldsymbol{E}$ 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 $\nabla \times \boldsymbol{E}$ is zero in a different way—by computing the components of $\nabla \times \boldsymbol{E}$ for the field of a point charge, as given by Eq. (4.11). 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 *radial* force the work done is independent of the path, and there exists 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 force from a single charge was radial and spherically symmetric. It did not depend on the fact that the dependence on distance was as $1/r^2$—there could have been any r dependence. The existence of a potential, and the fact that the curl of $\boldsymbol{E}$ 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.

4–5 The flux of $\boldsymbol{E}$

We will now derive a field equation that depends specifically and directly 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 "only natural," because "that's the way things spread out." Take a light source with light streaming out: the amount of light that passes through a surface cut out by a cone with its apex at the source is the same no matter at what radius the surface is placed. It must be so if there is to be conservation of light energy. The amount of light per unit area—the intensity—must vary inversely as the area cut by the cone, i.e., inversely as the square of the distance from the source. Certainly the electric field should vary inversely as the square of the distance for the same reason! But there is no such thing as the "same reason" here. Nobody 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 conserved, that none could ever disappear once it was shot out of a charge, 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 physical idea. If the electric field *were* like conserved bullets going out, then it would vary inversely as the square of the distance and we would be able to describe that behavior by an equation—which is purely mathematical. Now there is no harm in thinking this way, so long as we do not say that the electric field *is made* out of bullets, but realize that we are using a model to help us find the right mathematics.

Suppose, indeed, 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 whatever it is flows out of a charge into the space around. If $\boldsymbol{E}$ were the vector of such a flow (as $\boldsymbol{h}$ is for heat flow), it would have a $1/r^2$ dependence near a point source. Now we wish to use this model to find out how to state the inverse square law in a deeper or more abstract way, rather than simply saying "inverse square." (You may wonder why we should want to avoid the direct statement of such a simple law, and want instead to imply the same thing sneakily in a different way. Patience! It will turn out to be useful.)

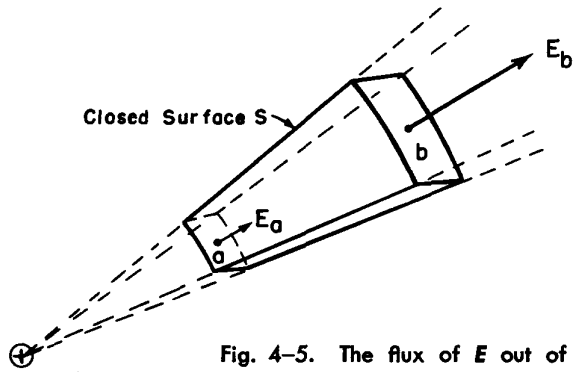


Fig. 4–5. The flux of E out of the surface S is zero.

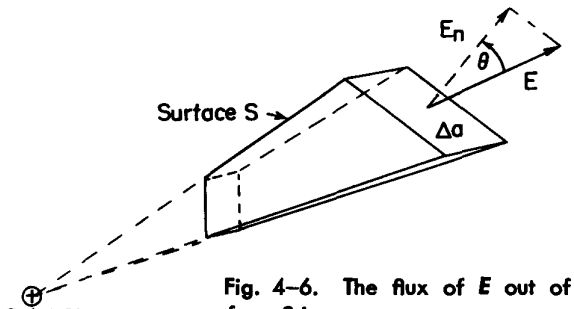


Fig. 4–6. The flux of E out of the surface S is zero.

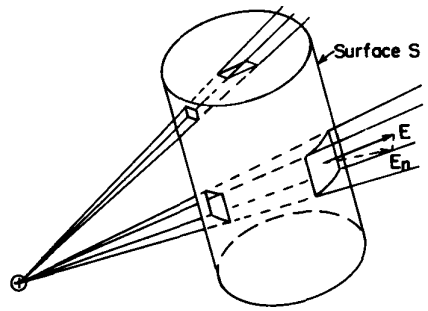


Fig. 4–7. Any volume can be thought of as completely made up of infinitesimal truncated cones. The flux of E from one end of each conical segment is equal and opposite to the flux from the other end. The total flux from the surface S is therefore zero.

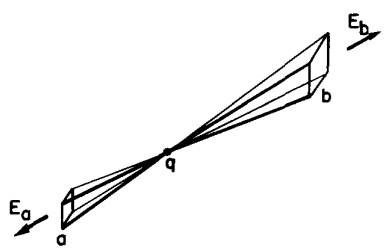


Fig. 4–8. If a charge is inside a surface, the flux out is not zero.

We ask: What is the "flow" 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 a flow, the net flow out of this box should be zero. That is what we get if by the "flow" 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—minus for the smaller face and plus for the larger face. The magnitude of E decreases 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 cancelled by the flux out of face b. The total flow out of S is zero, which is to say that for this surface

$$\int_S E_n \, da = 0. \tag{4.30}$$

Next we show that the two end surfaces may be tilted with respect to the radial line without changing the integral (4.30). 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, an infinitesimal angle. In Fig. 4–6 we show a surface S whose "sides" are radial, but whose "ends" are tilted. The end surfaces are not small in the figure, but you are to imagine the situation for very small end surfaces. Then 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 $E_n \, \Delta a$ is unchanged. The flux out of the whole surface S is still zero.

Now it is easy to see that the flux out of a volume enclosed by *any* 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 since the fluxes in and out of these end surfaces cancel by pairs, the total flux out of the surface will be zero. The idea is illustrated in Fig. 4–7. We have the completely general result that the total flux of E out of *any* surface S in the field of a point charge is zero.

But notice! 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 areas 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 arguments as before—only now they have the *same* sign. The flux out of a surface that *surrounds* 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.

We can choose any shape we wish 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 $\boldsymbol{E}$ everywhere on its surface is

$$\frac{1}{4\pi\epsilon_0}\frac{q}{r^2},$$

and is directed always normal to the surface. We find the total flux through S' if we multiply this normal component of $\boldsymbol{E}$ by the surface area:

$$\text{Flux through the suface } S' = \left(\frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\right)(4\pi r^2) = \frac{q}{\epsilon_0}, \tag{4.31}$$

a number independent of the radius of the sphere! We know then that the flux outward through S is also q/ϵ_0—a value independent of the shape of S so long as the charge q is inside.

We can write our conclusions as follows:

$$\int_{\text{any surface } S} E_n\, da = \begin{cases} 0; & q \text{ outside } S \\ \dfrac{q}{\epsilon_0}; & q \text{ inside } S \end{cases} \tag{4.32}$$

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 enclosed in a surface, whatever size and shape it is, the number of bullets passing through is the same—it is given by the rate at which bullets are generated at the gun. It all seems quite reasonable for conserved bullets. But does the model tell us anything more than we get simply by writing Eq. (4.32)? No one has succeeded in making these "bullets" do anything else but produce this one law. After that, they produce nothing but errors. That is why today we prefer to represent the electromagnetic field purely abstractly.

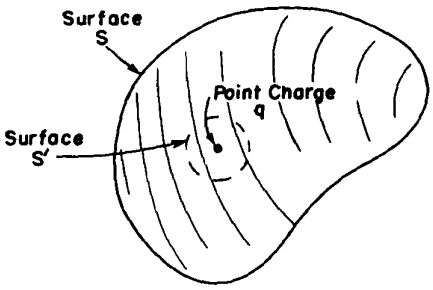


Fig. 4–9. The flux through S is the same as the flux through S'.

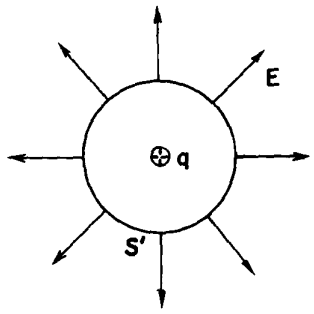


Fig. 4–10. The flux through a spherical surface containing a point charge q is q/ϵ_0.

4–6 Gauss' law; the divergence of E

Our nice result, Eq. (4.32), was proved for a single point charge. Now suppose that 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 field due to both charges. That is, if $\boldsymbol{E}_1$ represents the electric field that would have been produced by q_1 alone, and $\boldsymbol{E}_2$ represents the electric field produced by q_2 alone, the total electric field is $\boldsymbol{E} = \boldsymbol{E}_1 + \boldsymbol{E}_2$. The flux through any closed surface S is

$$\int_S (E_{1n} + E_{2n})\, da = \int_S E_{1n}\, da + \int_S E_{2n}\, da. \tag{4.33}$$

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 q_1 is inside S but q_2 is outside, then the first integral gives q_1/ϵ_0 and the second integral gives zero. If the surface encloses both charges, each will give its contribution and we have that the flux is $(q_1 + q_2)/\epsilon_0$. The general rule is clearly that 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.

Gauss' law:

$$\int_{\substack{\text{any closed}\\ \text{surface } S}} E_n\, da = \frac{\text{sum of charges inside}}{\epsilon_0}, \tag{4.34}$$

or

$$\int_{\substack{\text{any closed}\\ \text{surface } S}} \boldsymbol{E}\cdot\boldsymbol{n}\, da = \frac{Q_{\text{int}}}{\epsilon_0}, \tag{4.35}$$

where

$$Q_{\text{int}} = \sum_{\text{inside } S} q_i. \tag{4.36}$$

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 $\rho\, dV$. The sum over all charges is then the integral

$$Q_{\text{int}} = \int_{\substack{\text{volume}\\ \text{inside } S}} \rho\, dV. \tag{4.37}$$

From our derivation you see that Gauss' law follows from the fact that the exponent in Coulomb's law is exactly two. A $1/r^3$ 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 quite equivalent so long as we keep in mind the rule that the forces 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 showed in Chapter 3 that the flux of $\boldsymbol{E}$ out of such a cube is $\nabla \cdot \boldsymbol{E}$ times the volume dV of the cube. The charge inside of dV, by the definition of ρ, is equal to $\rho\, dV$, so Gauss' law gives

$$\nabla \cdot \boldsymbol{E}\, dV = \frac{\rho\, dV}{\epsilon_0},$$

or

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}. \tag{4.38}$$

The differential form of Gauss' law is the first of our fundamental field equations of electrostatics, Eq. (4.5). 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 consider one example of the use of Gauss' law. (We will come later to many more examples.)

4–7 Field of a sphere of charge

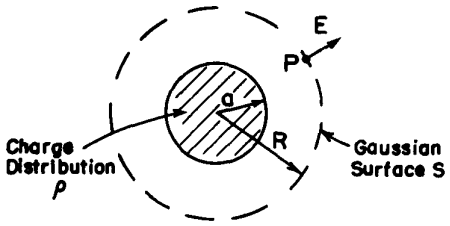


Fig. 4–11. Using Gauss' law to find the field of a uniform sphere of charge.

One of the difficult problems we had when we studied the theory of gravitational attractions was to prove that the force produced by a solid sphere of matter was the same at the surface of the sphere as it would be if all the matter were concentrated at the center. For many years Newton didn't make public his theory of gravitation, because he couldn't be sure this theorem was true. We 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 most simple fashion. Only this time we will prove the corresponding theorem for a uniform sphere of electrical charge. (Since the laws of electrostatics are the same as those of gravitation, the same proof could be done for the gravitational field.)

We ask: What is the electric field $\boldsymbol{E}$ at a point P anywhere outside the surface of a sphere filled with a uniform distribution of charge? Since there is no "special" direction, we can assume that $\boldsymbol{E}$ 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

$$\int E_n\, da = E \cdot 4\pi R^2.$$

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

$$E = \frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}, \tag{4.39}$$

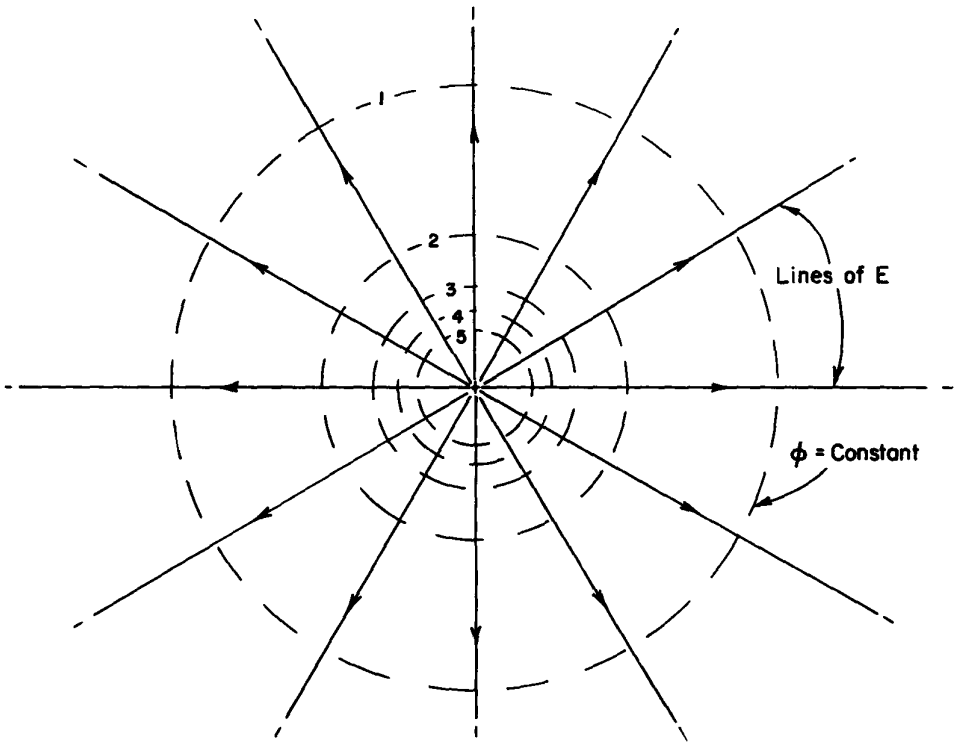


Fig. 4–12. Field lines and equipotential surfaces for a positive point charge.

which is the same formula we would have for a point charge Q. We have proved Newton's problem more easily than by doing the integral. It is, of course, a false kind of easiness—it has taken you some time to be able to understand Gauss' law, so you may think that no time has really been saved. But after 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 electrostatic field. The two laws of electrostatics, one that the flux 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 direction of the electric vector. But we also wish to represent the magnitude of the vector. We can make the rule that the strength of the electric field will be represented by the "density" of the lines. By the density of the lines we mean the number of lines per unit area through a surface perpendicular to the lines. With these two rules we can have 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 perpendicular to the lines at any radius r *increases* as r^2, so if we always keep the same *number* of lines for *all* distances from the charge, the *density* will remain in proportion to the magnitude of the field. We can guarantee that there are the same number of lines at every distance if we insist that the lines be *continuous*—that once a line is started 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 which *leave* a charge q must be equal to q/ϵ_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 what is the geometri-

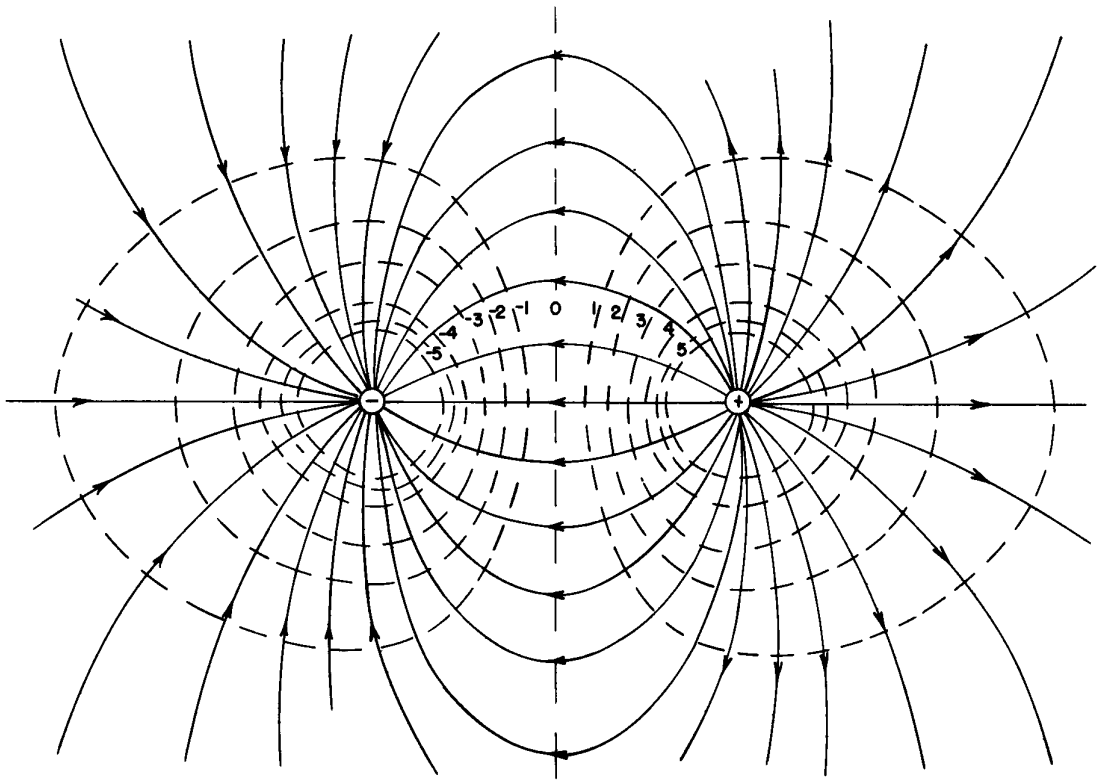


Fig. 4–13. Field lines and equipotentials for two equal and opposite point charges.

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 $\boldsymbol{E}$ were *not* perpendicular to the surface, it would have a component *in* the surface. The potential would be changing in the surface, but then it wouldn't be an equipotential. The equipotential surfaces must then be everywhere at right angles to the electric field lines.

For a point charge all by itself, the equipotential surfaces are spheres centered at the charge. We have shown in Fig. 4–12 the intersection 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 pictures like Fig. 4–12 and superimpose them—impossible! Then we would have field lines crossing each other, and that's not possible, because $\boldsymbol{E}$ can't have *two* directions at the same point. The disadvantage of the field-line picture is now evident. By geometrical arguments it is impossible to analyze in a very simple way where the new 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, in 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 (and opposite) charges. If we calculate the fields from Eq. (4.13) and the potentials from (4.23), we can draw the field lines and equipotentials. Figure 4–13 shows the result. But we first had to solve the problem mathematically!

A Note about Units

Quantity	*Unit*
F	newton
Q	coulomb
L	meter
W	joule
$\rho \sim Q/L^3$	coulomb/meter^3
$1/\epsilon_0 \sim FL^2/Q^2$	newton·meter^2/coulomb^2
$E \sim F/Q$	newton/coulomb
$\phi \sim W/Q$	joule/coulomb = volt
$E \sim \phi/L$	volt/meter
$1/\epsilon_0 \sim EL^2/Q$	volt·meter/coulomb

5

Application of Gauss' Law

5–1 Electrostatics is 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 line 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 fields of a conductor

5–10 The field in a cavity of a conductor

5–1 Electrostatics is 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—$\boldsymbol{E}$ is a gradient. From these two laws, all the predictions of electrostatics follow. But to say these things mathematically is one thing; to use them easily, and 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 directly. 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 too. So when we use Gauss' law for the solution of particular problems, we will have to add something to it. We will have to presuppose, 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 other charges? 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, for a charge to be in equilibrium at any particular point P_0, 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_0 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_0. But that is in violation of Gauss' law if there is no charge at P_0, as we can easily see.

Consider a tiny imaginary surface that encloses P_0, as in Fig. 5–1. If the electric field everywhere in the vicinity is pointed toward P_0, 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. But Gauss' law says that the flux of electric field through any surface is proportional to the total charge inside. If there is no charge at P_0, the field we have imagined violates Gauss' law. It is impossible to balance a positive charge in empty space—at a point where there is not 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 than electrical forces!

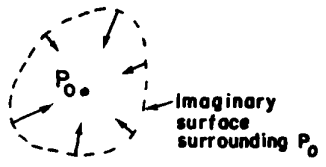


Fig. 5–1. If P_0 were a position of stable equilibrium for a positive charge, the electric field everywhere in the neighborhood would point toward P_0.

Our result has been obtained for a point charge. Does the same conclusion hold for a complicated arrangement of charges 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 restoring for displacements in every direction.

Call $\boldsymbol{F}$ the total force on the rod in any position—$\boldsymbol{F}$ is then a vector field. Following the argument used above, we conclude that at a position of stable equilibrium, the divergence of $\boldsymbol{F}$ 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:

$$\boldsymbol{F} = q_1\boldsymbol{E}_1 + q_2\boldsymbol{E}_2. \tag{5.1}$$

The divergence of $\boldsymbol{F}$ is given by

$$\nabla \cdot \boldsymbol{F} = q_1 (\nabla \cdot \boldsymbol{E}_1) + q_2 (\nabla \cdot \boldsymbol{E}_2).$$

If each of the two charges q_1 and q_2 is in free space, both $\nabla \cdot \boldsymbol{E}_1$ and $\nabla \cdot \boldsymbol{E}_2$ are zero, and $\nabla \cdot \boldsymbol{F}$ is zero—not negative, as would be required for equilibrium. You can see that an extension of the argument shows that no rigid combination of any number of charges can have a position of stable equilibrium in an electrostatic field in free space.

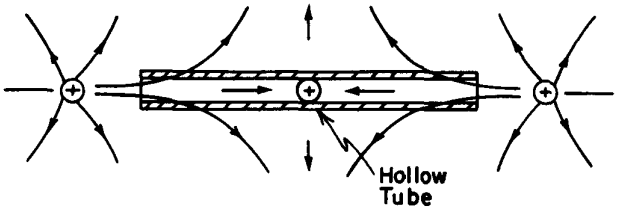


Fig. 5–2. A charge can be in equilibrium if there are mechanical constraints.

Now we have not shown that equilibrium is forbidden if there are pivots or other mechanical constraints. As an example, consider a hollow tube in which a charge can move back and forth freely, but not sideways. Now it is very easy to devise an electric field that points inward at both ends of the tube if it is allowed that the field may point laterally outward near the center of the tube. We simply place positive charges at each end of the tube, as in Fig. 5–2. There can now be an equilibrium point even though the divergence of $\boldsymbol{E}$ is zero. The charge, of course, would not be in stable equilibrium for sideways motion were it not for "non-electrical" forces from the tube walls.

5–3 Equilibrium with conductors

There is no stable 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 stable 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 can move freely around in them. Perhaps when the point charge is displaced slightly, the other charges on the conductors will move in a way that will give a restoring force to the point charge? The answer is still no—although the proof we have just given doesn't show 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 to heat 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 force is *away* from P_0). Any readjustment of the charges on the conductors can only lower the potential energy still more, so (by the principle of virtual work) their motion will only *increase* the force in that particular direction away from P_0, and not reverse 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 locations or the sizes of the supporting charges with suitable devices. You know that a rod standing on its point in a gravitational field is unstable, but this does not prove that it cannot be balanced on the end of a finger. Similarly, a charge can be held in one spot by electric fields if they are *variable*. But not with a passive—that is, a *static*—system.

5–4 Stability of atoms

If charges cannot be held stably in position, it is surely not proper to imagine matter to be made up of static *point* charges (electrons and protons) governed only by the laws of electrostatics. Such a static configuration is impossible; it would collapse!

It was once suggested that the positive charge of an atom could be distributed uniformly in a sphere, and the negative charges, the electrons, could be at rest inside the positive charge, as shown in Fig. 5–3. This was the first atomic model, proposed by Thompson. But Rutherford concluded from the experiment of Geiger and Marsden that the positive charges were very much concentrated, in what he called the nucleus. Thompson'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 in 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 the kinetic energy required to stay in orbit, and would spiral in toward the nucleus. Again unstable!

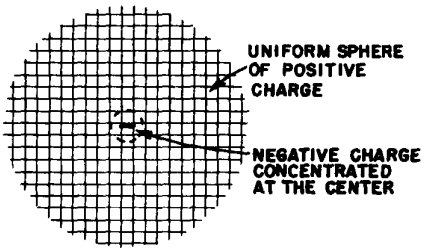


Fig. 5–3. The Thompson model of an atom.

The stability of the atoms 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 spread out in space over a distance given by the uncertainty principle. If it were confined in too small a space, it would have a great uncertainty in momentum. But that means that it would have a high expected energy—which it would use to escape from the electrical attraction. The net result is an electrical equilibrium not too different from the idea of Thompson —only it is the *negative* charge that is spread out (because the mass of the electron is so much smaller than the mass of the proton).

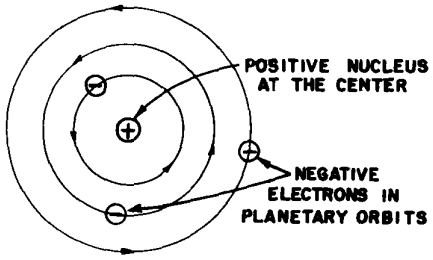


Fig. 5–4. The Rutherford-Bohr model of an atom.

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 planar 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 rod. By this we mean that electric charges are distributed uniformly along an 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 without integrating, by using Gauss' law and some guesswork. First, we surmise 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 axial component from charges on the other side. The result could 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 obvious. (It may not be easy to prove, but it is true if space is symmetric—as we believe it is.)

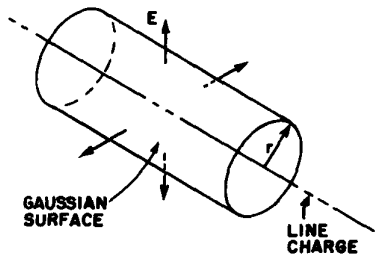


Fig. 5–5. A cylindrical gaussian surface coaxial with a line charge.

We can use Gauss' law in the following way. We consider an *imaginary* surface in the shape of a cylinder coaxial with the line, as shown in Fig. 5–5. According to 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 normal to the surface, the normal component is the magnitude of the field. Let's call it E. Also, let the radius of the cylinder be r, and its length be taken as one unit, for convenience. The flux through the cylindrical surface is equal to E times the area of the surface, which is $2\pi r$. The flux through the two end faces is zero because the electric field is tan-

gential to them. The total charge inside our surface is just λ, because the length of the line inside is one unit. Gauss' law then gives

$$E \cdot 2\pi r = \lambda/\epsilon_0,$$

$$E = \frac{\lambda}{2\pi\epsilon_0 r}. \tag{5.2}$$

The electric field of a line charge depends inversely on the *first* power of the distance from the line.

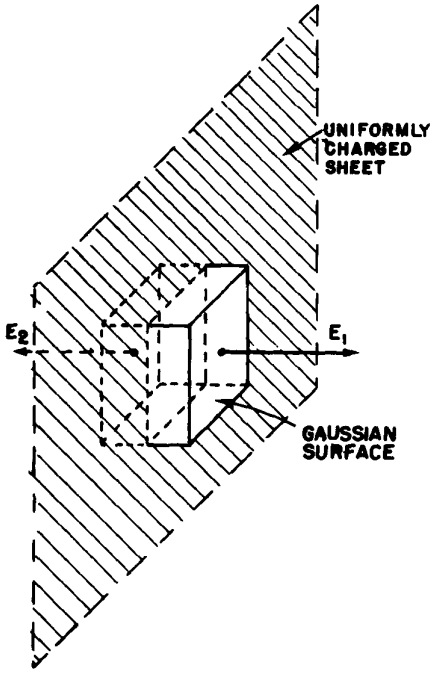


Fig. 5–6. The electric field near a uniformly charged sheet can be found by applying Gauss' law to an imaginary box.

5–6 A sheet of charge; two sheets

As another example, we will 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 we have no field from any other charges in the world*, the fields 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 first 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 flux to the charge inside, we have

$$EA + EA = \frac{\sigma A}{\epsilon_0},$$

from which

$$E = \frac{\sigma}{2\epsilon_0}, \tag{5.3}$$

a simple but important result.

You may remember that the same result was obtained in an earlier chapter by an integration over the entire surface. Gauss' law gives us the answer, in this instance, 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 that

$$E_1 + E_2 = \frac{\sigma}{\epsilon_0}, \tag{5.4}$$

where E_1 and E_2 are the fields directed outward on each side of the sheet.

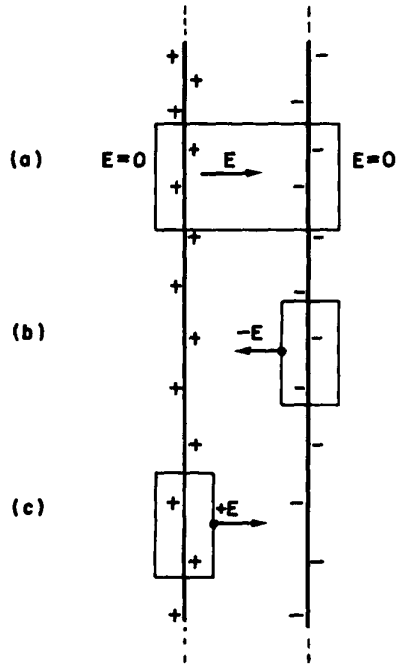


Fig. 5–7. The field between two charged sheets is σ/ϵ_0.

The problem of two parallel sheets with equal and opposite charge densities, $+\sigma$ and $-\sigma$, is equally simple if we assume again that the outside world is quite symmetric. Either by superposing two solutions for a single sheet or by constructing a gaussian box that includes both 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 or the other, 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

$$E \text{ (between the sheets)} = \sigma/\epsilon_0, \tag{5.5}$$

$$E \text{ (outside)} = 0. \tag{5.6}$$

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 forces, spread nearly uniformly throughout the body of the nucleus.

Suppose that we have a sphere of radius R filled uniformly with 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 the volume inside times ρ, or

$$\tfrac{4}{3}\pi r^2 \rho.$$

Using Gauss' law, it follows that the magnitude of the field is given by

$$E = \frac{\rho r}{3\epsilon_0} \qquad (r < R). \tag{5.7}$$

You can see that this 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 thin spherical shell of charge. Assuming that the field is everywhere radial and is spherically symmetric, one gets immediately from Gauss' law that the field outside the shell is like that of a point charge, while the field everywhere inside the shell is zero. (A gaussian surface inside the shell will contain no charge.)

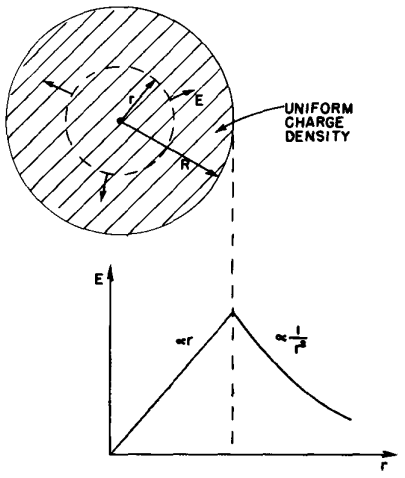


Fig. 5–8. Gauss' law can be used to find the field inside a uniformly charged sphere.

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 square of the distance. Consider any point P inside a uniform spherical shell of charge. Imagine a small cone whose apex is at P and which extends to the surface of the sphere, where it cuts out a small surface area Δa_1, as in Fig. 5–9. An exactly symmetric cone diverging from the opposite side of P would cut out the surface area Δa_2. If the distances from P to these two elements of area are r_1 and r_2, the areas are in the ratio

$$\frac{\Delta a_2}{\Delta a_1} = \frac{r_2^2}{r_1^2}.$$

(You can show this by geometry for any point P inside the sphere.)

If the surface of the sphere is uniformly charged, the charge Δq on each of the elements of area is proportional to the area, so

$$\frac{\Delta q_2}{\Delta q_1} = \frac{\Delta a_2}{\Delta a_1}.$$

Coulomb's law then says that the magnitudes of the fields produced at P by these two surface elements are in the ratio

$$\frac{E_2}{E_1} = \frac{q_2/r_2^2}{q_1/r_1^2} = 1.$$

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 not exactly two.

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 exactly zero. For instance, if the force varied more rapidly, like, say, the inverse cube of r, that portion of the surface which is nearer to an interior point would produce a field which is larger than that which is farther away, resulting in a radial inward field for a positive surface

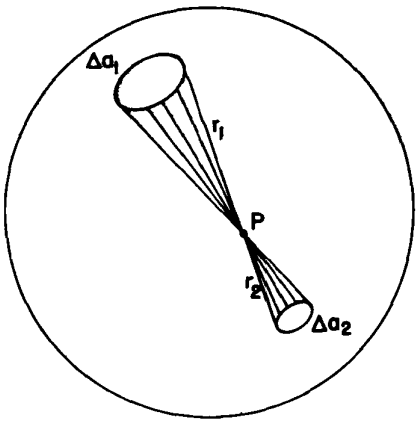


Fig. 5–9. The field is zero at any point P inside a spherical shell of charge.

charge. These conclusions suggest an elegant way of finding out whether the inverse square law is precisely correct. We need only determine whether or not the field inside of a uniformly charged spherical shell is precisely zero.

It is lucky 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 to an ideal inverse square. Such comparisons of things that are equal, or nearly so, are usually the bases of the most precise physical measurements.

How shall we observe 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 we touch a small metal ball to a charged object and then touch it to an electrometer the meter will become charged and the pointer will move from zero (Fig. 5–10a). The ball picks up charge because there are electric fields outside the charged sphere that cause charges to run onto (or off) 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 carried to the electrometer. 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.

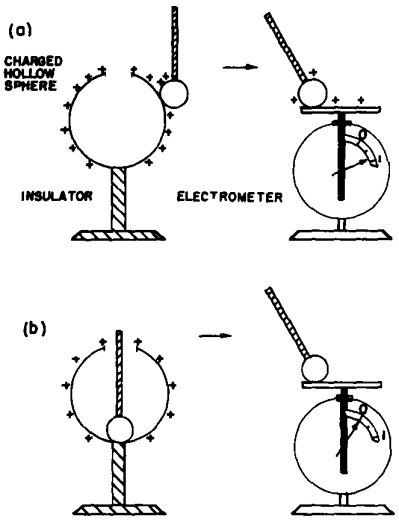


Fig. 5–10. The electric field is zero inside a closed conducting shell.

It appears that Benjamin Franklin was the first to notice 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 until 18 years later, and Gauss' law came even later still.

Gauss' law has been checked carefully by putting an electrometer inside a large sphere and observing whether any deflections occur when the sphere is charged to a high voltage. A null result is always obtained. Knowing the geometry of the apparatus and the sensitivity of the meter, it is possible to compute the minimum 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 $r^{-2+\epsilon}$, we can place an upper bound on ϵ. By this method Maxwell determined that ϵ was less than 1/10,000. The experiment was repeated and improved upon in 1936 by Plimpton and Laughton. 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 law to be in various circumstances? The experiments we just described measure the dependence of the field on distance for distances of some tens of centimeters. 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 formulation 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 which varies inversely with the first power of the distance. How accurately is the exponent known for such small distances? As a result of very careful measurements in 1947 by Lamb and Retherford on the relative positions of the energy levels of hydrogen, we know that the exponent is correct again to one part in a billion on the atomic scale—that is, at distances of the order of one angstrom (10^{-8} centimeter).

The accuracy of the Lamb-Retherford measurement was possible again because of a physical "accident." Two of the states of a hydrogen atom are expected to have almost indentical energies *only* if the potential varies exactly as $1/r$. A measurement was made of the very slight *difference* in energies by finding

the frequency ω of the photons that are emitted or absorbed in the transition from one state to the other, using for the energy difference $\Delta E = \hbar\omega$. Computations showed that ΔE would have been noticeably different from what was observed if the exponent in the force law $1/r^2$ differed from 2 by as much as one part in a billion.

Is the same exponent correct at still shorter distances? From measurements in nuclear physics it is found that there are electrostatic forces at typical nuclear distances—at about 10^{-13} centimeter—and that they still vary approximately as the inverse square. We shall look at some of the evidence in a later chapter. Coulomb's law is, we know, still valid, at least to some extent, at distances of the order of 10^{-13} centimeter.

How about 10^{-14} centimeter? This range can be investigated by bombarding protons with very energetic electrons and observing how they are scattered. Results to date seem to indicate that the law fails at these distances. The electrical force seems to be about 10 times too weak at distances less than 10^{-14} centimeter. Now there are two possible explanations. One is that the Coulomb law does not work at such small distances; the other is that our objects, the electrons and protons, are not point charges. Perhaps either the electron or proton, or both, is some kind of a smear. Most physicists prefer to think 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 act—on the average—like a little sphere of positive charge. We know that the field from a sphere of charge does not vary as $1/r^2$ all the way into the center. It is quite likely that the proton charge is smeared, but the theory of pions is still quite incomplete, so it may also be that Coulomb's law fails at very small distances. The 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 yes; at least to an accuracy of 15 parts in a million.

We go back now to an important matter that we slighted when we spoke of the experimental verification of Gauss' law. You may have wondered how the experiment of Maxwell or of Plimpton and Laughton could give such an accuracy unless the spherical conductor they used was a perfect sphere. An accuracy of one part in a billion is really something to achieve, and you might well ask whether they could make a sphere which was that precise. There are certain to be slight irregularities in any real sphere and if there are irregularities, will they not produce fields inside? We wish to show now that it is not necessary to have a perfect sphere. It is possible, in fact, to show that there is no field inside a closed conducting shell of *any* shape. In other words, the experiments depended on $1/r^2$, but had nothing to do with the surface being a sphere (except that with a sphere it is easier to calculate what the fields *would* be if Coulomb had been wrong), so we take up that subject now. To show this, it is necessary to know some of the properties of electrical conductors.

5–9 The fields of a conductor

An electrical conductor is a solid that contains many "free" electrons. The electrons can move around freely *in* the material, but cannot leave the surface. In a metal there are so many free electrons that any electric field will set large numbers of them into motion. Either the current of electrons so set up must be continually kept moving by external sources of energy, or the motion of the electrons will cease as they discharge the sources producing the initial field. In "electrostatic" situations, we do not consider continuous sources of current (they will be considered later when we study magnetostatics), so the electrons move only until they have arranged themselves to produce zero electric field everywhere inside the conductor. (This usually happens in a small fraction of a second.) If there were any field left, this field would urge still more electrons to move; the only electrostatic solution is that the field is everywhere zero inside.

Now consider the *interior* of a charged conducting object. (By "interior" we mean in the *metal* itself.) Since the metal is a conductor, the interior field must

be zero, and so the gradient of the potential ϕ is zero. That means that ϕ does not vary from point to point. Every conductor is an equipotential *region*, and its surface is an equipotential surface. Since in a conducting material the electric field is everywhere zero, the divergence of $\boldsymbol{E}$ 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 at 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 charge is put on, or *in*, 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 are no forces preventing that. Saying it another way: we know that the electric field lines must always go at right angles to an equipotential surface.

We can also, using Gauss' law, relate the field strength just outside a conductor to the local density of the charge at the surface. For a gaussian surface, we take a small cylindrical box half inside and half outside the surface, like the one shown in Fig. 5–11. There is a contribution to the total flux of $\boldsymbol{E}$ only from the side of the box outside the conductor. The field just outside the surface of a conductor is then

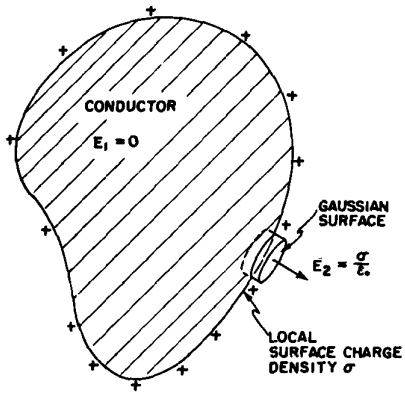


Fig. 5–11. The electric field just outside the surface of a conductor is proportional to the local surface density of charge.

Outside a conductor:

$$E = \frac{\sigma}{\epsilon_0}, \tag{5.8}$$

where σ is the *local* surface charge density.

Why does a sheet of charge on a conductor produce a different field than *just* a sheet of charge? In other words, why is (5.8) twice as large as (5.3)? The reason, of course, is that we have *not* said for the conductor that there are no "other" charges around. There must, in fact, be some to make $\boldsymbol{E} = 0$ in the conductor. The charges in the immediate neighborhood of a point P on the surface do, in fact, give a field $E_{\text{local}} = \sigma_{\text{local}}/2\epsilon_0$ both inside and outside the surface. But all the rest of the charges on the conductor "conspire" to produce an additional field at the point P equal in magnitude to E_{local}. The total field inside goes to zero and the field outside to $2E_{\text{local}} = \sigma/\epsilon_0$.

5–10 The field in a cavity of a conductor

We return now to the problem of the hollow container—a conductor with a cavity. There is no field in the *metal*, 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 conductor or the cavity—say for the 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, so there is no flux through S and 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. There *could* be a positive surface charge on one part and a negative one somewhere else, as indicated in Fig. 5–12. Such a thing cannot be ruled out by Gauss' law.

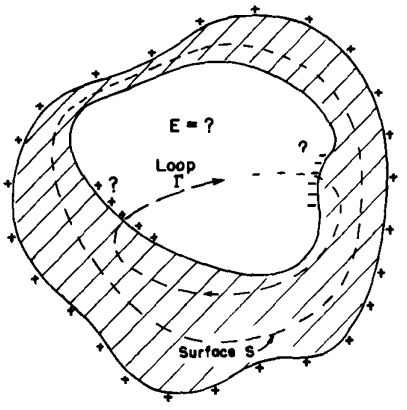


Fig. 5–12. What is the field in an empty cavity of a conductor, for any shape?

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 law that the circulation of $\boldsymbol{E}$ is always zero (electrostatics). 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 lines of $\boldsymbol{E}$ would have to start on the

positive charges and end on the negative charges (since we are considering only the case that there are no free charges in the cavity). Now imagine 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 charges would not be zero. The integral through the metal is zero, since $\boldsymbol{E} = 0$. So we would have

$$\oint \boldsymbol{E} \cdot d\boldsymbol{s} \neq 0???$$

But the line integral of $\boldsymbol{E}$ around any closed loop in an electrostatic field is always zero. So there can 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 "inside an *empty*" cavity. If some charges are *placed* at some fixed locations in the cavity—as on an insulator or on a small conductor insulated from the main one—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. The same arguments 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 such a great precision. The shape of the hollow shell used doesn't matter. It doesn't need to be spherical; it could be square! If Gauss' law is exact, 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.

6

The Electric Field in Various Circumstances

6–1 Equations of the electrostatic potential

6–2 The electric 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 method of images

6–8 A point charge near a conducting plane

6–9 A point charge near a conducting sphere

6–10 Condensers; parallel plates

6–11 High-voltage breakdown

6–12 The field emission microscope

Review: Chapter 23, Vol. I, *Resonance*

6–1 Equations of the electrostatic potential

This chapter will describe the behavior of the electric field in a number of different circumstances. It will provide some experience with the way the electric field behaves, and will describe some of the mathematical methods which are used to find this field.

We begin by pointing out that the whole mathematical problem is the solution of two equations, the Maxwell equations for electrostatics:

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}, \tag{6.1}$$

$$\nabla \times \boldsymbol{E} = 0. \tag{6.2}$$

In fact, the two can be combined into a single equation. From the second equation, we know at once that we can describe the field as the gradient of a scalar (see Section 3–7):

$$\boldsymbol{E} = -\nabla\phi. \tag{6.3}$$

We may, if we wish, completely describe any particular electric field in terms of its potential ϕ. We obtain the differential equation that ϕ must obey by substituting Eq. (6.3) into (6.1), to get

$$\nabla \cdot \nabla\phi = -\frac{\rho}{\epsilon_0}. \tag{6.4}$$

The divergence of the gradient of ϕ is the same as ∇^2 operating on ϕ:

$$\nabla \cdot \nabla\phi = \nabla^2\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2}, \tag{6.5}$$

so we write Eq. (6.4) as

$$\nabla^2\phi = -\frac{\rho}{\epsilon_0}. \tag{6.6}$$

The operator ∇^2 is called the Laplacian, and Eq (6 6) is called the Poisson equation. The entire subject of electrostatics, from a mathematical point of view, is merely a study of the solutions of the single equation (6.6). Once ϕ is obtained by solving Eq. (6.6) we can find $\boldsymbol{E}$ immediately from Eq. (6.3).

We take up first the special class of problems in which ρ is given as a function of x, y, z. In that case the problem is almost trivial, for we already know the solution of Eq. (6.6) for the general case. We have shown that if ρ is known at every point, the potential at point (1) is

$$\phi(1) = \int \frac{\rho(2)\, dV_2}{4\pi\epsilon_0 r_{12}}, \tag{6.7}$$

where $\rho(2)$ is the charge density, dV_2 is the volume element at point (2), and r_{12} is the distance between points (1) and (2). The solution of the *differential* equation (6.6) is reduced to an *integration* over space. The solution (6.7) should be especially noted, because there are many situations in physics that lead to equations like

$$\nabla^2 (\text{something}) = (\text{something else}),$$

and Eq. (6.7) is a prototype of the solution for any of these problems.

The solution of electrostatic field problems is thus completely straightforward when the positions of all the charges are known. Let's see how it works in a few examples.

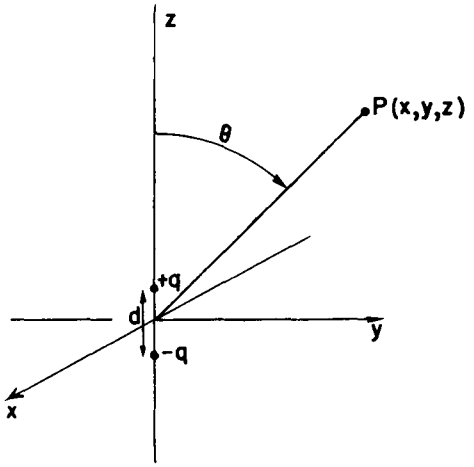


Fig. 6–1. A dipole: two charges $+q$ and $-q$ the distance d apart.

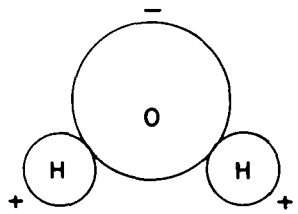


Fig. 6–2. The water molecule H_2O. The hydrogen atoms have slightly less than their share of the electron cloud; the oxygen, slightly more.

6–2 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 ask about 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 on one side and a little more positive charge on the other. This arrangement is certainly not as simple as two point charges, but when seen from 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 terms of (6.8) in a power series in the small quantity d (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,

$$\frac{1}{\sqrt{[z + (d/2)]^2 + x^2 + y^2}} \approx \frac{1}{r}\left(1 - \frac{1}{2}\frac{zd}{r^2}\right).$$

The difference of these two terms gives for the potential

$$\phi(x, y, z) = \frac{1}{4\pi\epsilon_0}\frac{z}{r^3}\,qd. \tag{6.9}$$

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 momentum!):

$$p = qd. \tag{6.10}$$

Equation (6.9) can also be written as

$$\phi(x, y, z) = \frac{1}{4\pi\epsilon_0}\frac{p\cos\theta}{r^2}, \tag{6.11}$$

since $z/r = \cos\theta$, 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 (whereas for a point charge it goes as $1/r$). The electric field $\boldsymbol{E}$ of the dipole will then decrease as $1/r^3$.

We can put our formula into a vector form if we define $\boldsymbol{p}$ as a vector whose magnitude is p and whose direction is along the axis of the dipole, pointing from q_- toward q_+. Then

$$\cos\theta = \boldsymbol{p}\cdot\boldsymbol{e}_r, \tag{6.12}$$

where $\boldsymbol{e}_r$ is the unit radial vector (Fig. 6–3). We can also represent the point (x, y, z) by $\boldsymbol{r}$. Then

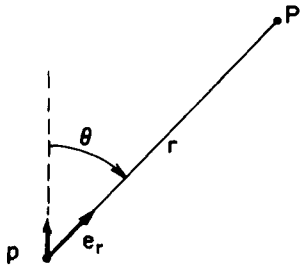


Fig. 6–3. Vector notation for a dipole.

Dipole potential:

$$\phi(r) = \frac{1}{4\pi\epsilon_0}\frac{\boldsymbol{p}\cdot\boldsymbol{e}_r}{r^2} = \frac{1}{4\pi\epsilon_0}\frac{\boldsymbol{p}\cdot\boldsymbol{r}}{r^3}. \tag{6.13}$$

This formula is valid for a dipole with any orientation and position if $\boldsymbol{r}$ 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 z-component of the field is $-\partial\phi/\partial z$. For a dipole oriented along the z-axis we can use (6.9):

$$-\frac{\partial\phi}{\partial z} = -\frac{p}{4\pi\epsilon_0}\frac{\partial}{\partial z}\left(\frac{z}{r^3}\right) = -\frac{p}{4\pi\epsilon_0}\left(\frac{1}{r^3} - \frac{3z^2}{r^5}\right),$$

or

$$E_z = \frac{p}{4\pi\epsilon_0}\frac{3\cos^2\theta - 1}{r^3}. \tag{6.14}$$

The x- and y-components are

$$E_x = \frac{p}{4\pi\epsilon_0}\frac{3zx}{r^5}, \qquad E_y = \frac{p}{4\pi\epsilon_0}\frac{3zy}{r^5}.$$

These two can be combined to give one component directed *perpendicular* to the z-axis, which we will call the transverse component $E_\perp$:

$$E_\perp = \sqrt{E_x^2 + E_y^2} = \frac{p}{4\pi\epsilon_0}\frac{3z}{r^5}\sqrt{x^2 + y^2}$$

or

$$E_\perp = \frac{p}{4\pi\epsilon_0}\frac{3\cos\theta\sin\theta}{r^3}. \tag{6.15}$$

The transverse component $E_\perp$ is in the x-y plane and points directly away from the *axis* of the dipole. The total field, of course, is

$$E = \sqrt{E_z^2 + E_\perp^2}\,.$$

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 $\theta = 90°$. At both of these special angles the electric field has only a z-component, but of opposite sign at the two places (Fig. 6–4).

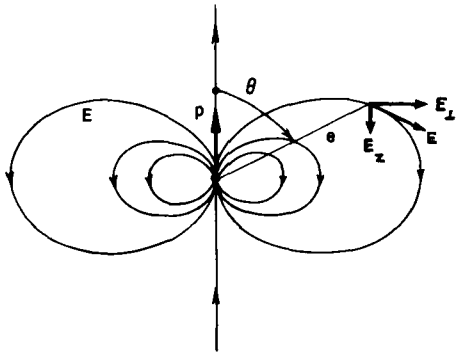


Fig. 6–4. The electric field of a dipole.

6–3 Remarks on vector equations

This is a good place to make a general remark about vector analysis. The fundamental proofs can be expressed by elegant 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 equations in vector form so that they would no longer depend on any particular coordinate system. After that, we are allowed to choose any coordinate system we wish, knowing that the relation is, in general, true. It clearly doesn't make any sense to bother with an arbitrary coordinate system at some complicated angle when you can choose a neat system for the particular problem—provided that the result can finally be expressed as a vector equation. So by all means 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 $\nabla \cdot \boldsymbol{E}$ and wondering what it is, don't forget that it can always be spread out 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 defeat 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 going on. 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 cleverness 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 print.

6–4 The dipole potential as a gradient

We would like to point out a rather amusing thing about the dipole formula, Eq. (6.13). The potential can also be written as

$$\phi = -\frac{1}{4\pi\epsilon_0}\,\boldsymbol{p} \cdot \nabla\left(\frac{1}{r}\right). \tag{6.16}$$

If you calculate the gradient of $1/r$, you get

$$\nabla\left(\frac{1}{r}\right) = -\frac{\boldsymbol{r}}{r^3} = -\frac{\boldsymbol{e}_r}{r^2}\,,$$

and Eq. (6.16) is the same as Eq. (6.13).

How did we think of that? We just remembered that $\boldsymbol{e}_r/r^2$ 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 $1/4\pi\epsilon_0$ while we make these arguments; we can stick it in at the end.) Now if we move the charge $+q$ up a distance Δz, the potential at P will change a little, by, say, $\Delta\phi_+$. How much is $\Delta\phi_+$? Well, it is just the amount that the potential *would* change if we were to *leave* the charge at the origin and move P *downward* by the same distance Δz (Fig. 6–5). That is,

$$\Delta\phi_+ = -\frac{\partial\phi_0}{\partial z}\Delta z,$$

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

$$\phi_+ = \frac{q}{r} - \frac{\partial}{\partial z}\left(\frac{q}{r}\right)\frac{d}{2}. \tag{6.17}$$

Applying the same reasoning for the potential from the negative charge, we can write

$$\phi_- = \frac{-q}{r} + \frac{\partial}{\partial z}\left(\frac{-q}{r}\right)\frac{d}{2}. \tag{6.18}$$

The total potential is the sum of (6.17) and (6.18):

$$\phi = \phi_+ + \phi_- = -\frac{\partial}{\partial z}\left(\frac{q}{r}\right)d \tag{6.19}$$

$$= -\frac{\partial}{\partial z}\left(\frac{1}{r}\right)qd.$$

For other orientation of the dipole, we could represent the displacement of the positive charge by the vector $\Delta\boldsymbol{r}_+$. We should then write Eq. (6.17) as

$$\Delta\phi_+ = -\nabla\phi_0 \cdot \Delta\boldsymbol{r}_+,$$

where $\Delta\boldsymbol{r}$ is then to be replaced by $\boldsymbol{d}/2$. Completing the derivation as before, Eq. (6.19) would then become

$$\phi = -\nabla\left(\frac{1}{r}\right)\cdot q\boldsymbol{d}.$$

This is the same as Eq. (6.16), if we replace $q\boldsymbol{d} = \boldsymbol{p}$, 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

$$\phi = -\boldsymbol{p}\cdot\nabla\Phi_0, \tag{6.20}$$

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,

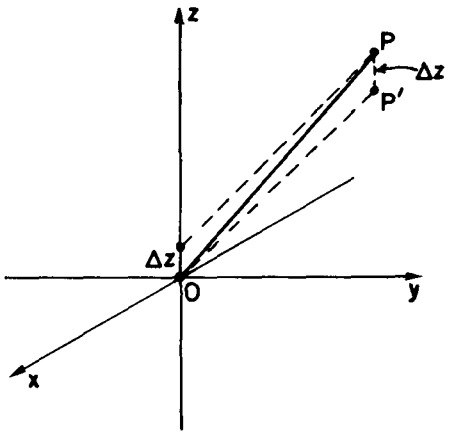


Fig. 6–5. The potential at P from a point charge at Δz above the origin is the same as the potential at P'(Δz below P) from the same charge at the origin.

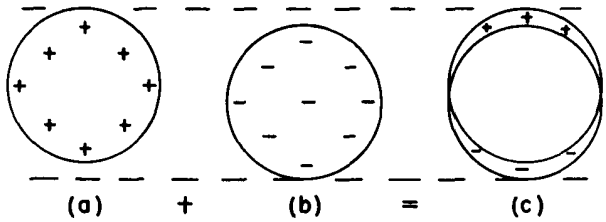


Fig. 6–6. Two uniformly charged spheres, superposed with a slight displacement, are equivalent to a nonuniform distribution of surface charge.

originally superposed to make a neutral—that is, uncharged—sphere. If the positive sphere is then displaced slightly with respect to the negative sphere, the body of the uncharged sphere would remain neutral, but a little positive charge will appear on one side, and some negative charge will appear on the opposite side, as illustrated in Fig. 6–6. If the relative displacement of the two spheres is small, the net charge is equivalent to a surface charge (on a spherical surface), and the surface charge density will be proportional to the cosine of the polar angle.

Now if we want the potential from this distribution, we do not need to do an integral. We know that the potential from each of the spheres of charge is—for points outside the sphere—the same as from a point charge. The two displaced spheres are like two point charges; the potential is just that of a dipole.

In this way you can show that a charge distribution on a sphere of radius a with a surface charge density

$$\sigma = \sigma_0 \cos\theta$$

produces a field outside the sphere which is just that of a dipole whose moment is

$$p = \frac{4\pi\sigma_0 a^3}{3}.$$

It can also be shown that inside the sphere the field is constant, with the value

$$E = \frac{\sigma_0}{3\epsilon_0}.$$

If θ is the angle from the positive z-axis, the electric field inside the sphere is in the *negative* z-direction. The example we have just considered is not as artificial as it may appear; we will encounter it again in the theory of dielectrics.

6–5 The dipole approximation for an arbitrary distribution

The dipole field appears in another circumstance both interesting and important. Suppose that we have an object that has a complicated distribution of charge—like the water molecule (Fig. 6–2)—and we are interested only in the fields far away. We will show that it is possible to find a relatively simple expression for the fields which is appropriate for distances large compared with the size of the object.

We can think of our object as an assembly of point charges q_i in a certain limited region, as shown in Fig. 6–7. (We can, later, replace q_i by $\rho\, dV$ if we wish.) Let each charge q_i be located at the displacement $\boldsymbol{d}_i$ from an origin chosen somewhere

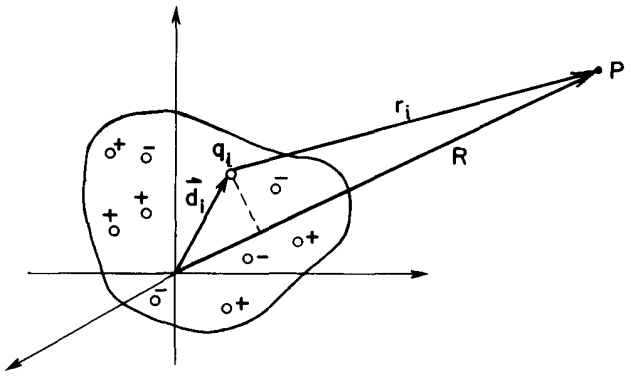


Fig. 6–7. Computation of the potential at a point P at a large distance from a set of charges.

in the middle of the group of charges. What is the potential at the point P, located at $\boldsymbol{R}$, where $\boldsymbol{R}$ is much larger than the maximum $\boldsymbol{d}_i$? The potential from the whole collection is given by

$$\phi = \frac{1}{4\pi\epsilon_0} \sum_i \frac{q_i}{r_i}, \tag{6.21}$$

where r_i is the distance from P to the charge q_i (the length of the vector $\boldsymbol{R} - \boldsymbol{d}_i$). Now if the distance from the charges to P, the point of observation, is enormous, each of the r_i's can be approximated by R. Each 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

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{1}{R} \sum q_i = \frac{Q}{4\pi\epsilon_0 R}, \tag{6.22}$$

where Q is just the total charge of the whole object. Thus we find that for points far enough from any lump of charge, the lump 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. This is not an unusual case; in fact, as we know, 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 effects of the separate charges. We need a better approximation than (6.22) for the potential from an arbitrary distribution of charge in a neutral object. Equation (6.21) is still precise, but we can no longer just set $r_i = R$. We need a more accurate expression for r_i. If the point P is at a large distance, r_i will differ from R to an excellent approximation by the projection of $\boldsymbol{d}$ on $\boldsymbol{R}$, as can be seen from Fig. 6–7. (You should imagine that P is really farther away than is shown in the figure.) In other words, if $\boldsymbol{e}_r$ is the unit vector in the direction of $\boldsymbol{R}$, then our next approximation to r_i is

$$r_i \approx R - \boldsymbol{d}_i \cdot \boldsymbol{e}_r. \tag{6.23}$$

What we really want is $1/r_i$, which, since $d_i \ll R$, can be written to our approximation as

$$\frac{1}{r_i} \approx \frac{1}{R}\left(1 + \frac{\boldsymbol{d}_i \cdot \boldsymbol{e}_r}{R}\right). \tag{6.24}$$

Substituting this in (6.21), we get that the potential is

$$\phi = \frac{1}{4\pi\epsilon_0}\left(\frac{Q}{R} + \sum_i q_i \frac{\boldsymbol{d}_i \cdot \boldsymbol{e}_r}{R^2} + \cdots\right). \tag{6.25}$$

The three dots indicate the terms of higher order in d/R that we have neglected. These, as well as the ones we have already obtained, are successive terms in a Taylor expansion of $1/r_i$ about $1/R$ in powers of d_i/R.

The first term in (6.25) is what we got before; it drops out if the object is neutral. The second term depends on $1/R^2$, just as for a dipole. In fact, if we *define*

$$\boldsymbol{p} = \sum q_i \boldsymbol{d}_i \tag{6.26}$$

as a property of the charge distribution, the second term of the potential (6.25) is

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{\boldsymbol{p} \cdot \boldsymbol{e}_r}{R^2}, \tag{6.27}$$

precisely a dipole potential. The quantity $\boldsymbol{p}$ is called the dipole moment of the distribution. It is a generalization of our earlier definition, and reduces to it for the special case of two point charges.

Our result is that, far enough away from *any* mess of charges that is as a whole neutral, the potential is a dipole potential. It decreases as $1/R^2$ and varies as $\cos\theta$—and its strength depends on the dipole moment of the distribution of charge. It is for these reasons that dipole fields are important, since the simple case of a pair of point charges is quite rare.

The water molecule, for example, 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, obtaining another term in the potential which decreases as $1/R^3$, and 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 integrations. We turn now to an entirely new kind of problem, the determination of the fields near charged conductors.

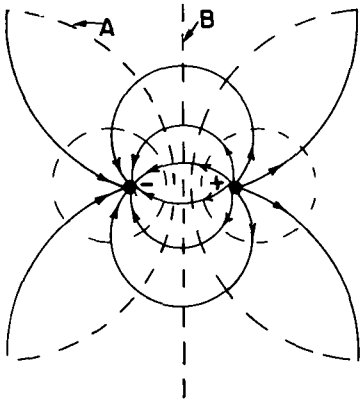


Fig. 6–8. The field lines and equipotentials for two point charges.

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—hopefully 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 rather elaborate numerical methods. Such numerical computations, these days, are set up on a computing machine that will do the work for us, once we have told it how to proceed.

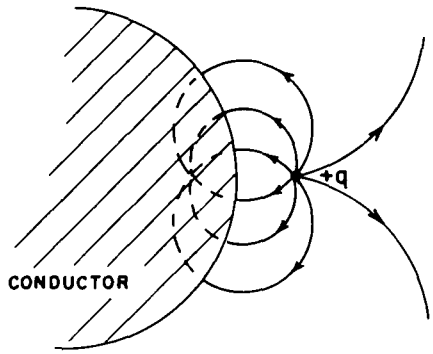


Fig. 6–9. The field outside a conductor shaped like the equipotential A of Fig. 6–8.

On the other hand, there are a lot of little practical cases where it would be nice to be able to find the answer by some more direct method—without having to write a program for a computer. Fortunately, there are a number of cases where the answer can be obtained by squeezing it out of Nature by some trick or other. The first trick we will describe involves making use of solutions we have already obtained for 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 6–8 shows some of the field lines and equipotential surfaces we obtained by the computations in Chapter 5. Now consider the equipotential surface marked *A*. 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 near a point charge. If the metal sheet we placed at the equipotential surface eventually closes on itself (or, in practice, if it goes far enough) we have the kind of situation considered in Section 5–10, 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 have the same fields outside our curved conductor no matter what is inside. We can even fill up

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—as 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 computing the field due to q and to an imaginary 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 hyperbolic-shaped conductors and other 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 that particular shape can be 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 B 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 picture. The field lines for our solution are shown in Fig. 6–10. Notice that the plane, since it was halfway between the two charges, has zero potential. We have solved the problem of a positive charge next to a grounded conducting sheet.

We have now solved for the total field, but what about the *real* 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-

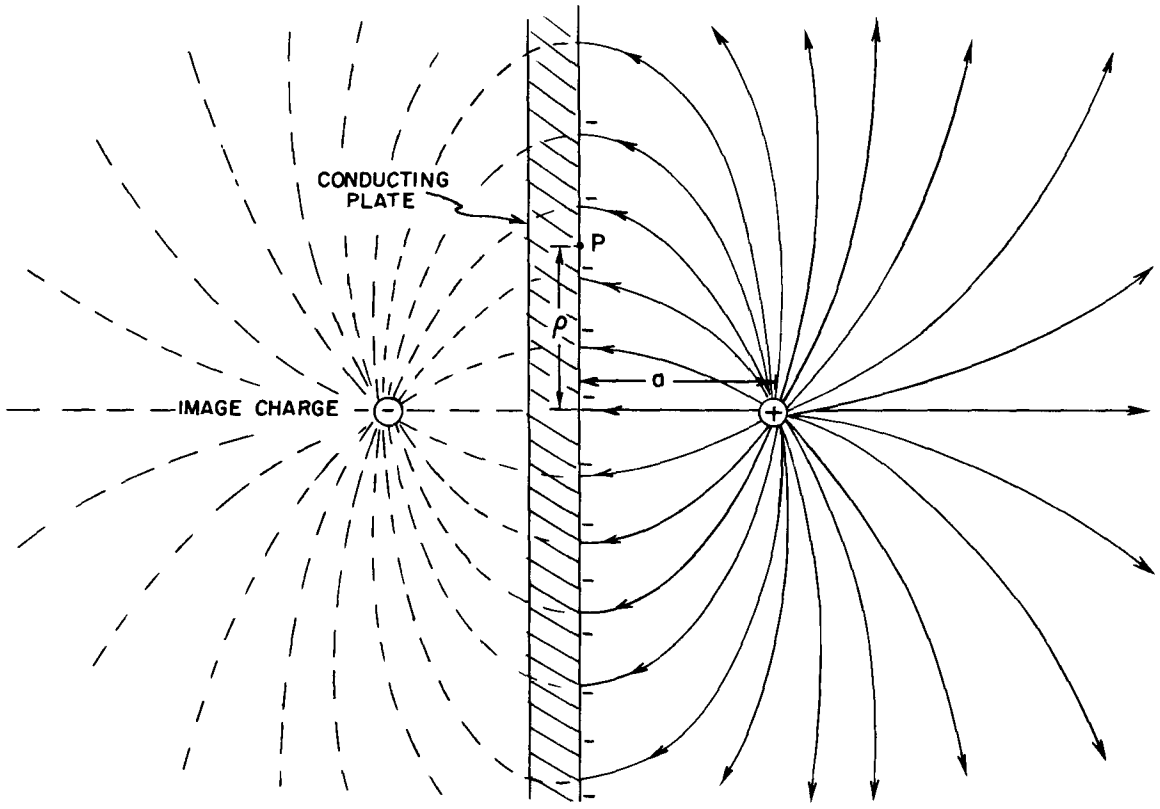


Fig. 6–10. The field of a charge near a plane conducting surface, found by the method of images.

ponent of the electric field just outside a conductor is equal to the density of surface charge σ divided by ϵ_0. We can obtain the density of charge at any point on the surface by working backwards from the normal component of the electric field at the surface. We know that, because we know the field everywhere.

Consider a point on the surface at the distance ρ from the point directly beneath the positive charge (Fig. 6–10). The electric field at this point is normal to the surface and is directed into it. The component normal to the surface of the field from the *positive* point charge is

$$E_{n+} = -\frac{1}{4\pi\epsilon_0}\frac{aq}{(a^2 + \rho^2)^{3/2}} \tag{6.28}$$

To this we must add the electric field produced by the negative image charge. That just doubles the normal component (and cancels all others), so the charge density σ at any point on the surface is

$$\sigma(\rho) = \epsilon_0 E(\rho) = -\frac{2aq}{4\pi(a^2 + \rho^2)^{3/2}}. \tag{6.29}$$

An interesting check on our work is to integrate σ over the whole surface. We find that the total induced charge is $-q$, as it should be.

One further question: Is there a force on the point charge? Yes, because there is an attraction from the induced negative surface charge on the plate. Now that we know what the surface charges are (from Eq. (6.29)), we could compute the force on our positive point charge by an integral. But we also know that the force acting on the positive charge is exactly the same as it *would be* with the negative image charge instead of the plate, because the fields in the neighborhood are the same in both cases. The point charge feels a force toward the plate whose magnitude is

$$F = \frac{1}{4\pi\epsilon_0}\frac{q^2}{(2a)^2}. \tag{6.30}$$

We have found the force much more easily than by integrating over all the negative charges.

6–9 A point charge near a conducting sphere

What other surfaces besides a plane have a simple solution? The next most simple shape is a sphere. Let's find the fields around a metal sphere which has a point charge q near it, as shown in Fig. 6–11. Now we must look for a simple physical situation which gives a sphere for an equipotential surface. If we look around at problems people have already solved, we find that someone has noticed that the field of two *unequal* point charges has an equipotential that is a sphere Aha! If we choose the location of an image charge—and pick the right amount of charge—maybe we can make the equipotential surface fit our sphere. Indeed, it can be done with the following prescription.

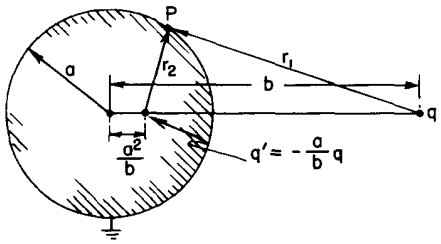


Fig. 6–11. The point charge q induces charges on a grounded conducting sphere whose fields are those of an image charge q′ placed at the point shown.

Assume that you want the equipotential surface to be a sphere of radius a with its center at the distance b from the charge q. Put an image charge of strength $q' = -q(a/b)$ on the line from the charge to the center of the sphere, and at a distance a^2/b from the center. The sphere will be at zero potential.

The mathematical reason stems from the fact that a sphere is the locus of all points for which the distances from two points are in a constant ratio Referring to Fig. 6–11, the potential at P from q and q' is proportional to

$$\frac{q}{r_1} + \frac{q'}{r_2}.$$

The potential will thus be zero at all points for which

$$\frac{q'}{r_2} = -\frac{q}{r_1} \quad \text{or} \quad \frac{r_2}{r_1} = -\frac{q'}{q}.$$

If we place q' at the distance a^2/b from the center, the ratio r_2/r_1 has the constant value a/b. Then if

$$\frac{q'}{q} = -\frac{a}{b}, \tag{6.31}$$

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 its 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 add a point charge q'' at the center of the sphere. The sphere still remains 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 insulated 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 using an image charge q' as before, but, in addition, adding a charge q'' at the center of the sphere, choosing

$$q'' = -q' = \frac{a}{b}q. \tag{6.32}$$

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 neutral sphere. Where 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 closer to itself and leaves positive charges on the surface of the far side. The attraction by the negative charges exceeds the repulsion from the positive charges, there is a net attraction. We can find out how large the attraction is by computing the force on q in the field produced by q' and q''. The total force is the sum of the attractive force between q and a charge $q' = -(a/b)q$, at the distance $b - (a^2/b)$, and the repulsive force between q and a charge $q'' = +(a/b)q$ at the distance b.

Those who were entertained in childhood by the baking powder box which has on its label a picture of a baking powder box which has on its label a picture of a baking powder box which has . may be interested in the following problem. Two equal spheres, one with a total charge of $+Q$ and the other with a total charge of $-Q$, are placed at some distance from each other. What is the force between them? The problem can be solved with an infinite number of images. One first approximates each sphere by a charge at its center. These charges will have image charges in the other sphere. The image charges will have images, etc , etc , etc The solution is like the picture on the box of baking powder—and it converges pretty fast.

6–10 Condensers; parallel plates

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 σ/ϵ_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":

$$\phi_1 - \phi_2 = V.$$

(You will find that sometimes people use V for the potential, but we have chosen to use ϕ.)

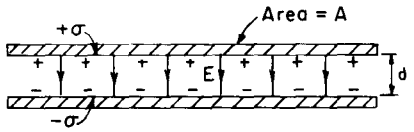


Fig. 6–12 A parallel-plate condenser.

The potential difference V is the work per unit charge required to carry a small charge from one plate to the other, so that

$$V = Ed = \frac{\sigma}{\epsilon_0} d = \frac{d}{\epsilon_0 A} Q, \tag{6.33}$$

where $\pm Q$ is the total charge on each plate, A is the area of the plates, and d is the separation.

We find that the voltage is proportional to the charge. Such a proportionality between V and Q is found for any two conductors in space if there is a plus charge on one and an equal minus charge on the other. The potential difference between them—that is, the voltage—will be proportional to the charge. (We are assuming that there are no other charges around.)

Why this proportionality? Just the superposition principle. Suppose we know the solution for one set of charges, and then we superimpose two such solutions. The charges are doubled, the fields are doubled, and the work done in carrying a unit charge from one point to the other is also doubled. Therefore the potential difference between any two points is proportional to the charges. In particular, the potential difference between the two conductors is proportional to the charges on them. Someone originally wrote the equation of proportionality the other way. That is, they wrote

$$Q = CV,$$

where C is a constant. This coefficient of proportionality is called the *capacity*, and such a system of two conductors is called a *condenser*.* For our parallel-plate condenser

$$C = \frac{\epsilon_0 A}{d} \quad \text{(parallel plates)}. \tag{6.34}$$

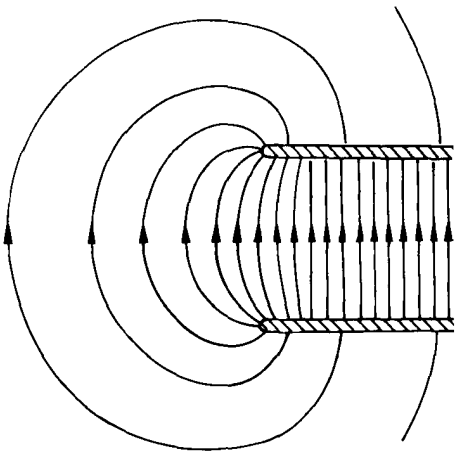


Fig. 6–13. The electric field near the edge of two parallel plates.

This formula is not exact, because the field is not really uniform everywhere between the plates, as we assumed. The field does not just suddenly quit at the edges, but really is more as shown in Fig. 6–13. The total charge is not σA, as we have assumed—there is a little correction for the effects at the edges. To find out what the correction is, we will have to calculate the field more exactly and find out just what does happen at the edges. That is a complicated mathematical problem which can, however, be solved by techniques which we will not describe now. The result of such calculations is that the charge density rises somewhat near the edges of the plates. This means that the capacity of the plates is a little higher than we computed. [A very good approximation for the capacity is obtained if we use Eq. (6.34) but take for A the area one *would* get if the plates were extended artificially by a distance 3/8 of the separation between the plates.]

We have talked about the capacity for two conductors only. Sometimes people talk about the capacity of a single object. They say, for instance, that the capacity of a sphere of radius a is $4\pi\epsilon_0 a$. What they imagine is that the other terminal is another sphere of infinite radius—that when there is a charge $+Q$ on the sphere, the opposite charge, $-Q$, is on an infinite sphere. One can also speak of capacities when there are three or more conductors, a discussion we shall, however, defer.

Suppose that we wish to have a condenser with a very large capacity. We could get a large capacity by taking a very big area and a very small separation. We could put waxed paper between sheets of aluminum foil and roll it up. (If we seal it in plastic, we have a typical radio-type condenser.) What good is it? It is good for storing charge. If we try to store charge on a ball, for example, its potential rises rapidly as we charge it up. It may even get so high that the charge begins to escape into the air by way of sparks. But if we put the same charge on a condenser whose capacity is very large, the voltage developed across the condenser will be small.

---

* Some people think the words "capacitance" and "capacitor" should be used, instead of "capacity" and "condensor." We have decided to use the older terminology, because it is still more commonly heard in the physics laboratory—even if not in textbooks!

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 condenser (or "capacitor") does just that. There are also many applications in electronic instruments and in computers where a condenser 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 23, Vol. I, where we described the properties of resonant circuits.

From the definition of C, we see that its unit is one coul/volt. This unit is also called a *farad*. Looking at Eq. (6.34), we see that one can express the units of ϵ_0 as farad/meter, which is the unit most commonly used. Typical sizes of condensers run from one micro-microfarad (= 1 picofarad) to millifarads. Small condensers of a few picofarads are used in high-frequency tuned circuits, and capacities up to hundreds or thousands of microfarads are found in power-supply filters. A pair of plates one square centimeter in area with a one millimeter separation have a capacity of roughly one micro-microfarad.

$$\epsilon_0 \approx \frac{1}{36\pi \times 10^9} \text{ farad/meter}$$

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 one that 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 just outside.

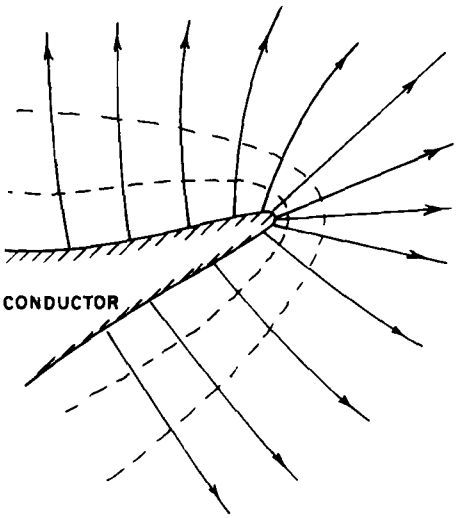


Fig. 6–14. The electric field near a sharp point on a conductor is very high.

One way to see that the field 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 a somewhat idealized version of the conductor 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, which ball has the biggest 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 $\phi_1 = \phi_2$, so

$$\frac{Q}{a} = \frac{q}{b}.$$

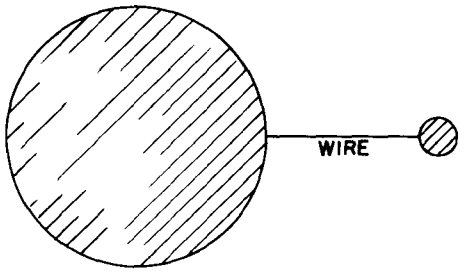


Fig. 6–15. The field of a pointed object can be approximated by that of two spheres at the same potential.

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

$$\frac{E_a}{E_b} = \frac{Q/a^2}{q/b^2} = \frac{b}{a}. \tag{6.35}$$

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 loose 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 hits another atom to be able to knock an

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 on 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 1000 angstroms, 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.

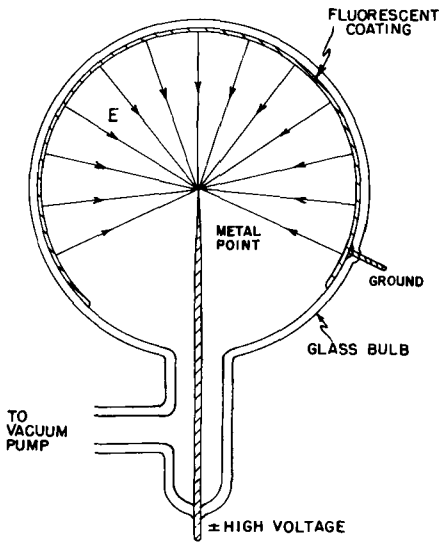


Fig. 6–16. Field-emission microscope.

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 40 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 some kind of an image of the tip of the needle. More precisely, we see a picture of the *emissivity* of the surface of the needle—that is the ease 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, this resolution is not possible for the following reasons. First, there is quantum-mechanical diffraction of the electron waves which blurs 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 transverse component of the velocity causes some smearing of the image. The combination of these two effects limits the resolution to 25 A or so.

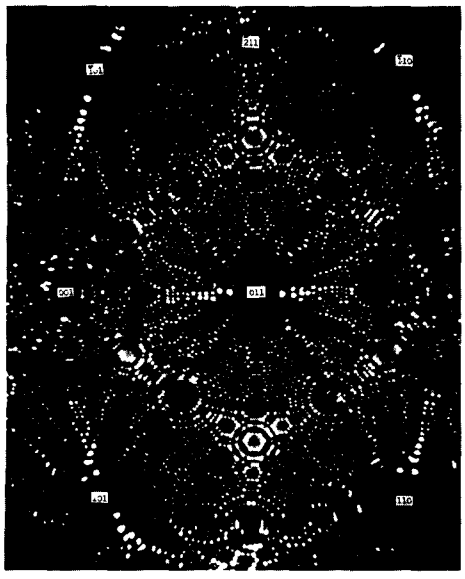


Fig. 6–17. Image produced by a field-emission microscope. [Courtesy of Erwin W. Mueller, Research Prof. of Physics, Pennsylvania State University.]

If, however, we reverse the polarity and introduce a small amount of helium gas into the bulb, much higher 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 helium ion is then accelerated outward along a field line to the fluorescent screen. Since the helium ion is so much heavier than an electron, the quantum-mechanical wavelengths are much smaller. If the temperature is not too high, the effect of the thermal velocities is also smaller than in the electron case. With less smearing of the image a much sharper picture of the point is obtained. It has been possible to obtain magnifications up to 2,000,000 times with the positive ion field-emission microscope—a magnification ten times better than is obtained with the best electron microscope.

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 ionizes 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 packed 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 E. W. Mueller: "The field-ion microscope," *Advances in Electronics and Electron Physics*, **13**, 83–179 (1960). Academic Press, New York

7

The Electric Field in Various Circumstances (Continued)

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 an electrolyte

7-5 The electrostatic field of a grid

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 direct 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,

$$\nabla^2\phi = 0, \tag{7.1}$$

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 closed 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 measurements of a physical analog. 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,

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 electrostatics. 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

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = 0. \tag{7.2}$$

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 $\mathfrak{z}$ is defined as

$$\mathfrak{z} = x + iy.$$

(Do not confuse $\mathfrak{z}$ 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 $\mathfrak{z}$. We can use $\mathfrak{z}$ as a single (complex) variable, and with it write the usual kinds of mathematical functions $F(\mathfrak{z})$. For example,

$$F(\mathfrak{z}) = \mathfrak{z}^2,$$

or

$$F(\mathfrak{z}) = 1/\mathfrak{z}^3,$$

or

$$F(\mathfrak{z}) = \mathfrak{z} \log \mathfrak{z},$$

and so forth.

Given any particular $F(\mathfrak{z})$ we can substitute $\mathfrak{z} = x + iy$, and we have a function of x and y—with real and imaginary parts. For example,

$$\mathfrak{z}^2 = (x + iy)^2 = x^2 - y^2 + 2ixy. \tag{7.3}$$

Any function $F(\mathfrak{z})$ can be written as a sum of a pure real part and a pure imaginary part, each part a function of x and y:

$$F(\mathfrak{z}) = U(x, y) + iV(x, y), \tag{7.4}$$

where $U(x, y)$ and $V(x, y)$ are real functions. Thus from any complex function $F(\mathfrak{z})$ two new functions $U(x, y)$ and $V(x, y)$ can be derived. For example, $F(\mathfrak{z}) = \mathfrak{z}^2$ gives us the two functions

$$U(x, y) = x^2 - y^2, \tag{7.5}$$

and

$$V(x, y) = 2xy. \tag{7.6}$$

Now we come to a miraculous 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 matter would

become too dull.) It is this. For any "ordinary function" (mathematicians will define it better) the functions U and V *automatically* satisfy the relations

$$\frac{\partial U}{\partial x} = \frac{\partial V}{\partial y}, \tag{7.7}$$

$$\frac{\partial V}{\partial x} = -\frac{\partial U}{\partial y}. \tag{7.8}$$

It follows immediately that each of the functions U and V satisfy Laplace's equation:

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0, \tag{7.9}$$

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0. \tag{7.10}$$

These equations are clearly true for the functions of (7.5) and (7.6).

Thus, starting with any ordinary 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 *any* function $F(\zeta)$ and it should represent *some* electric field problem—in fact, *two* problems, because U and V *each* represent solutions. We can write down as many solutions as we wish—by just making up functions—then we just have to find the *problem* that goes with each solution. It may sound backwards, but it's a possible approach.

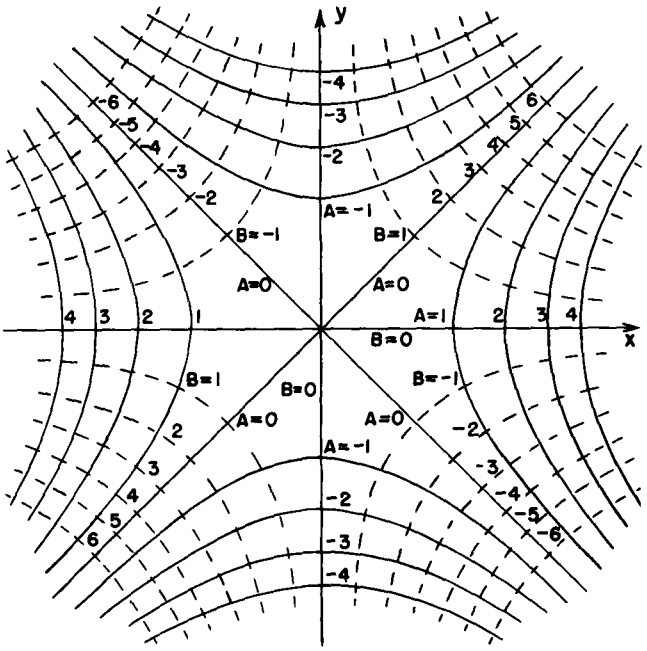


Fig. 7–1. Two sets of orthogonal curves which can represent equipotentials in a two-dimensional electrostatic field.

As an example, let's see what physics the function $F(\zeta) = \zeta^2$ gives us. From it we get the two potential functions of (7.5) and (7.6). To see what problem the function U 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 hyperbolas 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

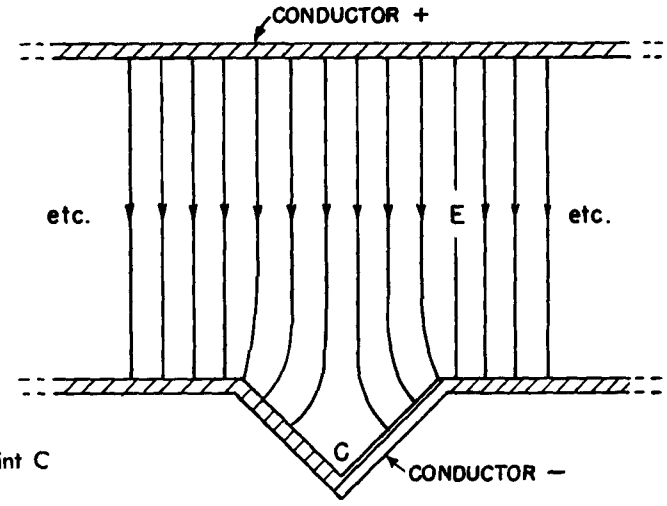


Fig. 7–2. The field near the point C is the same as that in Fig. 7–1.

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 or protuberances 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 hyperbolas 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 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 desired 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 = constant. We have a bonus! The curves for V = constant are orthogonal to the ones for U = constant because of the equations (7.7) and (7.8). Whenever we choose a function $F(\mathfrak{z})$, we get from U and V both the equipotentials and field lines. And you will remember that we have solved either of two problems, depending on which set of curves we call the equipotentials.

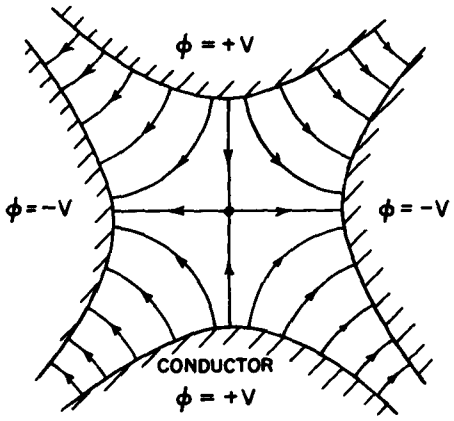


Fig. 7–3. The field in a quadrupole lens.

As a second example, consider the function

$$F(\mathfrak{z}) = \sqrt{\mathfrak{z}}. \tag{7.11}$$

If we write

$$\mathfrak{z} = x + iy = \rho e^{i\theta},$$

where

$$\rho = \sqrt{x^2 + y^2}$$

and

$$\tan\theta = y/x,$$

then

$$F(\mathfrak{z}) = \rho^{1/2} e^{i\theta/2}$$

$$= \rho^{1/2}\left(\cos\frac{\theta}{2} + i\sin\frac{\theta}{2}\right),$$

from which

$$F(\mathfrak{z}) = \left[\frac{(x^2+y^2)^{1/2}+x}{2}\right]^{1/2} + i\left[\frac{(x^2+y^2)^{1/2}-x}{2}\right]^{1/2}. \tag{7.12}$$

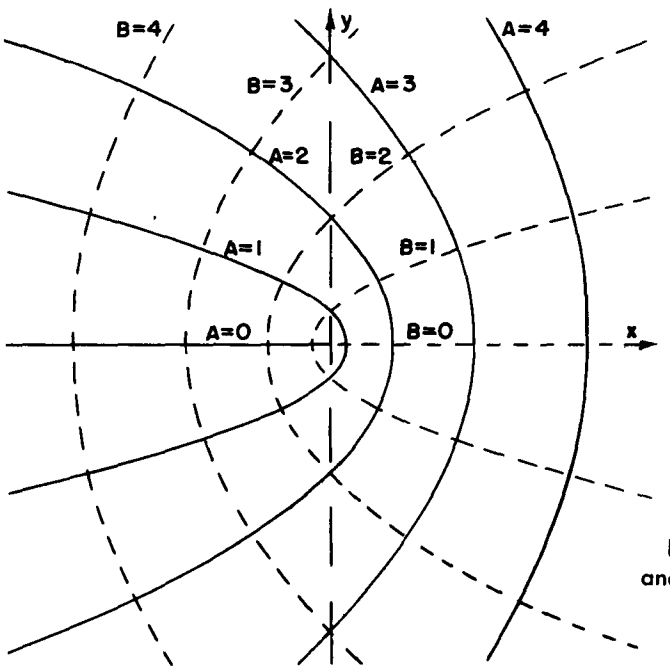


Fig. 7–4. Curves of constant $U(x, y)$ and $V(x, y)$ from Eq. (7.12).

The curves for $U(x, y) = A$ and $V(x, y) = B$, using 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 $B = 0$—to the right of the y-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.

Further examples are

$$F(\zeta) = z^{3/2}, \tag{7.13}$$

which yields the field *outside* a rectangular corner

$$F(\zeta) = \log \zeta, \tag{7.14}$$

which yields the field for a line charge, and

$$F(\zeta) = 1/\zeta, \tag{7.15}$$

which gives the field for the two-dimensional analog of an electric dipole, i.e., two parallel line charges with opposite polarities, 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.

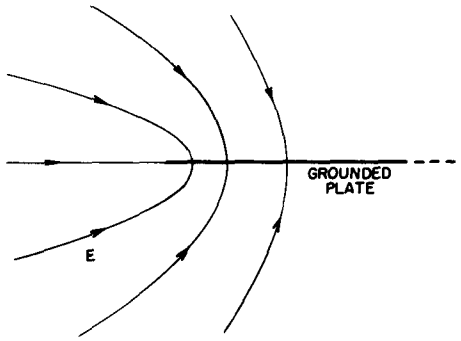


Fig. 7–5. The electric field near the edge of a thin grounded plate.

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 electrostatics 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—an upper layer of the atmosphere—is an example of such a plasma. The ultraviolet rays from the sun knock

electrons off the molecules of the air, creating free electrons and ions. In such a plasma the positive ions are very much heavier than the electrons, so we may neglect the ionic motion, in comparison to that of the electrons.

Let n_0 be the density of electrons in the undisturbed, equilibrium state. This must also be the density of positive ions, since the plasma is electrically neutral (when undisturbed). Now we suppose that the electrons are somehow moved from equilibrium and ask what happens. If the density of the electrons in one region is increased, they will repel each other and tend to return to their equilibrium positions. As the electrons move toward their original positions they pick up kinetic energy, and instead of coming 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 worry only about a situation in which the motions are all in one dimension, say x. Let us suppose that the electrons originally at x are, at the instant t, displaced from their equilibrium positions by a small amount $s(x, t)$. Since the electrons have been displaced, their density will, in general, be changed. The change in density is easily calculated. Referring to Fig. 7–6, the electrons initially contained between the two planes a and b have moved and are now contained between the planes a' and b'. The number of electrons that were between a and b is proportional to $n_0\Delta x$; the *same* number are now contained in the space whose width is $\Delta x + \Delta s$. The density has changed to

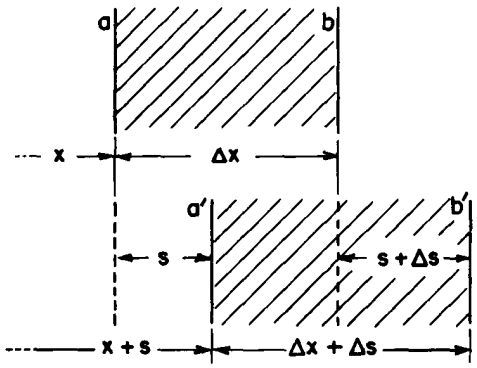


Fig. 7–6. Motion in a plasma wave. The electrons at the plane a move to a', and those at b move to b'.

$$n = \frac{n_0\Delta x}{\Delta x + \Delta s} = \frac{n_0}{1 + (\Delta s/\Delta x)}. \tag{7.16}$$

If the change in density is small, we can write [using the binomial expansion for $(1 + \epsilon)^{-1}$]

$$n = n_0\left(1 - \frac{\Delta s}{\Delta x}\right). \tag{7.17}$$

We assume that the positive ions do not move appreciably (because of the much larger inertia), so their density remains n_0. Each electron carries the charge $-q_e$, so the average charge density at any point is given by

$$\rho = -(n - n_0)q_e,$$

or

$$\rho = n_0 q_e \frac{ds}{dx} \tag{7.18}$$

(where we have written the differential form for $\Delta s/\Delta x$).

The charge density is related to the electric field by Maxwell's equations, in particular,

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}. \tag{7.19}$$

If the problem is indeed one-dimensional (and if there are no other fields but the one due to the displacements of the electrons), the electric field $\boldsymbol{E}$ has a single component E_x. Equation (7.19), together with (7.18), gives

$$\frac{\partial E_x}{\partial x} = \frac{n_0 q_e}{\epsilon_0}\frac{\partial s}{\partial x}. \tag{7.20}$$

Integrating Eq. (7.20) gives

$$E_x = \frac{n_0 q_e}{\epsilon_0} s + K. \tag{7.21}$$

Since $E_x = 0$ when $s = 0$, the integration constant K is zero.

The force on an electron in the displaced position is

$$F_x = -\frac{n_0 q_e^2}{\epsilon_0} s, \tag{7.22}$$

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

$$m_e \frac{d^2s}{dt^2} = -\frac{n_0 q_e^2}{\epsilon_0} s. \tag{7.23}$$

We find that s will vary harmonically. Its time variation will be as cos ωt, or—using the exponential notation of Vol. I—as

$$e^{i\omega_p t}. \tag{7.24}$$

The frequency of oscillation ω_p is determined from (7.23):

$$\omega_p^2 = \frac{n_0 q_e^2}{\epsilon_0 m_e}, \tag{7.25}$$

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 e^2 defined by

$$e^2 = \frac{q_e^2}{4\pi\epsilon_0} = 2.3068 \times 10^{-28} \text{ newton·meter}^2. \tag{7.26}$$

Using this convention, Eq. (7.25) becomes

$$\omega_p^2 = \frac{4\pi e^2 n_0}{m_e}, \tag{7.27}$$

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 one tries to propagate a radiowave through the ionosphere, one finds that it can penetrate only if its frequency is higher than the plasma frequency. Otherwise the signal is reflected back. We must 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 metals. In a metal we have a contained plasma of positive ions, and free electrons. The density n_0 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 lose the energy $\hbar\omega_p$ to the plasma oscillations. This does indeed happen. It was first observed experimentally in 1936 that electrons with energies of a few hundred to a few thousand electron volts lost energy in jumps when scattering from or going through a thin metal foil. The effect was not understood until 1953 when Bohm 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. B. Swann, *Phys. Rev.* **115**, 869 (1959).

7–4 Colloidal particles in an electrolyte

We turn to another phenomenon in which the locations 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 $\nabla^2\phi = -\rho/\epsilon_0$ or, for fields that vary in only one dimension, by

$$\frac{d^2\phi}{dx^2} = -\frac{\rho}{\epsilon_0}. \tag{7.28}$$

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 *also* satisfies (7.28).

According to statistical mechanics (see Chapter 40, Vol. I), 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

$$n(x) = n_0 e^{-U(x)/kT}, \tag{7.29}$$

where $U(x)$ 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 $q_e\phi(x)$, so that

$$U(x) = q_e\phi(x).$$

The density of positive ions, n_+, is then

$$n_+(x) = n_0 e^{-q_e\phi(x)/kT}.$$

Similarly, the density of negative ions is

$$n_-(x) = n_0 e^{+q_e\phi(x)/kT}.$$

The total charge density is

$$\rho = q_e n_+ - q_e n_-,$$

or

$$\rho = q_e n_0 (e^{-q_e\phi/kT} - e^{+q_e\phi/kT}). \tag{7.30}$$

Combining this with Eq. (7.28), we find that the potential ϕ must satisfy

$$\frac{d^2\phi}{dx^2} = -\frac{q_e n_0}{\epsilon_0}(e^{-q_e\phi/kT} - e^{+q_e\phi/kT}). \tag{7.31}$$

This equation is readily solved in general [multiply both sides by $2(d\phi/dx)$, 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

$$e^{\pm q_e\phi/kT} = 1 \pm \frac{q_e\phi}{kT}. \tag{7.32}$$

Equation (7.31) then gives

$$\frac{d^2\phi}{dx^2} = + \frac{2n_0q_e^2}{\epsilon_0 kT}\,\phi(x). \tag{7.33}$$

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

$$\phi = Ae^{-x/D} + Be^{+x/D}, \tag{7.34}$$

with

$$D^2 = \frac{\epsilon_0 kT}{2n_0q^2}. \tag{7.35}$$

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

$$\phi = Ae^{-x/D}, \tag{7.36}$$

in which A is the potential at $x = 0$, the surface of the colloidal particle.

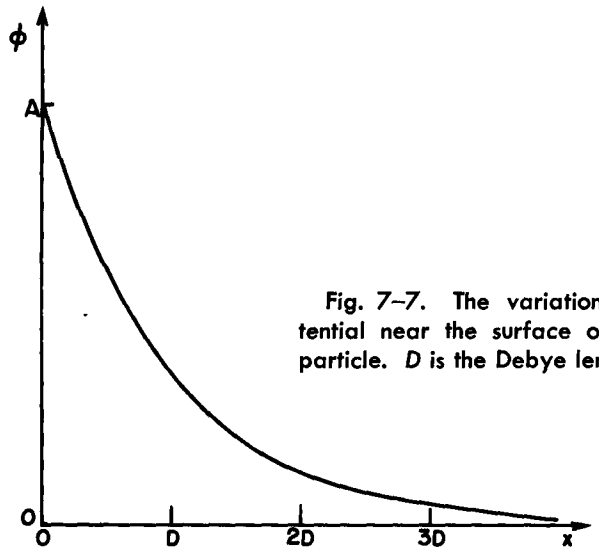


Fig. 7–7. The variation of the potential near the surface of a colloidal particle. D is the Debye length.

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 ion sheath that surrounds a large charged particle in an electrolyte. Equation (7.36) says that the sheath gets thinner with increasing concentration of the ions (n_0) or with decreasing temperature.

The constant A in Eq. (7.36) is easily obtained if we know the surface charge σ on the colloid particle. We know that

$$E_n = E_x(0) = \frac{\sigma}{\epsilon_0}. \tag{7.37}$$

But E is also the gradient of ϕ:

$$E_x(0) = -\left.\frac{\partial\phi}{\partial x}\right|_0 = + \frac{A}{D}, \tag{7.38}$$

from which we get

$$A = \frac{\sigma D}{\epsilon_0}. \tag{7.39}$$

Using this result in (7.36), we find (by taking $x = 0$) that the potential of the colloidal particle is

$$\phi(0) = \frac{\sigma D}{\epsilon_0}. \tag{7.40}$$

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 sheaths get thin enough, the particles have 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. Also, 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, decreasing 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 a uniform 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 found 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 7–8 shows a rough sketch of the equipotentials 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.

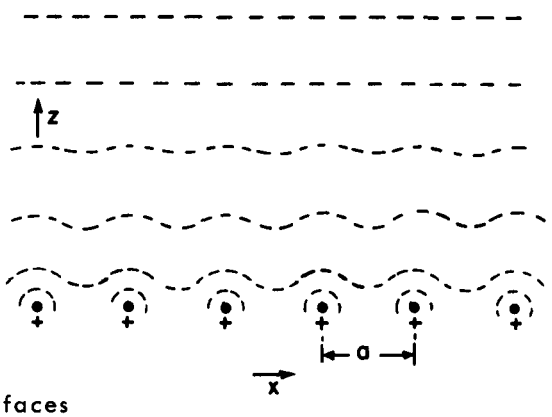


Fig. 7–8. Equipotential surfaces above a uniform grid of charged wires.

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 xy-plane and run parallel to the y-axis, then we might try terms like

$$\phi(x, z) = F_n(z)\cos\frac{2\pi nx}{a}, \tag{7.41}$$

where a is the spacing of the wires and n is the harmonic number. (We have assumed long wires, so there should be no variation with y.) A complete solution would be made up of a sum of such terms for $n = 1, 2, 3, \ldots$.

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

$$-\frac{4\pi^2n^2}{a^2}F_n(z)\cos\frac{2\pi nx}{a} + \frac{d^2F_n}{dz^2}\cos\frac{2\pi nx}{a} = 0, \tag{7.42}$$

or that $F_n(z)$ must satisfy

$$\frac{d^2F_n}{dz^2} = \frac{4\pi^2n^2}{a^2}F_n. \tag{7.43}$$

So we must have

$$F_n = A_ne^{-z/z_0}, \tag{7.44}$$

where

$$z_0 = \frac{a}{2\pi n}. \tag{7.45}$$

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 $z_0 = a/2\pi n$. For the first harmonic ($n = 1$), the amplitude falls by the factor $e^{-2\pi}$ (a large decrease) each time we increase z 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

$$\phi_0 = -E_0z$$

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 differentiated, 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—lighter and cheaper than copper sheet—is often used to shield sensitive electrical equipment from external disturbing fields.

8

Electrostatic Energy

8–1 The electrostatic energy of charges. A uniform sphere

8–2 The energy of a condenser. Forces on charged conductors

8–3 The electrostatic energy of an ionic crystal

8–4 Electrostatic energy in nuclei

8–5 Energy in the electrostatic field

8–6 The energy of a point charge

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 electrostatics 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 some 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

$$\frac{q_1 q_2}{4\pi\epsilon_0 r_{12}}. \tag{8.1}$$

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 a 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

$$\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}. \tag{8.2}$$

The total electrostatic energy U is the sum of the energies of all possible pairs of charges:

$$U = \sum_{\text{all pairs}} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}. \tag{8.3}$$

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 a (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

$$dU = \frac{Q_r\, dQ}{4\pi\epsilon_0 r}. \tag{8.4}$$

Review: Chapter 4, Vol. I, *Conservation of Energy*
Chapters 13 and 14, Vol. I, *Work and Potential Energy*

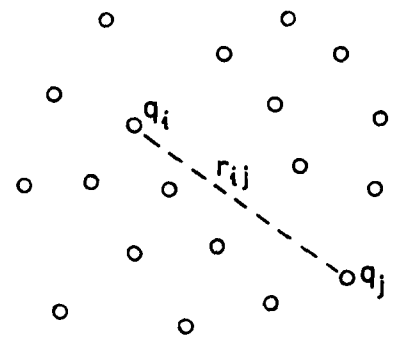


Fig. 8–1. The electrostatic energy of a system of particles is the sum of the electrostatic energy of each pair.

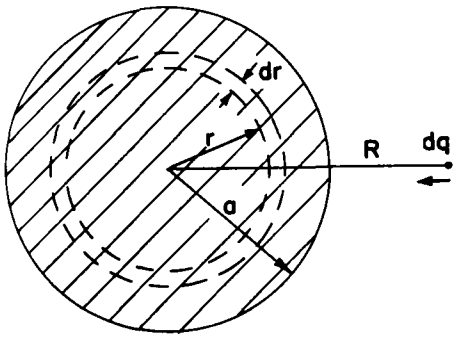


Fig. 8–2. The energy of a uniform sphere of charge can be computed by imagining that it is assembled from successive spherical shells.

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

$$dU = \frac{4\pi\rho^2 r^4\, dr}{3\epsilon_0}. \tag{8.5}$$

The total energy required to assemble the sphere is the integral of dU from $r = 0$ to $r = a$, or

$$U = \frac{4\pi\rho^2 a^5}{15\epsilon_0}. \tag{8.6}$$

Or if we wish to express the result in terms of the total charge Q of the sphere,

$$U = \frac{3}{5}\frac{Q^2}{4\pi\epsilon_0 a}. \tag{8.7}$$

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

$$V = \frac{Q}{C}, \tag{8.8}$$

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

$$U = \frac{1}{2}\frac{Q^2}{C}. \tag{8.9}$$

This energy can also be written as

$$U = \tfrac{1}{2}CV^2. \tag{8.10}$$

Recalling that the capacity of a conducting sphere (relative to infinity) is

$$C_{\text{sphere}} = 4\pi\epsilon_0 a,$$

we can immediately get from Eq. (8.9) the energy of a charged sphere,

$$U = \frac{1}{2}\frac{Q^2}{4\pi\epsilon_0 a}. \tag{8.11}$$

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

$$\Delta W = F\,\Delta z, \tag{8.12}$$

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

$$\Delta U = \frac{1}{2}\,Q^2\,\Delta\left(\frac{1}{C}\right). \tag{8.13}$$

Equating (8.12) and (8.13), we have

$$F\,\Delta z = \frac{Q^2}{2}\,\Delta\left(\frac{1}{C}\right). \tag{8.14}$$

This can also be written as

$$F\,\Delta z = -\frac{Q^2}{2C^2}\,\Delta C. \tag{8.15}$$

The force, of course, results from the attraction of the charges on the plates, but we see that we do not have to worry in detail 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 F 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.

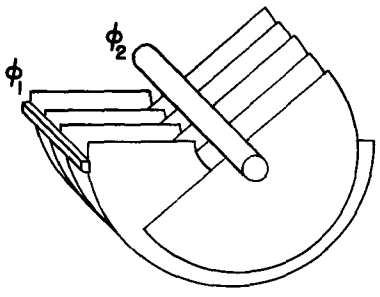


Fig. 8–3. What is the torque on a variable capacitor?

Returning to the special case of a parallel-plate condenser, we can use the formula we derived in Chapter 6 for the capacity:

$$\frac{1}{C} = \frac{d}{\epsilon_0 A}, \tag{8.16}$$

where A is the area of each plate. If we increase the separation by Δz,

$$\Delta\left(\frac{1}{C}\right) = \frac{\Delta z}{\epsilon_0 A}.$$

From Eq. (8.14) we get that the force between the plates is

$$F = \frac{Q^2}{2\epsilon_0 A}. \tag{8.17}$$

Let's look at Eq. (8.17) a little more closely and see if we can tell how the force arises. If for the charge on one plate we write

$$Q = \sigma A,$$

Eq. (8.17) can be rewritten as

$$F = \frac{1}{2} Q \frac{\sigma}{\epsilon_0}.$$

Or, since the electric field between the plates is

$$E_0 = \frac{\sigma}{\epsilon_0},$$

then

$$F = \tfrac{1}{2} Q E_0. \tag{8.18}$$

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 factor of one-half. The reason is that E_0 is not the field *at* the charges. If we imagine that the charge at the surface of the plate occupies a thin layer, as indicated in Fig. 8–4, the field will vary from zero at the inner boundary of the layer to E_0 in the space outside of the plate. The average field acting on the surface charges is $E_0/2$. That is why the factor one-half is in Eq. (8.18).

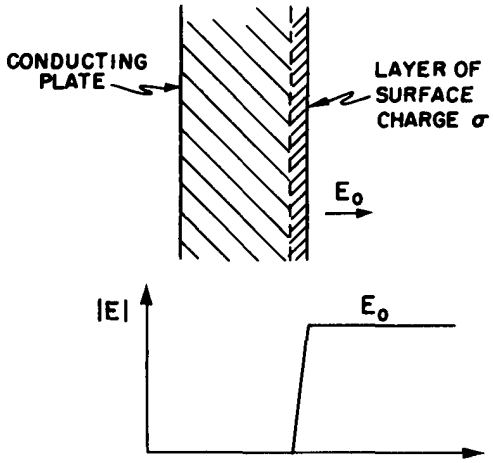


Fig. 8–4. The field at the surface of a conductor varies from zero to $E_0 = \sigma/\epsilon_0$, as one passes through the layer of surface charge.

You should notice that in computing the virtual work we have assumed that the charge on the condenser 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 condenser was held at a constant potential difference as we made the virtual displacement. Then we should have taken

$$U = \tfrac{1}{2} C V^2$$

and in place of Eq. (8.15) we would have had

$$F \Delta z = \tfrac{1}{2} V^2 \Delta C,$$

which gives a force equal in magnitude to the one in Eq. (8.15) (because $V = Q/C$), but with the opposite sign! Surely the force between the condenser plates doesn't reverse in sign as we disconnect 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 $F \Delta z$ *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 $F \Delta z$ is $-\frac{1}{2} V^2 \Delta C$, as before.

8–3 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 can be thought of as rigid spheres. They attract electrically until they begin to touch; then there is a repulsive 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 A ($=2.81 \times 10^{-8}$ 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,

$$N_0 = 6.02 \times 10^{23},$$

the energy of vaporization can also be given as

$$W = 7.64 \times 10^5 \text{ joules/mole}.$$

Physical chemists prefer for an energy unit 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

$$W = 183 \text{ kcal/mole}.$$

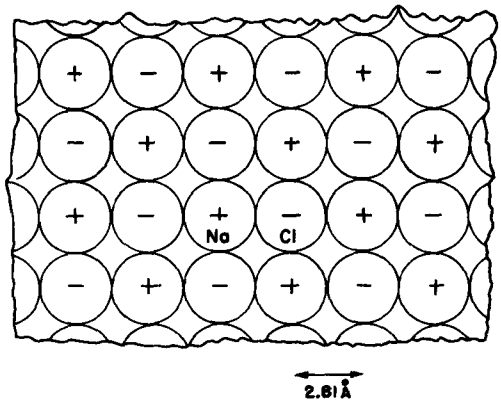


Fig. 8–5. Cross section of a salt crystal on an atomic scale. The checkerboard arrangement of Na and Cl ions is the same in the two cross sections perpendicular to the one shown. (See Vol. I, Fig. 1–7.)

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 *pairs* 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 $e^2 = q_e^2/4\pi\epsilon_0$ and a is the center-to-center 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 infinite 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 our special ion, 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

$$\begin{aligned} U_1 &= \frac{e^2}{a}\left(-\frac{2}{1}+\frac{2}{2}-\frac{2}{3}+\frac{2}{4}+\cdots\right) \\ &= -\frac{2e^2}{a}\left(1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\cdots\right). \end{aligned} \tag{8.19}$$

The series converges slowly, so it is difficult to evaluate numerically, but it is known to be equal to ln 2. So

$$U_1 = -\frac{2e^2}{a}\ln 2 = -1.386\,\frac{e^2}{a}. \tag{8.20}$$

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 $\sqrt{5}\,a$, the next at $\sqrt{10}\,a$, and so on. So for the whole line we get the series

$$\frac{e^2}{a}\left(-\frac{1}{1}+\frac{2}{\sqrt{2}}-\frac{2}{\sqrt{5}}+\frac{2}{\sqrt{10}}\cdots\right). \tag{8.21}$$

There are *four* such lines: above, below, in front, and in back. Then there are the four lines which are the nearest lines on diagonals, and 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 (8.20) 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 correct. 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 this 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 some 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.

8–4 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 hold 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 once one had determined this law of force, and the corresponding ones 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 an 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' spin. The protons have a spin, and any two interacting protons may be spinning with their angular momenta 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 (a) and (b).

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 nucleon 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 generator (for example, in Caltech's Kellogg 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 one-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^{11}. The energies of the lowest sixteen excited states of C^{11} 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 pattern 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

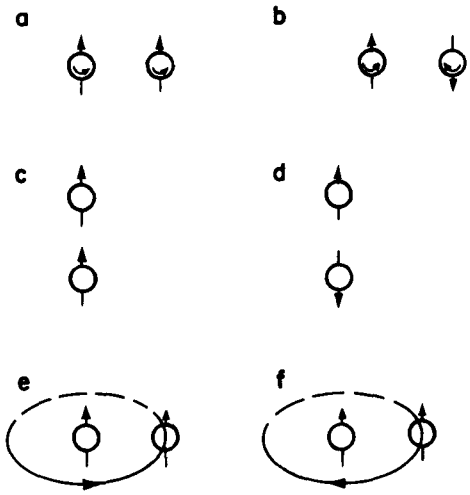


Fig. 8–6. The force between two protons depends on every possible parameter.

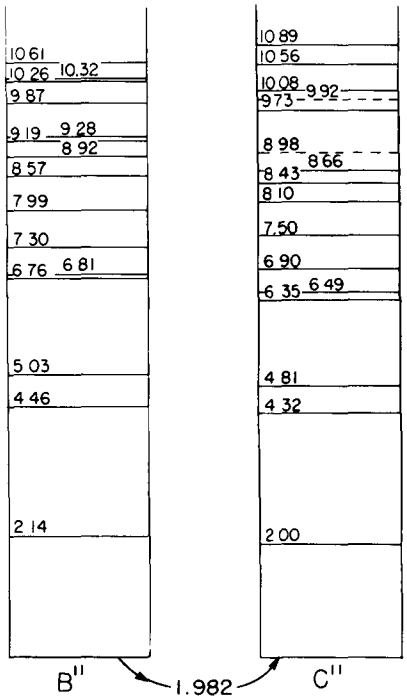


Fig. 8–7. The energy levels of B^{11} and C^{11} (energies in Mev). The ground state of C^{11} is 1.982 Mev higher than that of B^{11}.

order of 0.1 Mev. And so on. After about the tenth level, the correspondence seems to become lost, 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 pattern 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-neutron and proton-proton 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 *electrical* 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 even 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 consider that a nucleus is like a sphere with uniform charge density, we would expect the electrostatic energy (from Eq. 8.7) to be

$$U = \frac{3}{5}\frac{(Zq_e)^2}{4\pi\epsilon_0 r}, \tag{8.22}$$

where q_e 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

$$U = \frac{3}{5}\frac{Z(Z-1)q_e^2}{4\pi\epsilon_0 r} = \frac{3}{5}\frac{Z(Z-1)e^2}{r}. \tag{8.23}$$

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 mc^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

$$1.982 + 0.784 = 2.786 \text{ Mev}.$$

Using this energy in Eq. (8.23), for the radius of either B^{11} or C^{11} we find

$$r = 3.12 \times 10^{-13} \text{ cm}. \tag{8.24}$$

Does this number have any meaning? To see whether it does, we should compare it with some other determination of the radius of these nuclei. For example, we can make another measurement of the radius of a nucleus by seeing how it scatters fast 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

$$r = A^{1/3} r_0, \tag{8.25}$$

where

$$r_0 = 1.2 \times 10^{-13} \text{ cm}. \tag{8.26}$$

From these measurements we find that the radius of a B^{11} (or a C^{11}) nucleus is expected to be

$$r = (1.2 \times 10^{-13})(11)^{1/3} = 2.7 \times 10^{-13} \text{ cm}.$$

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—makes a kind of *core;* when one more particle is added to this core, it revolves around on 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

$$\frac{Z_B q_e^2}{4\pi\epsilon_0 a},$$

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. (8.23) 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^{-13} 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 $\rho\, dV$. Then Eq. (8.3) should be written

$$U = \frac{1}{2} \int_{\substack{\text{all} \\ \text{space}}} \frac{\rho(1)\rho(2)}{4\pi\epsilon_0 r_{12}}\, dV_1\, dV_2. \tag{8.27}$$

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

$$U = \frac{1}{2}\int \rho\phi\, dV. \tag{8.28}$$

This equation can be interpreted as follows. The potential energy of the charge $\rho\, dV$ is the product of this charge and the potential at the same point. The total energy is therefore the integral over $\phi\rho\, dV$. 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

$$U = q_2\phi(2) = q_2 \frac{q_1}{4\pi\epsilon_0 r_{12}}.$$

Notice that we could also write

$$U = \tfrac{1}{2}[q_1\phi(1) + q_2\phi(2)]. \tag{8.29}$$

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 at 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 gravitation 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 radiowaves 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 came. 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

$$U = \frac{\epsilon_0}{2} \int \boldsymbol{E} \cdot \boldsymbol{E}\, dV. \tag{8.30}$$

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

$$u = \frac{\epsilon_0}{2} \boldsymbol{E} \cdot \boldsymbol{E} = \frac{\epsilon_0 E^2}{2}. \tag{8.31}$$

This idea is illustrated in Fig. 8–8.

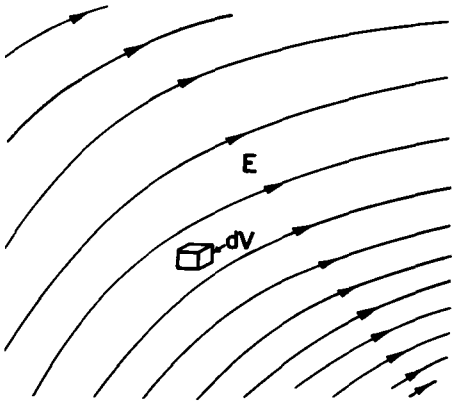


Fig. 8–8. Each volume element $dV = dx\, dy\, dz$ in an electric field contains the energy $(\epsilon_0/2)E^2\, dV$.

To show that Eq. (8.30) is consistent with our laws of electrostatics, 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

$$U = -\frac{\epsilon_0}{2} \int \phi \nabla^2 \phi\, dV. \tag{8.32}$$

Writing out the components of the integrand, we see that

$$\begin{aligned} \phi \nabla^2 \phi &= \phi \left(\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} \right) \\ &= \frac{\partial}{\partial x} \left(\phi \frac{\partial \phi}{\partial x} \right) - \left(\frac{\partial \phi}{\partial x} \right)^2 + \frac{\partial}{\partial y} \left(\phi \frac{\partial \phi}{\partial y} \right) - \left(\frac{\partial \phi}{\partial y} \right)^2 + \frac{\partial}{\partial z} \left(\phi \frac{\partial \phi}{\partial z} \right) - \left(\frac{\partial \phi}{\partial z} \right)^2 \\ &= \nabla \cdot (\phi \nabla \phi) - (\nabla \phi) \cdot (\nabla \phi). \end{aligned} \tag{8.33}$$

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:

$$\int_{\text{vol.}} \nabla \cdot (\phi \nabla \phi)\, dV = \int_{\text{surface}} (\phi \nabla \phi) \cdot \boldsymbol{n}\, da. \tag{8.34}$$

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 the net

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 ($R \to \infty$), the surface integral goes to zero and we have that

$$U = \frac{\epsilon_0}{2} \int_{\substack{\text{all}\\\text{space}}} (\nabla\phi)\cdot(\nabla\phi)\,dV = \frac{\epsilon_0}{2} \int_{\substack{\text{all}\\\text{space}}} \boldsymbol{E}\cdot\boldsymbol{E}\,dV. \tag{8.35}$$

We see that it is possible for us to represent the energy of any charge distribution as being the integral over an 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

$$E = \frac{q}{4\pi\epsilon_0 r^2}.$$

So the energy density at the distance r from the charge is

$$\frac{\epsilon_0 E^2}{2} = \frac{q^2}{32\pi^2\epsilon_0 r^4}.$$

We can take for an element of volume a spherical shell of thickness dr and area $4\pi r^2$. The total energy is

$$U = \int_{r=0}^{\infty} \frac{q^2}{8\pi\epsilon_0 r^2}\,dr = -\frac{q^2}{8\pi\epsilon_0}\frac{1}{r}\bigg|_{r=0}^{r=\infty}. \tag{8.36}$$

Now the limit at $r = \infty$ 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 locating 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. There are difficulties with either 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 momentum in an electromagnetic field, we will give a more complete account of these fundamental difficulties in our understanding of nature.

9

Electricity in the Atmosphere

9–1 The electric potential gradient of the amosphere

On an ordinary day over flat desert 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 outdoors 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 out 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 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.

9–1 The electric potential gradient of the atmosphere

9–2 Electric currents in the atmosphere

9–3 Origin of the atmospheric currents

9–4 Thunderstorms

9–5 The mechanism of charge separation

9–6 Lightning

Reference: Chalmers, J. Alan, *Atmospheric Electricity*, Pergamon Press, London (1957).

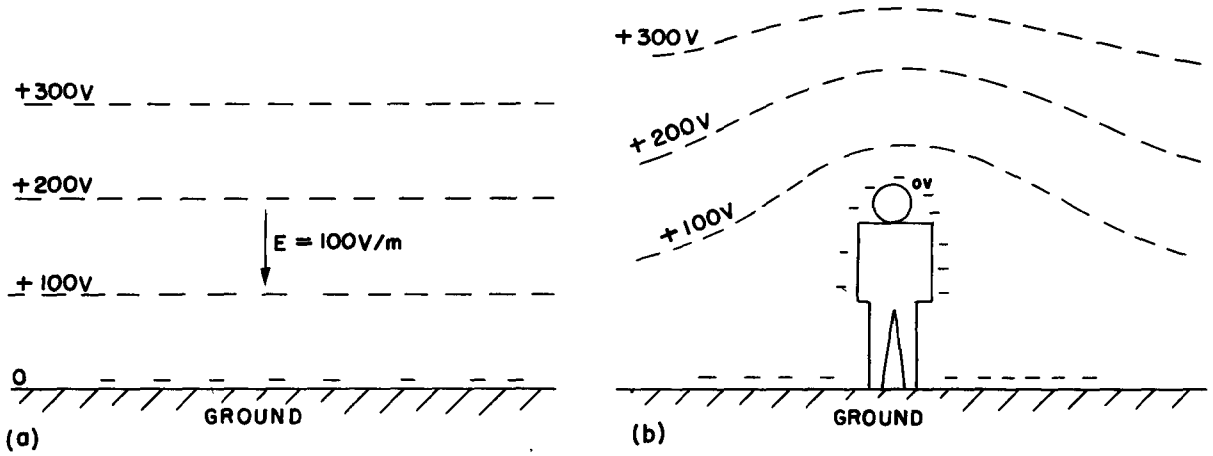


Fig. 9–1. (a) The potential distribution above the earth. (b) The potential distribution near a man in an open flat place.

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 onto) the conductor until it comes to the potential at its level. 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.

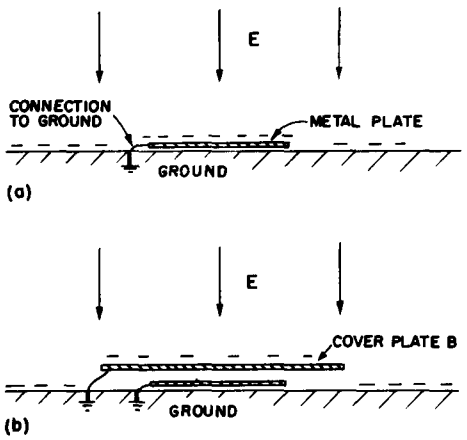


Fig. 9–2. (a) A grounded metal plate will have the same surface charge as the earth. (b) If the plate is covered with a grounded conductor it will have no surface charge.

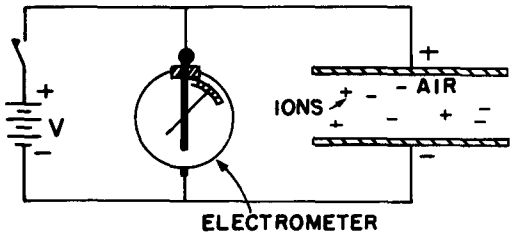


Fig. 9–3. Measuring the conductivity of air due to the motion of ions.

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.

Having 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 micromicroamperes crosses each 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—passes 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 do not stay as 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 (Hess, 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 now, 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 precise, 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 "large ions."

The small ions—those formed by cosmic rays—are the most mobile. Because they are so small, they move rapidly through the air—with a speed of about 1

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. Then, 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 throws all kinds of pollution into the air—than there is over water. It is not surprising that from day to day, from moment to moment, 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 micromicroamperes 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.

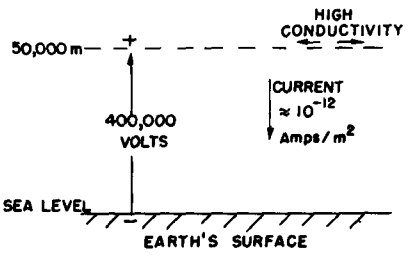


Fig. 9–4. Typical electrical conditions in a clear atmosphere.

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.

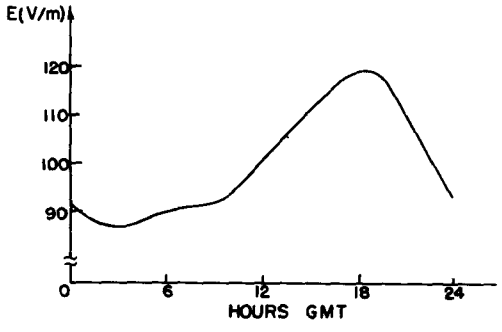


Fig. 9–5. The average daily variation of the atmospheric potential gradient on a clear day over the oceans; referred to Greenwich time.

9–3 Origin of the atmospheric currents

We must next 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. 9–6. It is the thunderstorm and its lightning. It turns out that the bolts of lightning do not "discharge" the potential we have been talking about (as you might at first guess). Lightning storms carry *negative* charges to the earth. When a lightning bolt strikes, ten-to-one it brings down negative charges to the earth in large amounts. It is the thunderstorms throughout the world that are charging the earth with an average of 1800 amperes, which is then being discharged through regions of fair weather.

There are about 300 thunderstorms per day all over the earth, and we can think of them as batteries pumping the electricity to the upper layer and maintaining the voltage difference. Then take into account the geography of the earth—there are thunderstorms in the afternoon 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 and were made only *after* it was known that the variation should have occurred. These things are very difficult because we don't have enough observations on the seas and over all parts of the world to know the number of thunderstorms accurately. But those people who think they "do it right" obtain the result that there is a peak in the activity at 7:00 P.M. Greenwich Mean Time.

Fig. 9–6. 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 about 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 important 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.

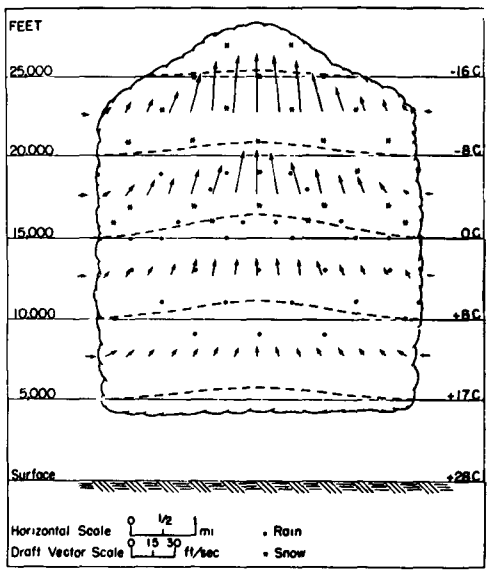


Fig. 9–7. A thunderstorm cell in the early stages of development. [From U.S. Department of Commerce Weather Bureau Report, June 1949.]

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 altitudes 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 mechanically unstable and the warm air rises." Of course, if the temperature is different at different heights, the air *is* unstable *thermodynamically*. Left to itself infinitely 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 *colder* 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.

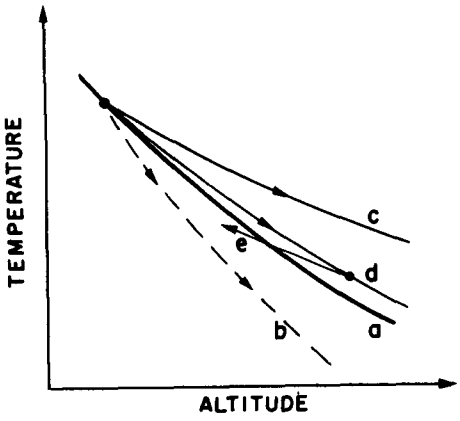


Fig. 9–8. Atmospheric temperature. (a) Static atmosphere; (b) adiabatic cooling of dry air; (c) adiabatic cooling of wet air; (d) wet air with some mixing of ambient air.

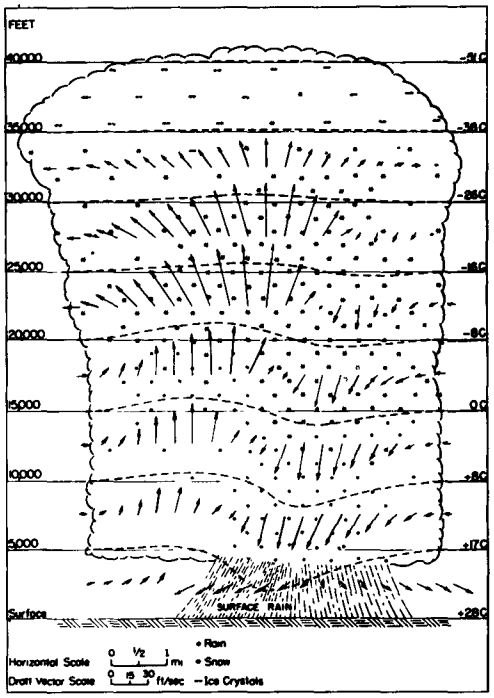


Fig. 9–9. A mature thunderstorm cell. [From U.S. Department of Commerce Weather Bureau Report, June 1949.]

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 enormous heights. This is the machinery 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 entrains air from the environment and is cooled 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 "mature" 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 carried 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 piece 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 becomes attached 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 in 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 gets down far enough, as indicated by curve (e) in the figure. The moment it does that, it is denser than the environment and continues 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 rain 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 when it arrives at the earth's surface. So just before the rain comes there is a certain little cold wind that gives

us a forewarning of the coming storm. In the storm itself there are rapid and irregular gusts of air, there is an enormous turbulence in the cloud, 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 usually spreads into an anvil shape. The cell comes to the end of its life.

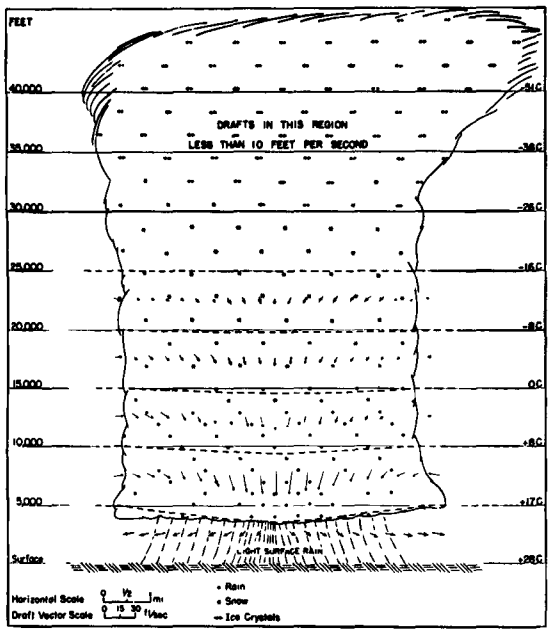


Fig. 9–10. The late phase of a thunderstorm cell. [From U.S. Department of Commerce Weather Bureau Report, June 1949.]

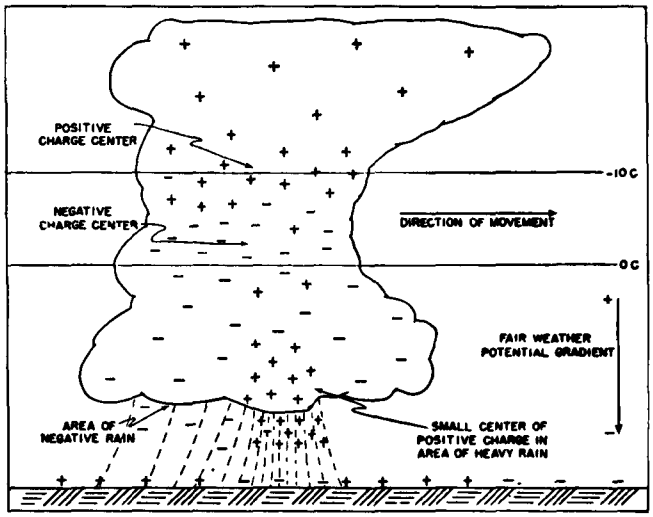


Fig. 9–11. The distribution of electrical charges in a mature thunderstorm cell. [From U.S. Department of Commerce Weather Bureau Report, June 1949.]

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 men!)—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 20, or 30, or even 100 million 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 thunderstorm only 5 seconds after each lightning stroke to build its charge up again. That doesn't necessarily mean that another stroke is going to occur in exactly 5 seconds every time, because, of course, 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.

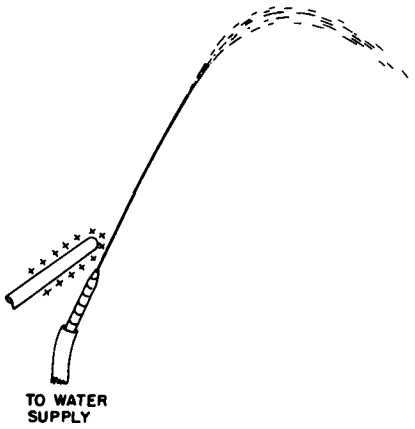


Fig. 9–12. A jet of water with an electric field near the nozzle.

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 strong 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 thunderstorm 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 large-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 breakup 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 stream fall on a large thin metal plate. The larger drops make a louder 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 drops begin? One of the older theories is called the "breaking-drop" theory. Somebody discovered that if you have a drop of water that breaks 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 concoct 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 an induced dipole moment—with the bottom of the drop positive and the top of the drop negative, as drawn in Fig. 9–13. Now there are in the air the "nuclei" 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 ions away if their motion through the air is slow enough. Thus the positive ions cannot attach at the top either. This 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 polarization in a 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.

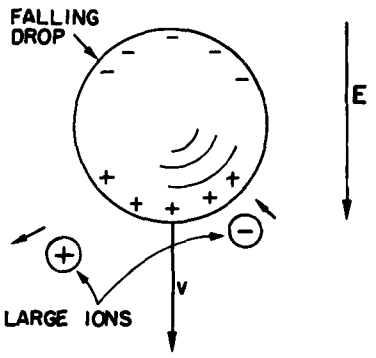


Fig. 9–13. C. T. R. Wilson's theory of charge separation in a thundercloud.

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 at that

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 on 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 drops on ice 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 strokes. 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!

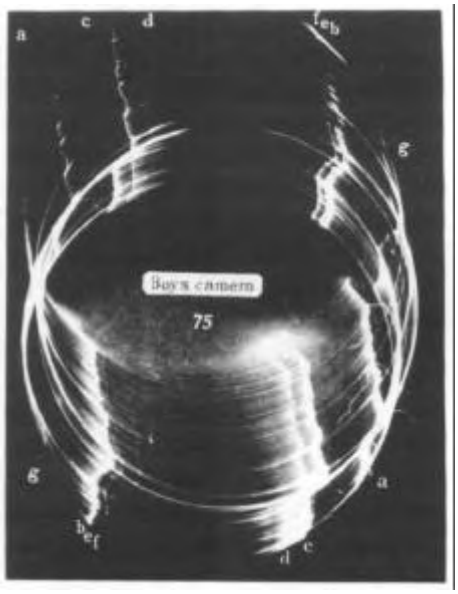


Fig. 9–14. Photograph of a lightning flash taken with a "Boys" camera. [From Schonland, Malan, and Collens, *Proc. Roy. Soc. London*, Vol. 152 (1935).]

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.

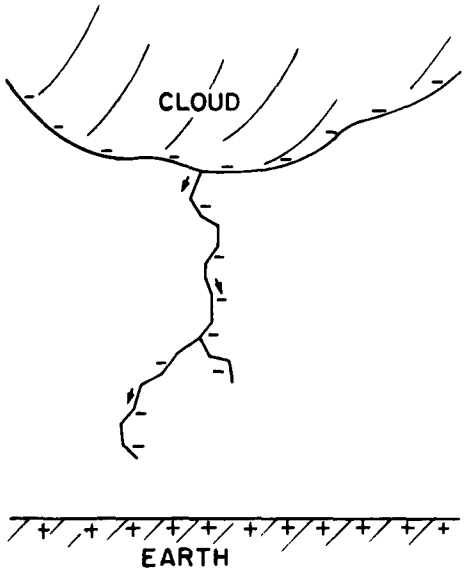


Fig. 9–15. The formation of the "step leader."

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 not 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 "wire" 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 column 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—by 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 pauses. It is called a "dark leader" this time, and it goes all the way down—from top to bottom in one swoop. It goes full steam 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—zing!—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—once as many as 42 times on the same track was seen—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 *two* steps—both toward 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 dumping 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, emptying 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 brush-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 Artabanis, the advisor to Xerxes, giving his master advice on a contemplated attack on the Greeks—during Xerxes' campaign to bring the entire known world under the control of the Persians. Artabanis said, "See how God with his lightning always smites the bigger animals and will not suffer them to wax insolent, while these of a lesser bulk chafe him not. How likewise his bolts fall ever on the highest houses and tallest trees." And then he explains the reason: "So, plainly, doth he love to bring down everything that exalts 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 Artabanis 2300 years ago? Do not exalt yourself. You could only do it less poetically.

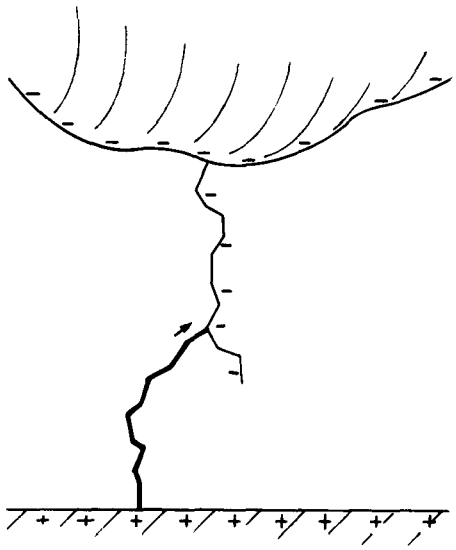


Fig. 9–16. The return lightning stroke runs back up the path made by the leader.

10

Dielectrics

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

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 an electric field 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 electroscope 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

$$C = \frac{\epsilon_0 A}{d}, \tag{10.1}$$

and the charge and voltage on the capacitor are related by

$$Q = CV. \tag{10.2}$$

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.

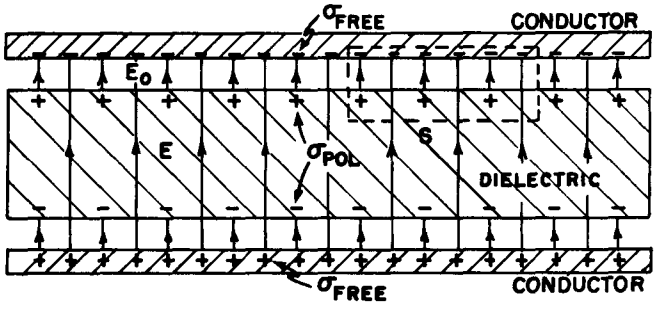


Fig. 10–1. A parallel-plate capacitor with a dielectric. The lines of *E* are shown.

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

be lower than it would be without the material. There is only one possible conclusion, and that is that there must be positive charges 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.

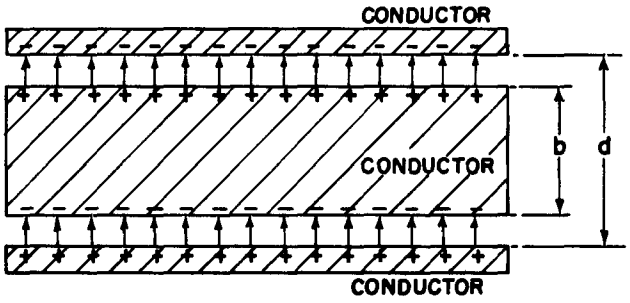


Fig. 10–2. If we put a conducting plate in the gap of a parallel-plate condenser, the induced charges reduce the field in the conductor to zero.

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. (10.1), with $(d - b)$ substituted for d:

$$C = \frac{\epsilon_0 A}{d[1 - (b/d)]}. \tag{10.3}$$

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 insulated 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.

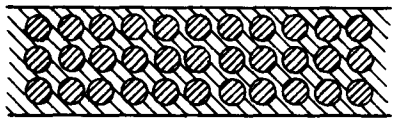


Fig. 10–3. A model of a dielectric: small conducting spheres embedded in an idealized insulator.

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 essential. 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.

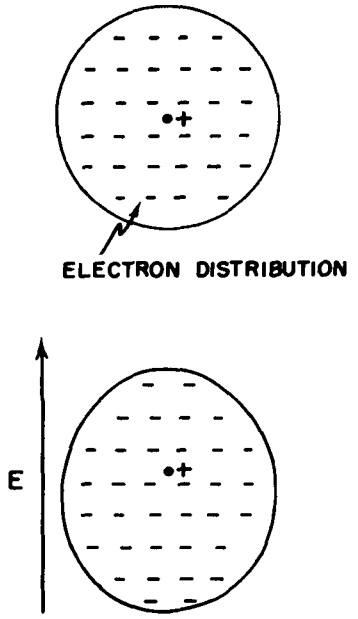


Fig. 10–4. An atom in an electric field has its distribution of electrons displaced with respect to the nucleus.

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 δ, so that $q\delta$ is the dipole moment per atom. (We use δ 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 $Nq\delta$. This dipole moment per unit volume will be represented by a vector, $\boldsymbol{P}$. Needless to say, it is in the direction of the individual dipole moments, i.e., in the direction of the charge separation δ:

$$\boldsymbol{P} = Nq\boldsymbol{\delta}. \tag{10.4}$$

In general, $\boldsymbol{P}$ will vary from place to place in the dielectric. However, at any point in the material, $\boldsymbol{P}$ is proportional to the electric field $\boldsymbol{E}$. 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 $\boldsymbol{P}$ 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 $\boldsymbol{P}$ 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 $\boldsymbol{P}$ 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 δ; at the other surface they have moved in, leaving some positive charge effectively out a distance δ. As shown in Fig. 10–5, we will have a surface density of charge, which will be called the surface *polarization charge*.

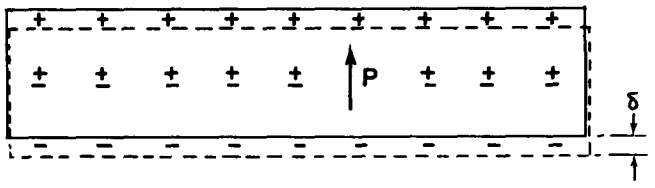


Fig. 10–5. A dielectric slab in a uniform field. The positive charges displaced the distance δ with respect to the negatives.

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

$$\sigma_{\text{pol}} = Nq_e\,\delta.$$

But this is just equal to the magnitude P of the polarization vector $\boldsymbol{P}$, Eq. (10.4):

$$\sigma_{\text{pol}} = P. \tag{10.5}$$

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_{free}, because they can move "freely" anywhere on the conductor. This is, of course, the charge that we put on when we charged the capacitor. It should be emphasized that σ_{pol} exists only because of σ_{free}. If σ_{free} is removed by discharging the capacitor, then σ_{pol} 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_{pol} and σ_{free} have opposite signs, so

$$E = \frac{\sigma_{\text{free}} - \sigma_{\text{pol}}}{\epsilon_0}. \tag{10.6}$$

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_{free} 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

$$E = \frac{\sigma_{\text{free}} - P}{\epsilon_0}. \tag{10.7}$$

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

$$\boldsymbol{P} = \chi\epsilon_0\boldsymbol{E}. \tag{10.8}$$

The constant χ (Greek "khi") is called the *electric susceptibility* of the dielectric.

Then Eq. (10.7) becomes

$$E = \frac{\sigma_{\text{free}}}{\epsilon_0}\frac{1}{(1+\chi)}, \tag{10.9}$$

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

$$V = Ed = \frac{\sigma_{\text{free}}d}{\epsilon_0(1+\chi)}.$$

The total charge on the capacitor is $\sigma_{\text{free}}A$, so that the capacitance defined by (10.2) becomes

$$C = \frac{\epsilon_0 A(1+\chi)}{d} = \frac{\kappa\epsilon_0 A}{d}. \tag{10.10}$$

We have explained the observed facts. When a parallel-plate capacitor is filled with a dielectric, the capacitance is increased by the factor

$$\kappa = 1 + \chi, \tag{10.11}$$

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 comes about.

Let's now consider something a little bit more complicated—the situation in which the polarization $\boldsymbol{P}$ is not everywhere the same. As mentioned earlier, if the polarization is not constant, we would expect in general to find a charge density in the volume, because more charge might come 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 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 $\boldsymbol{P}$ *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

$$\sigma_{\text{pol}} = \boldsymbol{P} \cdot \boldsymbol{n}. \tag{10.12}$$

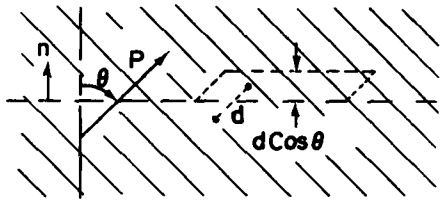


Fig. 10–6. The charge moved across an element of an imaginary surface in a dielectric is proportional to the component of $\boldsymbol{P}$ normal to the surface.

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 *volume* charge density. The total charge displaced *out* of any volume V by the polarization is the integral of the outward normal component of $\boldsymbol{P}$ 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

$$\Delta Q_{\text{pol}} = -\int_S \boldsymbol{P} \cdot \boldsymbol{n}\, da. \tag{10.13}$$

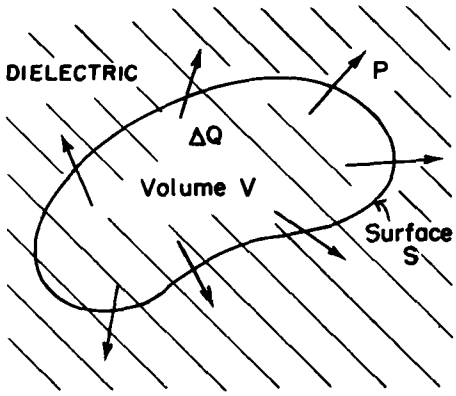


Fig. 10–7. A nonuniform polarization $\boldsymbol{P}$ can result in a net charge in the body of a dielectric.

We can attribute ΔQ_{pol} to a volume distribution of charge with the density ρ_{pol}, and so

$$\Delta Q_{\text{pol}} = \int_V \rho_{\text{pol}}\, dV. \tag{10.14}$$

Combining the two equations yields

$$\int_V \rho_{\text{pol}}\, dV = -\int_S \boldsymbol{P} \cdot \boldsymbol{n}\, da. \tag{10.15}$$

We have a kind of Gauss' theorem that relates the charge density from polarized materials to the polarization vector $\boldsymbol{P}$. We can see that it agrees with the result we got for the surface polarization charge or the dielectric in a parallel-plate capacitor. Using Eq. (10.15) with the gaussian surface of Fig. 10–1, the surface integral gives $P\,\Delta A$, and the charge inside is $\sigma_{\text{pol}}\,\Delta A$, so we get again that $\sigma = P$.

Just as we did for Gauss' law of electrostatics, we can convert Eq. (10.15) to a differential form—using Gauss' mathematical theorem:

$$\int_S \boldsymbol{P} \cdot \boldsymbol{n}\, da = \int_V \nabla \cdot \boldsymbol{P}\, dV.$$

We get

$$\rho_{\text{pol}} = -\nabla \cdot \boldsymbol{P}. \tag{10.16}$$

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.

10–4 The electrostatic equations with dielectrics

Now let's combine the above result with our theory of electrostatics. The fundamental equation is

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0}. \tag{10.17}$$

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_{pol} the charges due to nonuniform polarizations, and call ρ_{free} all the rest. Usually ρ_{free} is the charge we put on conductors, or at known places in space. Equation (10.17) then becomes

$$\nabla \cdot \boldsymbol{E} = \frac{\rho_{\text{free}} + \rho_{\text{pol}}}{\epsilon_0} = \frac{\rho_{\text{free}} - \nabla \cdot \boldsymbol{P}}{\epsilon_0},$$

or

$$\nabla \cdot \left(\boldsymbol{E} + \frac{\boldsymbol{P}}{\epsilon_0}\right) = \frac{\rho_{\text{free}}}{\epsilon_0}. \tag{10.18}$$

Of course, the equation for the curl of $\boldsymbol{E}$ is unchanged:

$$\nabla \times \boldsymbol{E} = 0. \tag{10.19}$$

Taking $\boldsymbol{P}$ from Eq. (10.8), we get the simpler equation

$$\nabla \cdot [(1 + \chi)\boldsymbol{E}] = \nabla \cdot (\kappa \boldsymbol{E}) = \frac{\rho_{\text{free}}}{\epsilon_0}. \tag{10.20}$$

These are the equations of electrostatics 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_{free} is known and the polarization $\boldsymbol{P}$ is proportional to $\boldsymbol{E}$.

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 electrostatics with the charge density ρ_{free} 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 matter 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_{pol} was not appreciated. The charge ρ_{free} was considered to be the entire charge density. In order to write Maxwell's equations in a simple form, a new vector $\boldsymbol{D}$ was defined to be equal to a linear combination of $\boldsymbol{E}$ and $\boldsymbol{P}$:

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}. \tag{10.21}$$

As a result, Eqs. (10.18) and (10.19) were written in an apparently very simple form:

$$\nabla \cdot \boldsymbol{D} = \rho_{\text{free}}, \qquad \nabla \times \boldsymbol{E} = 0. \tag{10.22}$$

Can one solve these? Only if a third equation is given for the relationship between $\boldsymbol{D}$ and $\boldsymbol{E}$. When Eq. (10.8) holds, this relationship is

$$\boldsymbol{D} = \epsilon_0(1 + \chi)\boldsymbol{E} = \kappa \epsilon_0 \boldsymbol{E}. \tag{10.23}$$

This equation was usually written

$$\boldsymbol{D} = \epsilon \boldsymbol{E}, \tag{10.24}$$

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,

$$\epsilon = \kappa \epsilon_0 = (1 + \chi)\epsilon_0. \tag{10.25}$$

Today we look upon 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 some 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 $\boldsymbol{D} = \epsilon \boldsymbol{E}$ is an attempt to describe a property of matter. But matter is extremely complicated, and such an equation is in fact not correct. For instance, if $\boldsymbol{E}$ gets too large, then $\boldsymbol{D}$ is no longer proportional to $\boldsymbol{E}$. For some substances, the proportionality breaks down even with relatively small fields. Also, the "constant" of proportionality may depend on how fast $\boldsymbol{E}$ 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 $\boldsymbol{E}$, (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

$$\nabla \cdot \boldsymbol{E}_0 = \frac{\rho_{\text{free}}}{\epsilon_0} \quad \text{and} \quad \nabla \times \boldsymbol{E}_0 = 0.$$

With the dielectric present, the first of these equations is modified; we have instead the equations

$$\nabla \cdot (\kappa \boldsymbol{E}) = \frac{\rho_{\text{free}}}{\epsilon_0} \quad \text{and} \quad \nabla \times \boldsymbol{E} = 0. \tag{10.26}$$

Now since we are taking κ to be everywhere the same, the last two equations can be written as

$$\nabla \cdot (\kappa \boldsymbol{E}) = \frac{\rho_{\text{free}}}{\epsilon_0} \quad \text{and} \quad \nabla \times (\kappa \boldsymbol{E}) = 0. \tag{10.27}$$

We therefore have the same equations for $\kappa \boldsymbol{E}$ as for $\boldsymbol{E}_0$, so they have the solution $\kappa \boldsymbol{E} = \boldsymbol{E}_0$. In other words, the field is everywhere smaller, 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, Eq. (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,

$$F_x = -\frac{\partial U}{\partial x} = -\frac{Q^2}{2}\frac{\partial}{\partial x}\left(\frac{1}{C}\right). \tag{10.28}$$

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 motion of conductors that are embedded in solid dielectric changes the mechanical stress conditions of the dielectric and alters its electrical

properties, as well as causing 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.

Many older books on electricity start with the "fundamental" law that the force between two charges is

$$F = \frac{q_1 q_2}{4\pi\epsilon_0 \kappa r^2}, \tag{10.29}$$

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 κ is a constant, which is only approximately true for most real materials. 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 sense, indeterminate. 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 got 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 casually 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 then the like charges repel. But that doesn't answer the original question. Why did the paper come toward the comb in the first place?

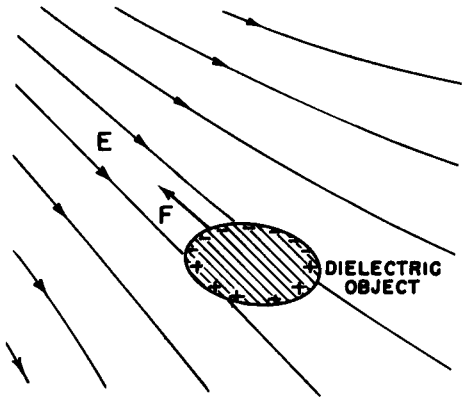


Fig. 10–8. A dielectric object in a nonuniform field feels a force toward regions of higher field strength.

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 fields, 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-

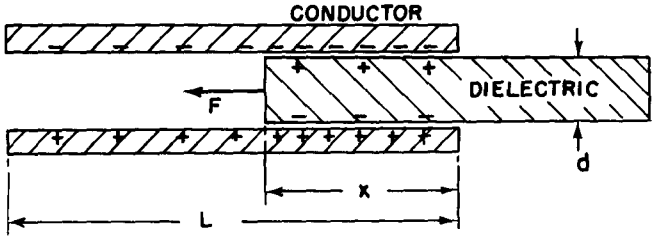


Fig. 10–9. The force on a dielectric sheet in a parallel-plate capacitor can be computed by applying the principle of energy conservation.

rived earlier. Equation (10.28) is equivalent to

$$F_x = -\frac{\partial U}{\partial x} = +\frac{V^2}{2}\frac{\partial C}{\partial x}. \tag{10.30}$$

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 W, that the plate separation and dielectric thickness are d, and that the distance to 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 $\kappa\epsilon_0 V/d$. So the total charge on the plates is

$$Q = \frac{\kappa\epsilon_0 V}{d}xW + \frac{\epsilon_0 V}{d}(L - x)W,$$

from which we get the capacitance:

$$C = \frac{\epsilon_0 W}{d}(\kappa x + L - x). \tag{10.31}$$

Using (10.30), we have

$$F_x = \frac{V^2}{2}\frac{\epsilon_0 W}{d}(\kappa - 1). \tag{10.32}$$

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 there would be in the present case.

Our discussion of the theory of dielectrics has dealt only with electrical phenomena, accepting the fact that the material has a polarization which is proportional 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 circumstances to obtain detailed information about atomic or molecular structure. This aspect will be treated in part in the next chapter.

11

Inside Dielectrics

11–1 Molecular dipoles

11–2 Electronic polarization

11–3 Polar molecules; orientation polarization

11–4 Electric fields in cavities of a dielectric

11–5 The dielectric constant of liquids; the Clausius-Mossotti equation

11–6 Solid dielectrics

11–7 Ferroelectricity; $BaTiO_3$

Review: Chapter 31, Vol. I, *The Origin of the Refractive Index*
Chapter 40, Vol. I, *The Principles of Statistical Mechanics*

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

$$\kappa - 1 = \frac{P}{\epsilon_0 E}. \tag{11.1}$$

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 material. We begin with the simplest possible example—the polarization of gases. But even gases already have complications: there are two types. The molecules of some gases, like oxygen, which has a symmetric pair of atoms in each molecule, have no inherent dipole moment. But the molecules of others, like water vapor (which has a nonsymmetric 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 forms of the two types of molecules are sketched in Fig. 11–1.

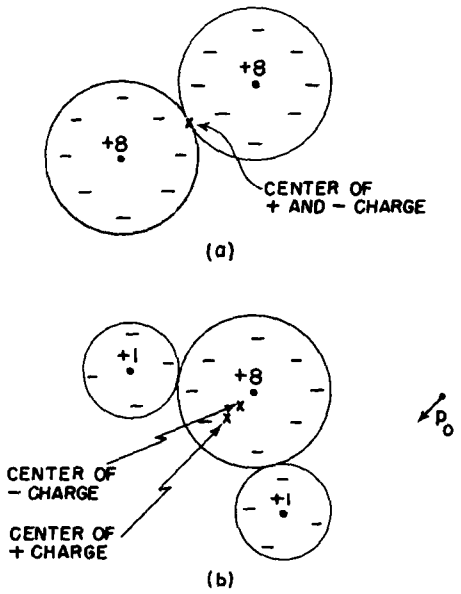


Fig. 11–1. (a) An oxygen molecule with zero dipole moment. (b) The water molecule has a permanent dipole moment p_0.

11–2 Electronic polarization

We will first discuss the polarization of nonpolar molecules. We can start with the simplest case of a monatomic 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 must do now is exactly the same as we did then. 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 obeys the equation

$$m\frac{d^2x}{dt^2} + m\omega_0^2 x = q_e E. \tag{11.2}$$

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

$$x = \frac{q_e E}{m(\omega_0^2 - \omega^2)}, \tag{11.3}$$

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

$$x = \frac{q_e E}{m\omega_0^2}. \tag{11.4}$$

From this we see that the dipole moment p of a single atom is

$$p = q_e x = \frac{q_e^2 E}{m\omega_0^2}. \tag{11.5}$$

In this theory the dipole moment p is indeed proportional to the electric field.

People usually write

$$\boldsymbol{p} = \alpha\epsilon_0 \boldsymbol{E}. \tag{11.6}$$

(Again the ϵ_0 is put in for historical reasons.) The constant α is called the polarizability of the atom, and has the dimensions L^3. It is a measure of how easy it is to induce a moment in an atom with an electric field. Comparing (11.5) and (11.6), our simple theory says that

$$\alpha = \frac{q_e^2}{\epsilon_0 m\omega_0^2} = \frac{4\pi e^2}{m\omega_0^2}. \tag{11.7}$$

If there are N atoms in a unit volume, the polarization P—the dipole moment per unit volume—is given by

$$\boldsymbol{P} = N\boldsymbol{p} = N\alpha\epsilon_0 \boldsymbol{E}. \tag{11.8}$$

Putting (11.1) and (11.8) together, we get

$$\kappa - 1 = \frac{P}{\epsilon_0 E} = N\alpha \tag{11.9}$$

or, using (11.7),

$$\kappa - 1 = \frac{4\pi N e^2}{m\omega_0^2}. \tag{11.10}$$

From Eq. (11.9) we would predict that the dielectric constant κ 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 38, Vol. I) that the energy needed to ionize the hydrogen atom should be approximately

$$E \approx \frac{1}{2}\frac{me^4}{\hbar^2}. \tag{11.11}$$

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

$$\omega_0 \approx \frac{1}{2}\frac{me^4}{\hbar^3}.$$

If we now use this value of ω_0 in Eq. (11.7), we find for the electronic polarizability

$$\alpha \approx 16\pi\left[\frac{\hbar^2}{me^2}\right]^3. \tag{11.12}$$

The quantity $(\hbar^2/me^2)$ is the radius of the ground-state orbit of a Bohr atom (see Chapter 38, Vol. I) and equals 0.528 angstroms. In a gas at standard pressure and temperature (1 atmosphere, 0°C) there are 2.69×10^{19} atoms/cm^3, so Eq. (11.9) gives us

$$\kappa = 1 + (2.69 \times 10^{19})16\pi\,(0.528 \times 10^{-8})^3 = 1.00020. \tag{11.13}$$

The dielectric constant for hydrogen gas is measured to be

$$\kappa_{\text{exp}} = 1.00026.$$

We see that our theory is about right. We should not expect any better, because the measurements were, of course, made with normal 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 α for hydrogen atoms gives a result about 12% higher than (11.12) (the 16π is changed to 18π), 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 13.5 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 α would be one-quarter as large. We expect that

$$\kappa_{\text{helium}} \approx 1.000050.$$

Experimentally,

$$\kappa_{\text{helium}} = 1.000068,$$

so you see that our rough estimates are coming out on the right track. So we have understood the dielectric constant of nonpolar gas, but only qualitatively, because we have not yet used a correct atomic theory of the motions of the atomic electrons.

11–3 Polar molecules; orientation polarization

Next we will consider a molecule which carries a permanent dipole moment p_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 an extra dipole moment induced because of the forces on the electrons; this part gives just the same kind of electronic polarizability 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 moment 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. At ordinary temperatures and electric fields the collisions 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.

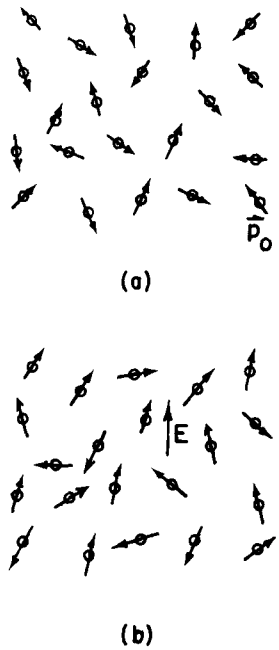


Fig. 11–2. (a) In a gas of polar molecules, the individual moments are oriented at random; the average moment in a small volume is zero. (b) When there is an electric field, there is some average alignment of the molecules.

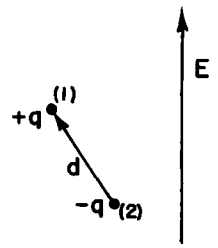


Fig. 11–3. The energy of a dipole p_0 in the field E is $-p_0 \cdot E$.

To use this method we need to know the energy of a dipole in an electric field. Consider a dipole of moment p_0 in an electric field, as shown in Fig. 11–3. 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

$$U = q\phi(1) - q\phi(2) = q\boldsymbol{d} \cdot \nabla\phi,$$

or

$$U = -\boldsymbol{p}_0 \cdot \boldsymbol{E} = -p_0 E \cos\theta, \tag{11.14}$$

where θ is the angle between $\boldsymbol{p}_0$ and $\boldsymbol{E}$. 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. I that in a state of thermal equilibrium, the relative number of molecules with the potential energy U is proportional to

$$e^{-U/kT}, \tag{11.15}$$

where $U(x, y, z)$ 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

$$n(\theta) = n_0 e^{+p_0 E\cos\theta/kT}. \tag{11.16}$$

For normal temperatures and fields, the exponent is small, so we can approximate by expanding the exponential:

$$n(\theta) = n_0\left(1 + \frac{p_0 E\cos\theta}{kT}\right). \tag{11.17}$$

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

$$n_0 = \frac{N}{4\pi}. \tag{11.18}$$

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 $\boldsymbol{E}$, we will just sum the components in that direction (the components at right angles to $\boldsymbol{E}$ will sum to zero):

$$P = \sum_{\substack{\text{unit}\\ \text{volume}}} p_0 \cos\theta_i.$$

We can evaluate the sum by integrating over the angular distribution. The solid angle at θ is $2\pi\sin\theta\,d\theta$, so

$$P = \int_0^\pi n(\theta) p_0 \cos\theta\, 2\pi\sin\theta\, d\theta. \tag{11.19}$$

Substituting for $n(\theta)$ from (11.17), we have

$$P = -\frac{N}{2}\int_0^\pi \left(1 + \frac{p_0 E}{kT}\cos\theta\right) p_0 \cos\theta\, d(\cos\theta),$$

which is easily integrated to give

$$P = \frac{N p_0^2 E}{3kT}. \tag{11.20}$$

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 p_0, and the mean moment that is produced by the lining up is again proportional to p_0. 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 P directly, but Eq. (11.20) does predict that $\kappa - 1$ should vary inversely as the temperature, and this we should check.

From (11.20) we get

$$\kappa - 1 = \frac{P}{\epsilon_0 E} = \frac{Np_0^2}{3\epsilon_0 kT}, \tag{11.21}$$

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.* [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^{-2} instead of as T^{-1}.] 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—its 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 in the electrons.

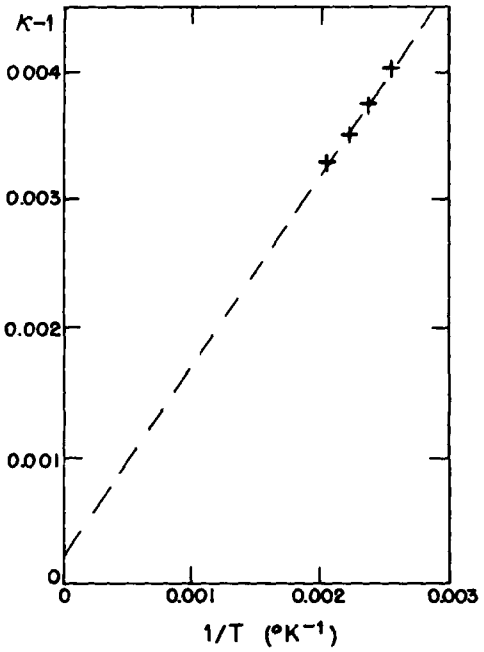


Fig. 11–4. Experimental measurements of the dielectric constant of water vapor at various temperatures.

11–4 Electric fields in cavities of a dielectric

We now turn to an interesting but complicated 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, $\boldsymbol{P}$ 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 $\boldsymbol{E}$ 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 $\nabla \times \boldsymbol{E} = 0$, the line integral of $\boldsymbol{E}$ around the curve, Γ, which goes as shown in (b) of the figure, should

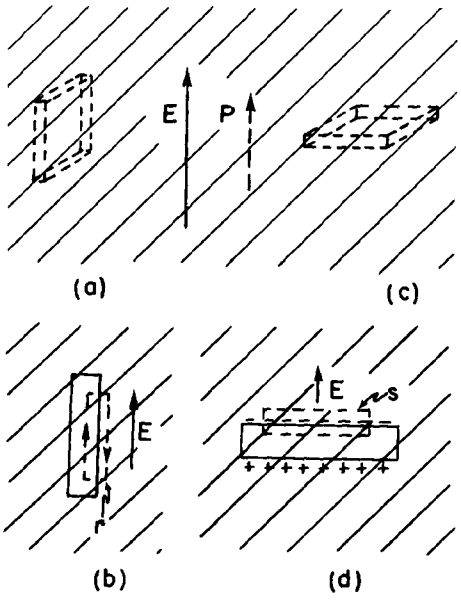


Fig. 11–5. The field in a slot cut in a dielectric depends on the shape and orientation of the slot.

* Sänger, Steiger, and Gächter, *Helvetica Physica Acta* **5**, 200 (1932).

be zero. The field inside the slot must 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 large sides 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

$$E_0 = E + \frac{P}{\epsilon_0}, \tag{11.22}$$

where E is again the electric field in the dielectric. (The gaussian surface contains the surface polarization charge $\sigma_{\text{pol}} = P$.) 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.

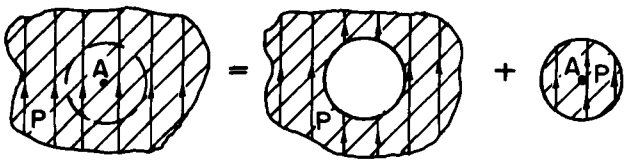


Fig. 11–6. The field at any point A in a dielectric can be considered as the sum of the field in a spherical hole plus the field due to a spherical plug.

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

$$E = E_{\text{hole}} + E_{\text{plug}}, \tag{11.23}$$

where E_{hole} is the field in the hole and E_{plug} 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

$$E_{\text{plug}} = -\frac{P}{3\epsilon_0}. \tag{11.24}$$

Using (11.23), we get

$$E_{\text{hole}} = E + \frac{P}{3\epsilon_0}. \tag{11.25}$$

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.)

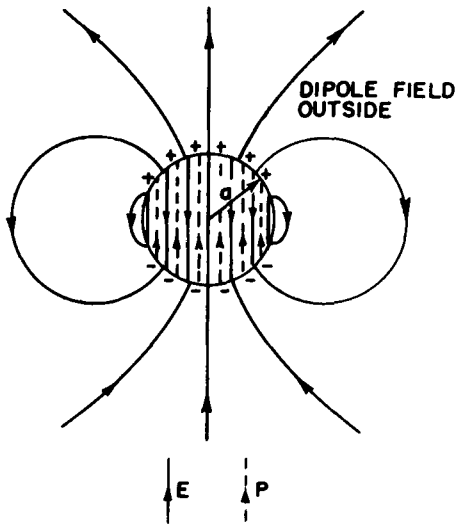


Fig. 11–7. The electric field of a uniformly polarized sphere.

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 E_{hole} of (11.25) for the polarizing field in

Eq. (11.6), then Eq. (11.8) becomes

$$P = N\alpha\epsilon_0\left(E + \frac{P}{3\epsilon_0}\right), \tag{11.26}$$

or

$$P = \frac{N\alpha}{1 - (N\alpha/3)}\,\epsilon_0 E. \tag{11.27}$$

Remembering that $\kappa - 1$ is just $P/\epsilon_0 E$, we have

$$\kappa - 1 = \frac{N\alpha}{1 - (N\alpha/3)}, \tag{11.28}$$

which gives us the dielectric constant of a liquid in terms of α, the atomic polarizability. This is called the *Clausius-Mossotti* equation.

Whenever $N\alpha$ is very small, as it is for a gas (because the density N is small), then the term $N\alpha/3$ can be neglected compared with 1, and we get our old result, Eq. (11.9), that

$$\kappa - 1 = N\alpha. \tag{11.29}$$

Let's compare Eq. (11.28) with some experimental results. It is first necessary to look at gases for which, using the measurement of κ, we can find α from Eq. (11.29). For instance, for carbon disulfide at zero degrees centigrade the dielectric constant is 1.0029, so $N\alpha$ 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 CS_2 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—$N\alpha$ in the liquid is equal to 381 times 0.0029, or 1.11. Notice that the $N\alpha/3$ 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 11–1 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 CS_2—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.

	Gas			Liquid				
Substance	κ (exp)	$N\alpha$	Density	Density	Ratio*	$N\alpha$	κ (predict)	κ (exp)
CS_2	1.0029	0.0029	0.00339	1.293	381	1.11	2.76	2.64
O_2	1.000523	0.000523	0.00143	1.19	832	0.435	1.509	1.507
CCl_4	1.0030	0.0030	0.00489	1.59	325	0.977	2.45	2.24
A	1.000545	0.000545	0.00178	1.44	810	0.441	1.517	1.54

* Ratio = density of liquid/density of gas.

Our derivation of Eq. (11.28) is valid only for *electronic* polarization in liquids. It is not right for a polar molecule like H_2O. If we go through the same calculations for water, we get 13.2 for $N\alpha$, which means that the dielectric constant for the liquid is *negative*, while the observed value of κ is 80. The problem has 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.

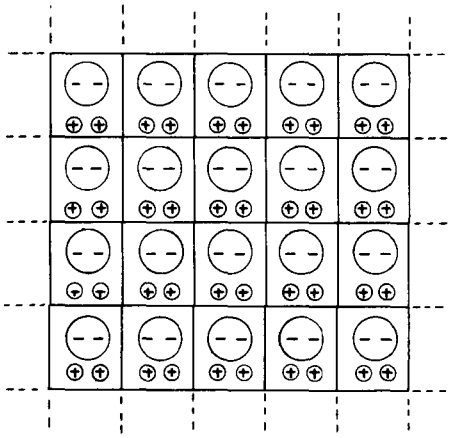


Fig. 11–8. A complex crystal lattice can have a permanent intrinsic polarization P.

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 also contribute to κ. In ionic crystals such as NaCl there is also *ionic 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.

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.

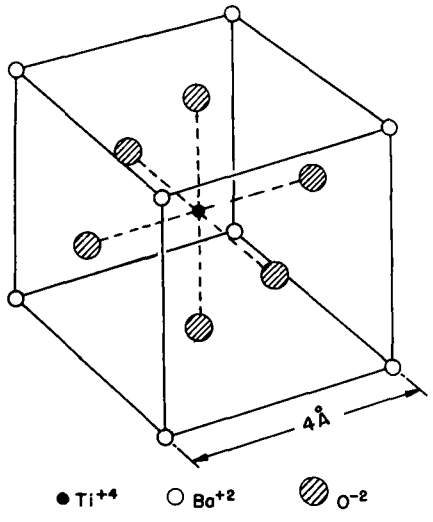


Fig. 11–9. The unit cell of $BaTiO_3$. The atoms really fill up most of the space; for clarity, only the positions of their centers are shown.

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 an enormous dielectric constant. Below this temperature, however, it suddenly takes on a permanent moment.

In working out the polarization of solid material, we must first find what are the local fields in each unit cell. We must include the fields from the polarization

itself, 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 what we would get in a spherical hole. If you work it out for a crystal, you find that the factor 1/3 in Eq. (11.24) becomes slightly different, but not far from 1/3. (For a simple cubic crystal, it is just 1/3.) 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 $N\alpha$ became greater than 3. It appears as though κ would become negative. But that surely cannot be right. Let's see what should happen if we were gradually to increase α in a particular crystal. As α gets larger, the polarization gets bigger, making a bigger local field. But a bigger local field will polarize each atom more, raising the local fields still more. If the "give" 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 $N\alpha = 3$. 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 can move a little within the cubic lattice. 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 dipole 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 $N\alpha$ is decreased just a little bit it comes unstuck. Since N decreases with increasing temperature—because of thermal expansion—we can vary $N\alpha$ 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 lock 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 $N\alpha$ 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

$$N\alpha = 3 - \beta(T - T_c), \tag{11.30}$$

where β is a small constant, of the same order of magnitude as the thermal expansion coefficient, or about 10^{-5} to 10^{-6} per degree C. Now if we substitute this relation into Eq. (11.28), we get that

$$\kappa - 1 = \frac{3 - \beta(T - T_c)}{\beta(T - T_c)/3}.$$

Since we have assumed that $\beta(T - T_c)$ is small compared with one, we can approximate this formula by

$$\kappa - 1 = \frac{9}{\beta(T - T_c)}. \tag{11.31}$$

This relation is right, of course, only for $T > T_c$. We see that just above the critical temperature κ is enormous. Because $N\alpha$ is so close to 3, there is a tremendous magnification effect, and the dielectric constant can easily be as high as 50,000 to 100,000. It is also very sensitive to temperature. For increases in temperature, the dielectric constant goes down inversely as the temperature, but, unlike the case of a dipolar gas, for which $\kappa - 1$ 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

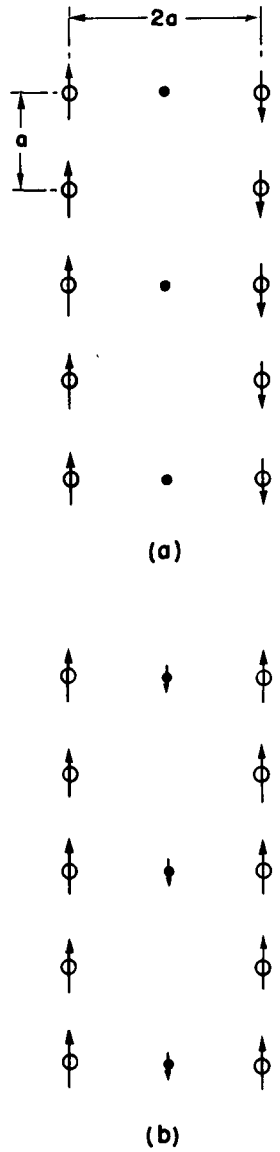


Fig. 11–10. Models of a ferroelectric: (a) corresponds to an antiferroelectric, and (b) to a normal ferroelectric.

to pick out chains of ions along vertical lines. One of them 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 imitate 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 is not a serious 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 is easier to figure out the first time—then when one understands how the thing works, it is 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 must find the sum of 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

$$E = \frac{1}{4\pi\epsilon_0}\frac{2p}{r^3}. \tag{11.32}$$

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

$$E_{\text{chain}} = \frac{p}{4\pi\epsilon_0}\frac{2}{a^3}\cdot\left(2+\frac{2}{8}+\frac{2}{27}+\frac{2}{64}+\cdots\right) = \frac{p}{\epsilon_0}\frac{0.383}{a^3}. \tag{11.33}$$

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 -0.050 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 -0.050 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 p 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

$$p = \alpha\epsilon_0 E_{\text{chain}}$$

and

$$E_{\text{chain}} = \frac{0.383}{a^3}\frac{p}{\epsilon_0}.$$

There are two solutions: E and p both zero, or

$$\alpha = \frac{a^3}{0.383},$$

with E and p both finite. Thus if α is as large as $a^3/0.383$, a permanent polarization sustained 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 2×10^{-8} cm, so we must expect that $\alpha = 21.8 \times 10^{-24}\,\text{cm}^3$. We can compare this with the known polarizabilities of the individual atoms. For oxygen, $\alpha = 30.2 \times 10^{-24}\,\text{cm}^3$; we're on the right track! But for titanium, $\alpha = 2.4 \times 10^{-24}\,\text{cm}^3$; 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 α(average) $= 16.3 \times 10^{-24}$, 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 9.2×10^{-24} cm^3. (A more precise computation using alternating atoms shows that actually 11.9×10^{-24} is needed.) To understand the properties of $BaTiO_3$, 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 out. 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 effects, there are still certain thermodynamic effects one could observe.) Such systems exist, and are called antiferroelectric. So what we have explained is really an 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 antiparallel to the oxygen-titanium chains. The small fields produced at the next oxygen-titanium chain will get it started parallel to the first. So $BaTiO_3$ 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 detailed report on our present understanding of the dielectric constants of gases, of liquids, and of solids.

12

Electrostatic Analogs

12–1 The same equations have the same solutions

12–2 The flow of heat; a point source near an infinite plane boundary

12–3 The stretched membrane

12–4 The diffusion of neutrons; a uniform spherical source in a homogeneous medium

12–5 Irrotational 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 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, really basically 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

$$\nabla \cdot (\kappa \boldsymbol{E}) = \frac{\rho_{\text{free}}}{\epsilon_0}, \tag{12.1}$$

$$\nabla \times \boldsymbol{E} = 0. \tag{12.2}$$

(We take the equations of electrostatics with dielectrics so as to have the most general situation.) The same physics can be expressed in another mathematical form:

$$\boldsymbol{E} = -\nabla\phi, \tag{12.3}$$

$$\nabla \cdot (\kappa \nabla\phi) = -\frac{\rho_{\text{free}}}{\epsilon_0}. \tag{12.4}$$

Now the point is that there are many physics problems whose mathematical equations have the same form. There is a potential (ϕ) whose gradient multiplied by a scalar function (κ) 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 flow of heat; a point source near an infinite plane boundary

We have discussed one example earlier (Section 3–4)—the flow of heat. Imagine a block of material, which need 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 $\boldsymbol{h}$. It represents the amount of heat energy which flows per unit time through a unit area perpendicular to the flow. The divergence of $\boldsymbol{h}$ represents the rate per unit volume at which heat is leaving a region:

$$\nabla \cdot \boldsymbol{h} = \text{rate of heat out per unit volume.}$$

(We could, of course, write the equation in integral form—just as we did in electrostatics 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 electrostatics.)

The rate at which heat is generated or absorbed at various places depends, of course, on the problem. Suppose, for example, that there is a source of heat inside the material (perhaps a radioactive source, or a resistor heated by an electrical 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 thermal energy to other internal energies in the volume. If u is the internal energy per unit volume, $-du/dt$ will also be a "source" of heat energy. We have, then,

$$\nabla \cdot \boldsymbol{h} = s - \frac{du}{dt}. \tag{12.5}$$

We are not going to discuss just now the complete equation in which things change with time, because we are making an analogy to electrostatics, 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,

$$\nabla \cdot \boldsymbol{h} = s. \tag{12.6}$$

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*.

$$\boldsymbol{h} = -K \nabla T. \tag{12.7}$$

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

$$\nabla \cdot (K \nabla T) = -s, \tag{12.8}$$

which has exactly the same form as (12.4). *Steady heat-flow problems and electrostatic problems are the same.* The heat flow vector $\boldsymbol{h}$ corresponds to $\boldsymbol{E}$, and the temperature T corresponds to ϕ. We have already noticed 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 statements from electrostatics 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 wire, or steampipe, or whatever it is in the center. Let the total amount of heat lost from a length L of the pipe be called G—which is what we are trying to find.

How can we solve this problem? We have the differential equations, but since these are the same as those of electrostatics, 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 concentric layer of dielectric material in between, as drawn in Fig. 12–1(b). Now since the heat flow $\boldsymbol{h}$ corresponds to the electric field $\boldsymbol{E}$, the quantity G 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 ϵ_0). 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 h 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 h 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 G:

$$2\pi rLh = G \qquad \text{or} \qquad h = \frac{G}{2\pi rL}. \tag{12.9}$$

The heat flow is proportional to the temperature gradient:

$$\boldsymbol{h} = -K\,\nabla T,$$

or, in this case, the magnitude of $\boldsymbol{h}$ is

$$h = -K\frac{dT}{dr}.$$

This, together with (12.9), gives

$$\frac{dT}{dr} = -\frac{G}{2\pi KLr}. \tag{12.10}$$

Integrating from $r = a$ to $r = b$, we get

$$T_2 - T_1 = -\frac{G}{2\pi KL}\ln\frac{b}{a}. \tag{12.11}$$

Solving for G, we find

$$G = \frac{2\pi KL(T_1 - T_2)}{\ln\,(b/a)}. \tag{12.12}$$

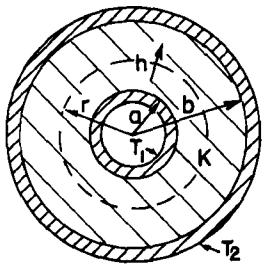
This result corresponds exactly to the result for the charge on a cylindrical condenser:

$$Q = \frac{2\pi\epsilon_0 L(\phi_1 - \phi_2)}{\ln\,(b/a)}.$$

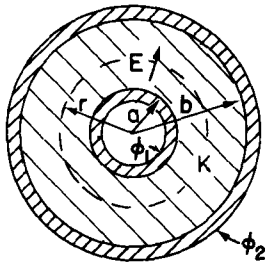
The problems are the same, and they have the same solutions. From our knowledge of electrostatics, we also know how much heat is lost by an insulated pipe.

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 localized 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 G 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 hot 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 κ on opposite sides of a plane boundary. Aha! 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 $\boldsymbol{h}$ 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 $\boldsymbol{E}$ (which is the analog of $\boldsymbol{h}$) 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 *methods*. 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.)

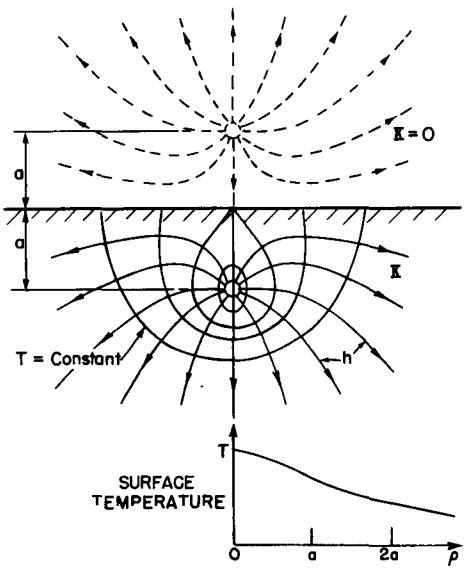


Fig. 12–2. The heat flow and isothermals near a point heat source at the distance a below the surface of a good thermal conductor. An image source is shown outside the material.

How shall we find an electric field that has *no* component perpendicular to the surface? That is, one which is always *tangent* 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 horizontal at the surface. The normal components of the two sources cancel out.

Thus our heat flow 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 in an infinite medium is

$$T = \frac{G}{4\pi Kr}. \tag{12.13}$$

(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

$$T = \frac{G}{4\pi Kr_1} + \frac{G}{4\pi Kr_2}. \tag{12.14}$$

This formula gives us the temperature everywhere in the block. Several isothermal surfaces are shown in Fig. 12–2. Also shown are lines of $\boldsymbol{h}$, which can be obtained from $\boldsymbol{h} = -K\nabla T$.

We originally asked for the temperature distribution on the surface. For a point on the surface at the distance ρ from the axis, $r_1 = r_2 = \sqrt{\rho^2 + a^2}$, so

$$T(\text{surface}) = \frac{1}{4\pi K}\frac{2G}{\sqrt{\rho^2 + a^2}}. \tag{12.15}$$

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 drumhead). 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.

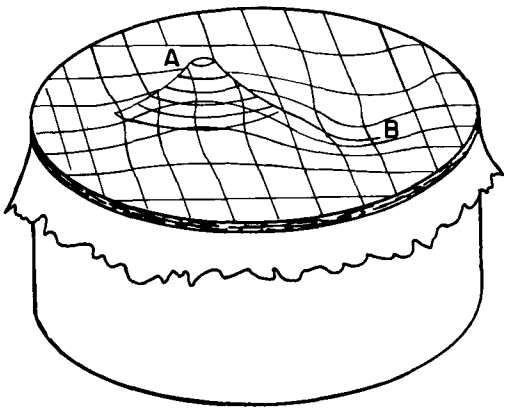


Fig. 12–3. A thin rubber sheet stretched over a cylindrical frame (like a drumhead). If the sheet is pushed up at A and down at B, what is the shape of the surface?

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. 12–4). 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.

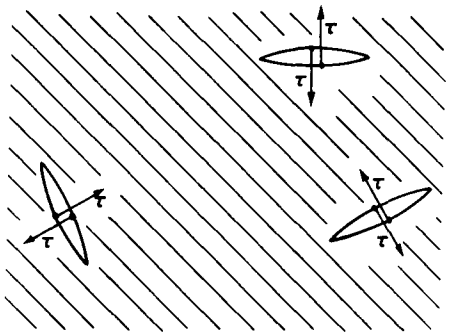


Fig. 12–4. The surface tension τ of a stretched rubber sheet is the force per unit length across a line.

Suppose now that we look at a vertical cross section of the membrane. It will appear as a curve, like the one in Fig. 12–5. 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 shown is parallel to the x-axis.)

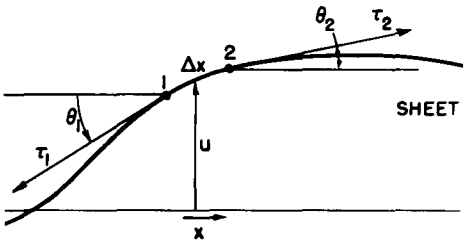


Fig. 12–5. Cross section of the deflected sheet.

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 of 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 θ by tan θ, which can be written as $\partial u/\partial x$. The force is then

$$\Delta F = \left[\tau_2 \left(\frac{\partial u}{\partial x}\right)_2 - \tau_1 \left(\frac{\partial u}{\partial x}\right)_1\right] \Delta y.$$

The quantity in brackets can be equally well written (for small Δx) as

$$\frac{\partial}{\partial x}\left(\tau \frac{\partial u}{\partial x}\right) \Delta x;$$

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 on the other two edges; the total is evidently

$$\Delta F = \left[\frac{\partial}{\partial x}\left(\tau \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\tau \frac{\partial u}{\partial y}\right)\right] \Delta x \, \Delta y. \tag{12.16}$$

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 then be written

$$f = -\nabla \cdot (\tau \nabla u), \tag{12.17}$$

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 $u(x, y)$ to the applied

forces $f(x, y)$ 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. (12.17),

$$\nabla^2 u = -\frac{f}{\tau}. \tag{12.18}$$

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 ρ/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 $\ln(1/r)$. (The *slope*, which corresponds to the electric field E, drops off as $1/r$.)

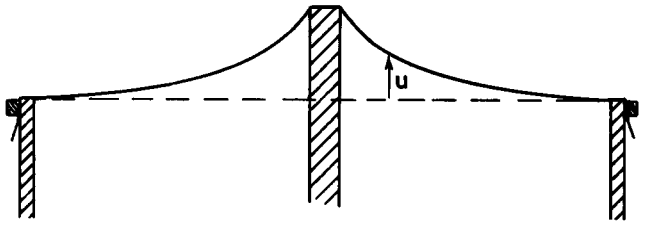


Fig. 12–6. Cross section of a stretched rubber sheet pushed up by a round rod. The function $u(x, y)$ is the same as the electric potential $\phi(x, y)$ near a very long charged rod.

The stretched rubber sheet 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 to 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 limited. 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 43 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 flow*.

Let $N(x, y, z)\,\Delta V$ 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 43 in Vol. I, we describe the flow by a flow vector $\boldsymbol{J}$. Its x-component J_x is the *net* number of neutrons that pass in unit time a unit area perpendicular to the x-direction. We found that

$$J_x = -D\frac{\partial N}{\partial x}, \tag{12.19}$$

where the diffusion constant D is given in terms of the mean velocity v, and the mean-free-path l between scatterings is given by

$$D = \frac{1}{3}lv.$$

The vector equation for $\boldsymbol{J}$ is

$$\boldsymbol{J} = -D\,\nabla N. \tag{12.20}$$

The rate at which neutrons flow across any surface element da is $\boldsymbol{J}\cdot\boldsymbol{n}\,da$ (where, as usual, $\boldsymbol{n}$ is the unit normal). The net flow *out of a volume element* is then (following the usual gaussian argument) $\nabla\cdot\boldsymbol{J}\,dV$. This flow would result in a decrease with time of the number in 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 $(S - \partial N/\partial t)\,\Delta V$. We have then that

$$\nabla\cdot\boldsymbol{J} = S - \frac{\partial N}{\partial t}. \tag{12.21}$$

Combining (12.21) with (12.20), we get the *neutron diffusion equation*

$$\nabla\cdot(-D\,\nabla N) = S - \frac{\partial N}{\partial t}. \tag{12.22}$$

In the static case—where $\partial N/\partial t = 0$—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 all the problems in electrostatics? We can do it *faster* this time because we *have* done the electrostatic problems!)

Suppose we have a block of material 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_0 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 $Q/4\pi\epsilon_0 r$, with the total charge Q given by $4\pi a^3\rho/3$. So

$$\phi_{\text{outside}} = \frac{\rho a^3}{3\epsilon_0 r}. \tag{12.23}$$

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

$$E = \frac{\rho r}{3\epsilon_0}. \tag{12.24}$$

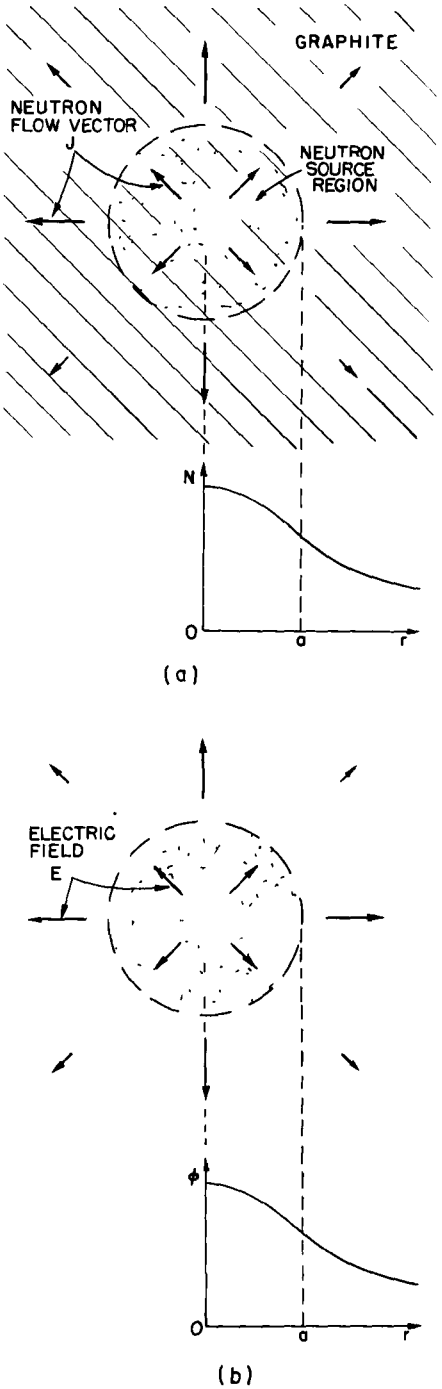


Fig. 12–7. (a) Neutrons are produced uniformly throughout a sphere of radius a in a large graphite block and diffuse outward. The neutron density N is found as a function of r, the distance from the center of the source. (b) The analogous electrostatic situation: a uniform sphere of charge, where N corresponds to ϕ and J corresponds to E.

The field increases linearly with r. Integrating E to get ϕ, we have

$$\phi_{\text{inside}} = -\frac{\rho r^2}{6\epsilon_0} + \text{a constant.}$$

At the radius a, ϕ_{inside} must be the same as ϕ_{outside}, so the constant must be $\rho a^2/2\epsilon_0$. (We are assuming that ϕ is zero at large distances from the source, which will correspond to N being zero for the neutrons.) Therefore,

$$\phi_{\text{inside}} = \frac{\rho}{3\epsilon_0}\left(\frac{3a^2}{2} - \frac{r^2}{2}\right). \tag{12.25}$$

We know immediately the neutron density in our other problem. The answer is

$$N_{\text{outside}} = \frac{Sa^3}{3Dr}, \tag{12.26}$$

and

$$N_{\text{inside}} = \frac{S}{3D}\left(\frac{3a^2}{2} - \frac{r^2}{2}\right). \tag{12.27}$$

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 $3a^2/2$. At the edge ($r = a$) it is proportional to $2a^2/2$, 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 on 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, obeys 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 equations were an approximation which was correct 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*, *circulation-free* liquid. Then we represent the flow by giving the velocity $v(\mathbf{r})$ as a function of position $\mathbf{r}$. If the motion is steady (the only case for which there is an electrostatic analog) $\mathbf{v}$ is independent of time. If ρ is the density of the fluid, then $\rho\mathbf{v}$ is the amount of mass which passes per unit time through a unit area. By the conservation of matter, the divergence of $\rho\mathbf{v}$ 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 $\nabla \cdot \rho\mathbf{v} = 0$. (It should, in general, be equal to $-\partial\rho/\partial t$, but since our fluid is incompressible, ρ cannot change.) Since ρ is everywhere the same, we can factor it out, and our equation is simply

$$\nabla \cdot \mathbf{v} = 0.$$

Good! We have electrostatics again (with no charges); it's just like $\nabla \cdot \mathbf{E} = 0$. Not so! Electrostatics is *not* simply $\nabla \cdot \mathbf{E} = 0$. 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 $\mathbf{v}$ is zero. But that is not generally true for real liquids. Most liquids will ordinarily develop some circulation. So we are restricted 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

$$\nabla \cdot v = 0 \tag{12.28}$$

and

$$\nabla \times v = 0. \tag{12.29}$$

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 John 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 40 and 41.)

Because $\nabla \times v = 0$, the velocity of "dry water" can be written as the gradient of some potential:

$$v = -\nabla\psi. \tag{12.30}$$

What is the physical meaning of ψ? 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, ψ 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

$$\nabla \cdot (\nabla\psi) = \nabla^2\psi = 0. \tag{12.31}$$

The velocity potential ψ obeys the same differential equation as the electrostatic potential in free space ($\rho = 0$).

Let's pick 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 circulation 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 the same. 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 quantative description for the velocity field, i.e., an expression for the velocity at any point P.

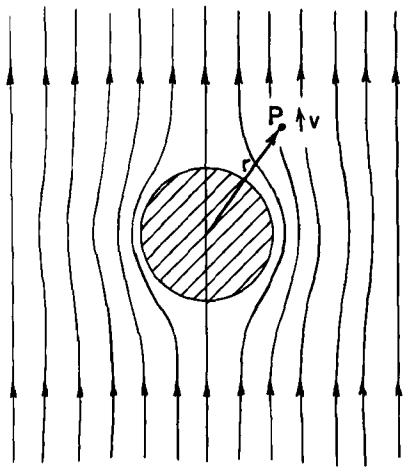


Fig. 12–8. The velocity field of irrotational fluid flow past a sphere.

We can find the velocity from the gradient of ψ, 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 $\partial\psi/\partial r$ is zero at $r = a$. To satisfy (2), we must have $\partial\psi/\partial z = v_0$ 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 ψ, 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,

$$\left.\frac{\partial\phi}{\partial r}\right|_{r=a} = 0. \tag{12.32}$$

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 $\boldsymbol{E}$ 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 $pz/4\pi\epsilon_0 r^3$. Thus we assume that

$$\phi = -E_0 z + \frac{pz}{4\pi\epsilon_0 r^3}. \tag{12.33}$$

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? To find out, we may use the other condition on ϕ, Eq. (12.32). We must differentiate ϕ with respect to r, but of course we must do so at a constant angle θ, so it is more convenient if we first express ϕ in terms of r and θ, rather than of z and r. Since $z = r\cos\theta$, we get

$$\phi = -E_0 r\cos\theta + \frac{p\cos\theta}{4\pi\epsilon_0 r^2}. \tag{12.34}$$

The radial component of $\boldsymbol{E}$ is

$$-\frac{\partial\phi}{\partial r} = +E_0\cos\theta + \frac{p\cos\theta}{2\pi\epsilon_0 r^3}. \tag{12.35}$$

This must be zero at $r = a$ for all θ. This will be true if

$$p = -2\pi\epsilon_0 a^3 E_0. \tag{12.36}$$

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 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) fell to zero at large r. The dipole field is the only one that does all three.

Using (12.36), our potential is

$$\phi = -E_0\cos\theta\left(r + \frac{a^3}{2r^2}\right). \tag{12.37}$$

The solution of the fluid flow problem can be written simply as

$$\psi = -v_0\cos\theta\left(r + \frac{a^3}{2r^2}\right). \tag{12.38}$$

It is straightforward to find 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

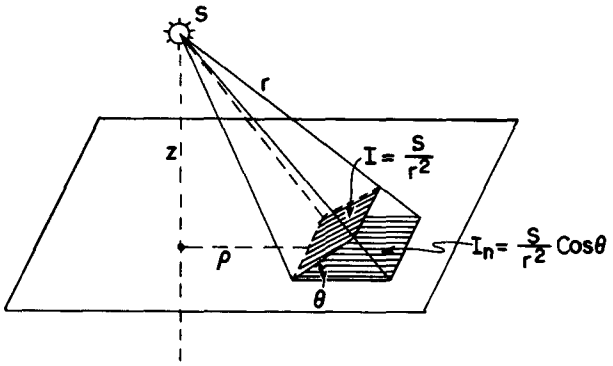


Fig. 12–9. The illumination I_n of a surface is the radiant energy per unit time arriving at a unit area of the surface.

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 flow 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 θ 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

$$I_n = \frac{S}{r^2}\, \mathbf{e}_r \cdot \mathbf{n}, \tag{12.39}$$

where $\mathbf{e}_r$ is the unit vector from the source and $\mathbf{n}$ 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 $4\pi\epsilon_0 S$. Knowing that, 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 z above the table. What is the widest spacing b from tube to tube that 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):

$$F_n = A_n e^{-2\pi n z/b}.$$

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 A_1, 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 $0.91z$. 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 arriving at the top of an opaque surface, so we must include in our integral only the sources which shine 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 $b = 0.8z$.) 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 electrostatics 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? That 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 electrostatics are the simplest vector equations that one can get which involve only the spatial derivatives of quantities. Any other *simple* problem—or simplification of a complicated problem—must look like electrostatics. What is common to all our problems is that they involve *space* and that we have *imitated* 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 smoothed-out 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 look wrong. It is possible that there is some as yet undiscovered underlying "machinery," and that the details of an underlying complexity are hidden in the smooth-looking equations—as is so

in the "smooth" diffusion of neutrons. 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 invention of an equation 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 some other such idiocy. No one has yet made up a theory of electricity for which $\nabla^2\phi = -\rho/\epsilon_0$ 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 $\nabla^2\phi = -\rho/\epsilon_0$ 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 no one yet knows an escape.

13

Magnetostatics

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 currents and charges

13–8 Superposition; the right-hand rule

Review: Chapter 15, Vol. I: *The Special Theory of Relativity*

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, $\boldsymbol{E}$. 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 $\boldsymbol{B}$, which specifies both the unique direction in space and the constant of proportionality with the velocity, and to write the magnetic force as $q\boldsymbol{v} \times \boldsymbol{B}$. The total electromagnetic force on a charge can, then, be written as

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{13.1}$$

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 $\boldsymbol{B}$ is evidently one newton·second per coulomb·meter. The same unit is also one volt·second per meter^2. 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 electrons or other 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 $\boldsymbol{j}$. 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

$$\boldsymbol{j} \cdot \boldsymbol{n}\, \Delta S, \tag{13.2}$$

where $\boldsymbol{n}$ 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 $\boldsymbol{v}$. 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 $\boldsymbol{v}$ times

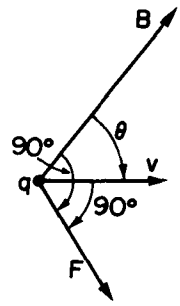


Fig. 13–1. The velocity-dependent component of the force on a moving charge is at right angles to **v** and to the direction of **B**. It is also proportional to the component of **v** at right angles to **B**, that is, to $v \sin \theta$.

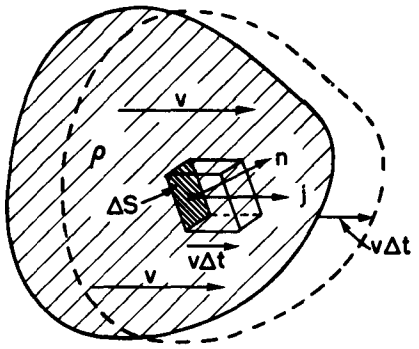


Fig. 13–2. If a charge distribution of density ρ moves with the velocity **v**, the charge per unit time through ΔS is $\rho \boldsymbol{v} \cdot \boldsymbol{n}\, \Delta S$.

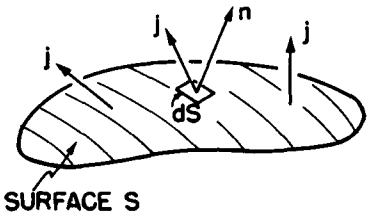


Fig. 13–3. The current I through the surface S is $\int \boldsymbol{j} \cdot \boldsymbol{n}\, dS$.

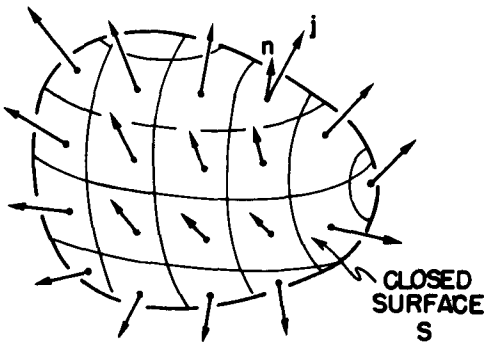


Fig. 13–4. The integral of $\boldsymbol{j} \cdot \boldsymbol{n}$ over a closed surface is the rate of change of the total charge Q inside.

$v\,\Delta t$, which when multiplied by the charge density ρ will give Δq. Thus

$$\Delta q = \rho \boldsymbol{v} \cdot \boldsymbol{n}\, \Delta S\, \Delta t.$$

The charge per unit time is then $\rho \boldsymbol{v} \cdot \boldsymbol{n}\, \Delta S$, from which we get

$$\boldsymbol{j} = \rho \boldsymbol{v}. \tag{13.3}$$

If the charge distribution consists of individual charges, say electrons, each with the charge q and moving with the mean velocity $\boldsymbol{v}$, then the current density is

$$\boldsymbol{j} = Nq\boldsymbol{v}, \tag{13.4}$$

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:

$$I = \int_S \boldsymbol{j} \cdot \boldsymbol{n}\, dS \tag{13.5}$$

(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

$$\int_{\substack{\text{any closed}\\ \text{surface}}} \boldsymbol{j} \cdot \boldsymbol{n}\, dS = -\frac{d}{dt}\,(Q_{\text{inside}}). \tag{13.6}$$

The charge inside can be written as a volume integral of the charge density:

$$Q_{\text{inside}} = \int_{\substack{V\\ \text{inside } S}} \rho\, dV. \tag{13.7}$$

If we apply (13.6) to a small volume ΔV, we know that the left-hand integral is $\nabla \cdot \boldsymbol{j}\, \Delta V$. The charge inside is $\rho\, \Delta V$, so the conservation of charge can also be written as

$$\nabla \cdot \boldsymbol{j} = -\frac{\partial \rho}{\partial t} \tag{13.8}$$

(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 $\boldsymbol{v}$ along the wire. Each charge feels a transverse force

$$\boldsymbol{F} = q\boldsymbol{v} \times \boldsymbol{B}$$

(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 $\Delta \boldsymbol{F}$ on the volume ΔV is the sum of the forces on the individual charges, that is,

$$\Delta \boldsymbol{F} = (N\,\Delta V)(q\boldsymbol{v} \times \boldsymbol{B}).$$

But $Nq\boldsymbol{v}$ is just $\boldsymbol{j}$, so

$$\Delta \boldsymbol{F} = \boldsymbol{j} \times \boldsymbol{B}\, \Delta V \tag{13.9}$$

(Fig. 13–5b). The force per unit volume is $\boldsymbol{j} \times \boldsymbol{B}$.

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

$$\Delta \boldsymbol{F} = j \times \boldsymbol{B} A \, \Delta L. \tag{13.10}$$

Now we can call jA the vector current $\boldsymbol{I}$ in the wire. (Its magnitude is the electric current in the wire, and its direction is along the wire.) Then

$$\Delta \boldsymbol{F} = \boldsymbol{I} \times \boldsymbol{B} \, \Delta L. \tag{13.11}$$

The force per unit length on a wire is $\boldsymbol{I} \times \boldsymbol{B}$.

This equation gives the important result that the magnetic force on a wire, due to the movement of charges 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).

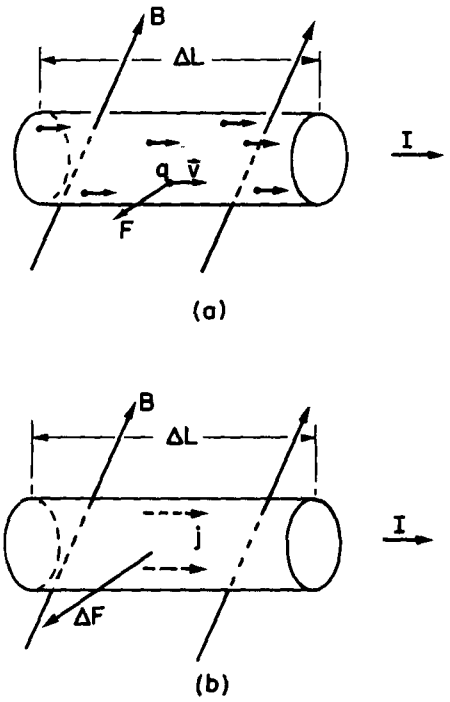


Fig. 13–5. The magnetic force on a current-carrying wire is the sum of the forces on the individual moving charges.

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.* There are indeed such 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 means 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 theoretical 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:*

$$\nabla \cdot \boldsymbol{B} = 0 \tag{13.12}$$

and

$$c^2 \nabla \times \boldsymbol{B} = \frac{j}{\epsilon_0}. \tag{13.13}$$

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 $\boldsymbol{j}$ 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 $\partial \boldsymbol{E}/\partial t$ and $\partial \boldsymbol{B}/\partial t$ from the complete Maxwell equations, Eqs. (2.41), 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 $\nabla \cdot \boldsymbol{j} = 0$. This is true, by Eq. (13.8), only if $\partial \rho/\partial t$ is zero. But that must be so if $\boldsymbol{E}$ 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!

The requirement that $\nabla \cdot j = 0$ means that we may only have charges which flow in paths that close back on themselves. They may, 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 condensers which are charging or discharging. (We will, of course, extend the theory later to include dynamic fields, but we want first to take the simpler case of steady currents.)

Now let us look at Eqs. (13.12) and (13.13) to see what they mean. The first one says that the divergence of $\boldsymbol{B}$ is zero. Comparing it to the analogous equation in electrostatics, which says that $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$, we can conclude that there is no magnetic analog of an electric charge. There are *no magnetic charges* from which lines of $\boldsymbol{B}$ can emerge. If we think in terms of "lines" of the vector field $\boldsymbol{B}$, 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 making loops around the currents. Since lines of $\boldsymbol{B}$ 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 $\nabla \cdot \boldsymbol{B} = 0$. This much is true not only for magnetostatics, it is *always* true—even for dynamic fields.

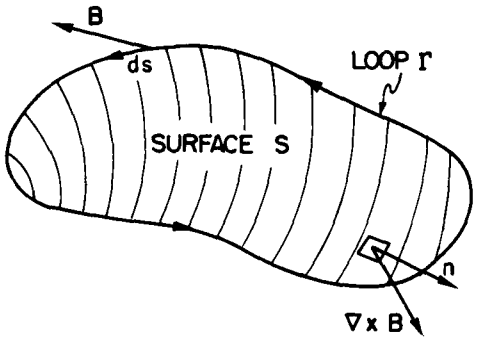


Fig. 13–6. The line integral of the tangential component of $\boldsymbol{B}$ is equal to the surface integral of the normal component of $\nabla \times \boldsymbol{B}$.

The connection between the $\boldsymbol{B}$ field and currents is contained in Eq. (13.13). Here we have a new kind of situation which is quite different from electrostatics, where we had $\nabla \times \boldsymbol{E} = 0$. That equation meant that the line integral of $\boldsymbol{E}$ around any closed path is zero:

$$\oint_{\text{loop}} \boldsymbol{E} \cdot d\boldsymbol{s} = 0.$$

We got that result from Stokes' theorem, which says that the integral around any closed path of *any* vector field is equal to the surface integral of the normal component of the curl 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

$$\oint_\Gamma \boldsymbol{B} \cdot d\boldsymbol{s} = \int_S (\nabla \times \boldsymbol{B}) \cdot \boldsymbol{n}\, dS. \tag{13.14}$$

Taking the curl of $\boldsymbol{B}$ from Eq. (13.13), we have

$$\oint_\Gamma \boldsymbol{B} \cdot d\boldsymbol{s} = \frac{1}{\epsilon_0 c^2} \int_S \boldsymbol{j} \cdot \boldsymbol{n}\, dS. \tag{13.15}$$

The integral over $\boldsymbol{j}$, 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 Γ, one usually speaks of "the current through the loop Γ." We have, then, a general law: the circulation of $\boldsymbol{B}$ around any closed curve is equal to the current I through the loop, divided by $\epsilon_0 c^2$:

$$\oint_\Gamma \boldsymbol{B} \cdot d\boldsymbol{s} = \frac{I_{\text{through }\Gamma}}{\epsilon_0 c^2}. \tag{13.16}$$

This law—called *Ampere's law*—plays the same role in magnetostatics that Gauss' law played in electrostatics. Ampere's law alone does not determine $\boldsymbol{B}$ from currents; we must, in general, also use $\nabla \cdot \boldsymbol{B} = 0$. 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.

13–5 The magnetic field of a straight wire and of a solenoid; atomic 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 $\boldsymbol{B}$ 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, $\boldsymbol{B}$ 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 $\boldsymbol{B} \cdot d\boldsymbol{s}$ quite easily; it is just the magnitude of $\boldsymbol{B}$ times the circumference. If r is the radius of the circle, then

$$\oint \boldsymbol{B} \cdot d\boldsymbol{s} = B \cdot 2\pi r.$$

The total current through the loop is merely the current I in the wire, so

$$B \cdot 2\pi r = \frac{I}{\epsilon_0 c^2},$$

or

$$B = \frac{1}{4\pi\epsilon_0 c^2}\frac{2I}{r}. \tag{13.17}$$

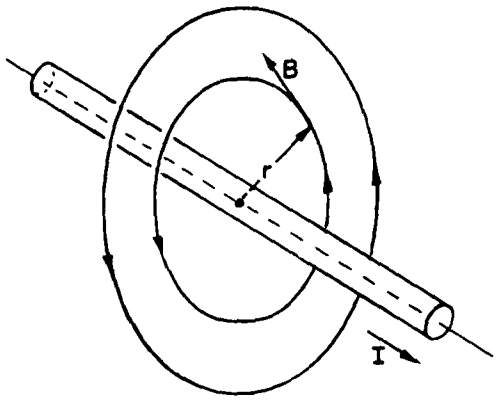


Fig. 13–7. The magnetic field outside of a long wire carrying the current I.

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 $\boldsymbol{B}$ is at right angles both to $\boldsymbol{I}$ and to $\boldsymbol{r}$, we have

$$\boldsymbol{B} = \frac{1}{4\pi\epsilon_0 c^2}\frac{2\boldsymbol{I} \times \boldsymbol{e}_r}{r}. \tag{13.18}$$

We have separated out the factor $1/4\pi\epsilon_0 c^2$, 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 $\boldsymbol{B}$ 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.

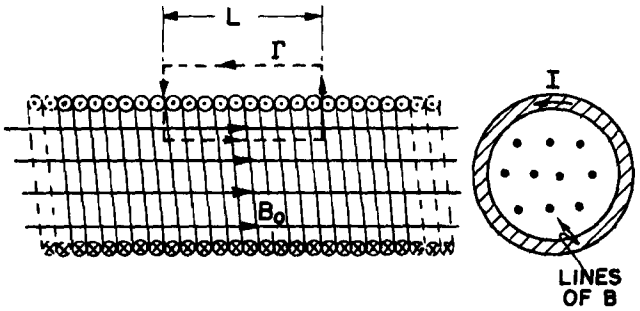


Fig. 13–8. The magnetic field of a long solenoid.

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 *stays* inside (and has 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" Γ shown in the figure. This loop goes the distance

L inside the solenoid, where the field is, say, $\boldsymbol{B}_0$, then goes at right angles to the field, and returns along the outside, where the field is negligible. The line integral of $\boldsymbol{B}$ for this curve is just B_0L, and it must be $1/\epsilon_0c^2$ times the total current through Γ, 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

$$B_0 = \frac{nI}{\epsilon_0c^2}. \tag{13.19}$$

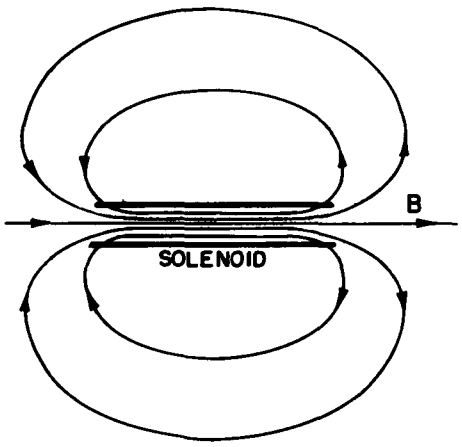


Fig. 13–9. The magnetic field outside of a solenoid.

What happens to the lines of $\boldsymbol{B}$ 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 $\boldsymbol{B}$ 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 magnetic iron" 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 $\boldsymbol{j}$ term.

Matter is very complex when looked at from a fundamental point of view—as we saw when we tried to understand dielectrics. 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 internal currents. In the case of iron, these currents come from 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 these spin and point every which way, so that there is no net effect. The miracle is that in a very few substances, like iron, a large fraction of the electrons spin with their axes in the same direction—for iron, two electrons from each atom takes 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 $\boldsymbol{B}$ given at the beginning of this chapter that what this vector is will depend on what we choose as a reference frame for our specification of the velocity of charges. But we have said nothing about which is the proper frame for specifying the magnetic field.

It turns out that *any* 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

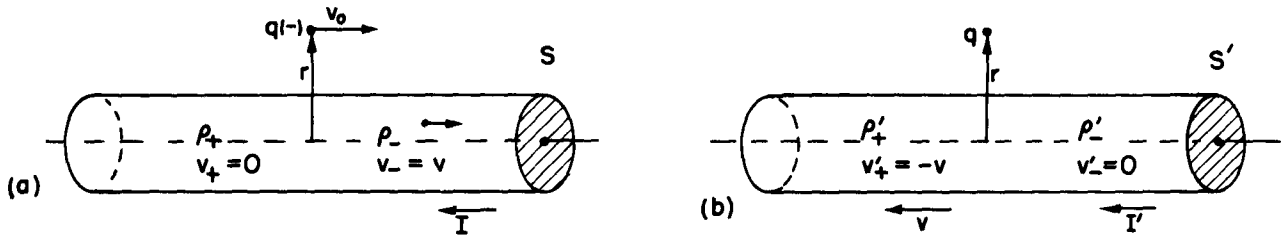


Fig. 13–10. The interaction of a current-carrying wire and a particle with the charge q as seen in two frames. In frame S (part a), the wire is at rest; in frame S' (part b), the charge is at rest.

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 can be no magnetic force 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_+, 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

$$\boldsymbol{F} = q\boldsymbol{v}_0 \times \boldsymbol{B}.$$

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

$$F = \frac{1}{4\pi\epsilon_0 c^2} \cdot \frac{2Iqv_0}{r}.$$

Using Eqs. (13.4) and (13.5), the current I can be written as $\rho_- vA$, where A is the area of a cross section of the wire. Then

$$F = \frac{1}{4\pi\epsilon_0 c^2} \cdot \frac{2q\rho_- Avv_0}{r}. \tag{13.20}$$

We could continue to treat the general case of arbitrary velocities for v and v_0, 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

$$F = \frac{q}{2\pi\epsilon_0} \frac{\rho_- A}{r} \frac{v^2}{c^2}. \tag{13.21}$$

Now we turn our attention to what happens in S', in which the particle is at rest and the wire is running 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 when 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 $1/\sqrt{1 - v^2/c^2}$. 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 matter 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

$$L = L_0\sqrt{1 - v^2/c^2}, \tag{13.22}$$

but with the same area A_0 (since 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 $\rho_0 L_0 A$, because charge is the same in any system, so that $\rho L = \rho_0 L_0$ or, from (13.22),

$$\rho = \frac{\rho_0}{\sqrt{1 - v^2/c^2}}. \tag{13.23}$$

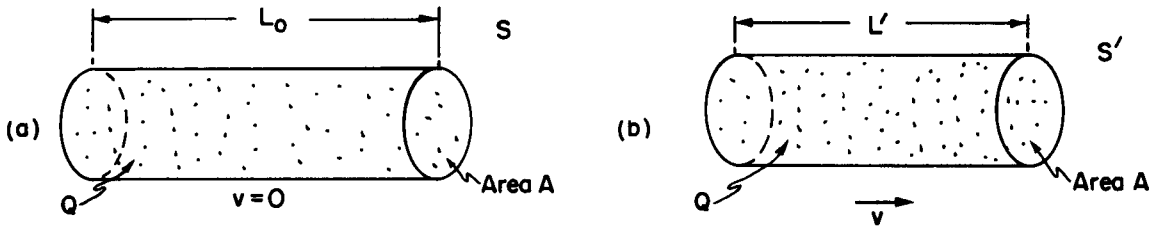


Fig. 13–11. If a distribution of charged particles at rest has the charge density ρ_0, the same charges will have the density $\rho = \rho_0/\sqrt{1 - v^2/c^2}$ when seen from a frame with the relative velocity v.

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

$$\rho'_+ = \frac{\rho_+}{\sqrt{1 - v^2/c^2}}. \tag{13.24}$$

The *negative* charges are at rest in S'. So they have their "rest density" ρ_0 in this frame. In Eq. (13.23) $\rho_0 = \rho'_-$, because they have the density ρ'_- when the *wire* is at rest, i.e., in frame S, where the speed of the negative charges is v. For the conduction electrons, we then have that

$$\rho_- = \frac{\rho'_-}{\sqrt{1 - v^2/c^2}}, \tag{13.25}$$

or

$$\rho'_- = \rho_-\sqrt{1 - v^2/c^2}. \tag{13.26}$$

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

$$\rho' = \rho_+ \frac{v^2/c^2}{\sqrt{1 - v^2/c^2}}. \tag{13.27}$$

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

$$E' = \frac{\rho' A}{2\pi\epsilon_0 r} = \frac{\rho_+ A v^2/c^2}{2\pi\epsilon_0 r\sqrt{1 - v^2/c^2}}. \tag{13.28}$$

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

$$F' = \frac{q}{2\pi\epsilon_0}\frac{\rho_+ A}{r}\frac{v^2/c^2}{\sqrt{1 - v^2/c^2}}. \tag{13.29}$$

Comparing this result for F' with our result for F in Eq. (13.21), we see that the magnitudes of the forces are almost identical from the two points of view. In fact,

$$F' = \frac{F}{\sqrt{1 - v^2/c^2}}, \tag{13.30}$$

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 do indeed 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 y, we want to compare Δp_y and $\Delta p'_y$. Using the relativistically correct equation of motion, $\boldsymbol{F} = d\boldsymbol{p}/dt$, we expect that after the time Δt our particle will have a transverse momentum Δp_y in the S-system given by

$$\Delta p_y = F\,\Delta t. \tag{13.31}$$

In the S'-system, the transverse momentum will be

$$\Delta p'_y = F'\,\Delta t'. \tag{13.32}$$

We must, of course, compare Δp_y and $\Delta p'_y$ for corresponding time intervals Δt and $\Delta t'$. We have seen in Chapter 15 of Vol. I that the time 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

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - v^2/c^2}}, \tag{13.33}$$

and everything comes out O.K. From (13.31) and (13.32),

$$\frac{\Delta p'_y}{\Delta p_y} = \frac{F'\,\Delta t'}{F\,\Delta t},$$

which is just $= 1$ if we combine (13.30) and (13.33).

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).

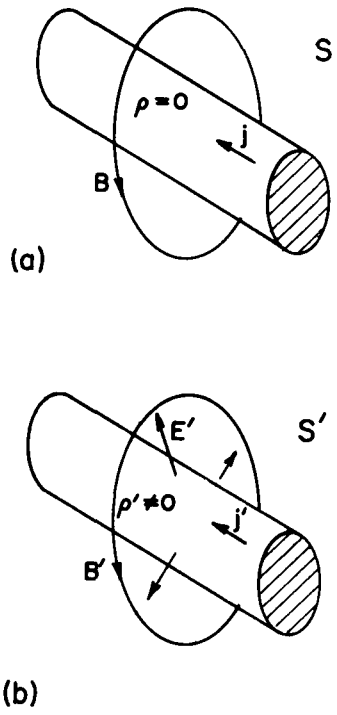


Fig. 13–12. In frame S the charge density is zero and the current density is j. There is only a magnetic field. In S', there is a charge density ρ', and a different current density j'. The magnetic field B' is different and there is an electric field E'.

If we had chosen still another coordinate system, we would have found a different mixture of $\boldsymbol{E}$ and $\boldsymbol{B}$ fields. Electric and magnetic forces are part of *one* physical 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 Einstein'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 $\boldsymbol{E}$ and $\boldsymbol{B}$. For instance, if we think of "lines" of $\boldsymbol{E}$ or $\boldsymbol{B}$, we must not attach 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 $\boldsymbol{B}$ 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 point in space. In particular, $\boldsymbol{E}$ and $\boldsymbol{B}$ 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 $\boldsymbol{E}$ and $\boldsymbol{B}$ at the charge, and the formula (13.1) is not to be altered if the *source* of $\boldsymbol{E}$ or $\boldsymbol{B}$ is moving (it is the values of $\boldsymbol{E}$ and $\boldsymbol{B}$ 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 later 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 place and later at another. Similarly, in an electromagnetic wave, the *wave* travels, but the magnitude of the fields *change*. So in the future when we—or someone else—speaks 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 $\boldsymbol{v}$, the density is

$$\rho = \frac{\rho_0}{\sqrt{1 - v^2/c^2}}.$$

In that frame their current density is

$$\boldsymbol{j} = \rho\boldsymbol{v} = \frac{\rho_0\boldsymbol{v}}{\sqrt{1 - v^2/c^2}}. \tag{13.34}$$

Now we know that the energy U and momentum p of a particle moving with velocity $\boldsymbol{v}$ are given by

$$U = \frac{m_0c^2}{\sqrt{1 - v^2/c^2}}, \qquad \boldsymbol{p} = \frac{m_0\boldsymbol{v}}{\sqrt{1 - v^2/c^2}},$$

where m_0 is its rest mass. We also know that U and $\boldsymbol{p}$ form a relativistic four-vector. Since ρ and $\boldsymbol{j}$ depend on the velocity $\boldsymbol{v}$ exactly as do U and $\boldsymbol{p}$, we can conclude that ρ and $\boldsymbol{j}$ 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 $\boldsymbol{v}_0$ of the particle different from the velocity of the conduction electrons.

If we wish to transform ρ and $\boldsymbol{j}$ to a coordinate system moving with a velocity $\boldsymbol{u}$ in the x-direction, we know that they transform just like t and (x, y, z), so that we have (see Chapter 15, Vol. I)

$$\begin{aligned}
x' &= \frac{x - ut}{\sqrt{1 - u^2/c^2}}, & j'_x &= \frac{j_x - u\rho}{\sqrt{1 - u^2/c^2}},\\
y' &= y, & j'_y &= j_y,\\
z' &= z, & j'_z &= j_z,\\
t' &= \frac{t - ux/c^2}{\sqrt{1 - u^2/c^2}}, & \rho' &= \frac{\rho - uj_x/c^2}{\sqrt{1 - u^2/c^2}}.
\end{aligned} \tag{13.35}$$

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 motions 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,

$$\nabla \cdot \boldsymbol{B} = 0, \qquad \nabla \times \boldsymbol{B} = \boldsymbol{j}/c^2\epsilon_0,$$

are linear in $\boldsymbol{B}$ and $\boldsymbol{j}$. That 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-hand 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 $\boldsymbol{B}$ 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 $\boldsymbol{E}$, 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 52, Vol. I). 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, or that currents in opposite directions repel. (Try working out the force using "left-hand rules.") An attraction or repulsion is a polar vector. This happens because in describing any complete interaction, we use the right-hand rule twice—once to find $\boldsymbol{B}$ from currents, again to find the force this $\boldsymbol{B}$ 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 conventions to a left-hand system all our $\boldsymbol{B}$ fields would be reversed, but all forces—or, what is perhaps more relevant, the observed accelerations of objects—would be unchanged.

Although physicists have recently found to their surprise that *all* the laws of nature are not always invariant for mirror reflections, the laws of electromagnetism do have such a basic symmetry.

14

The Magnetic Field in Various Situations

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

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

$$\nabla \cdot \boldsymbol{B} = 0, \tag{14.1}$$

$$c^2 \nabla \times \boldsymbol{B} = \frac{\boldsymbol{j}}{\epsilon_0}. \tag{14.2}$$

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. (4.25). 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 $\boldsymbol{B}$ if we know the current density $\boldsymbol{j}$ of all moving charges.

In electrostatics we saw that (because the curl of $\boldsymbol{E}$ was always zero) it was possible to represent $\boldsymbol{E}$ as the gradient of a scalar field ϕ. Now the curl of $\boldsymbol{B}$ is *not* always zero, so it is not possible, in general, to represent it as a gradient. However, the *divergence* of $\boldsymbol{B}$ *is* always zero, and this means that we can always represent $\boldsymbol{B}$ 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 $\boldsymbol{B}$ to a field we will call $\boldsymbol{A}$ by

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}. \tag{14.3}$$

Or, by writing out the components,

$$\begin{aligned} B_x &= (\nabla \times \boldsymbol{A})_x = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \\ B_y &= (\nabla \times \boldsymbol{A})_y = \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \\ B_z &= (\nabla \times \boldsymbol{A})_z = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}. \end{aligned} \tag{14.4}$$

Writing $\boldsymbol{B} = \nabla \times \boldsymbol{A}$ guarantees that Eq. (14.1) is satisfied, since, necessarily,

$$\nabla \cdot \boldsymbol{B} = \nabla \cdot (\nabla \times \boldsymbol{A}) = 0.$$

The field $\boldsymbol{A}$ 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 have different vector potentials $\boldsymbol{A}$ which give the same magnetic fields. Again, because $\boldsymbol{B}$ is obtained from $\boldsymbol{A}$ by differentiation, adding a

constant to A doesn't 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 an A that gives correctly the magnetic field B for some real situation, 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:

$$B = \nabla \times A' = \nabla \times A.$$

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,

$$A' = A + \nabla\psi \tag{14.5}$$

will be an equally satisfactory vector potential, leading to the same field B.

It is usually convenient to take some of the "latitude" out of A by arbitrarily placing some other condition on it (in much the same way that we found it convenient—often—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

$$\nabla \cdot A = 0. \tag{14.6}$$

(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 in the direction of B_0, we must have

$$\begin{aligned} B_x &= \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} = 0, \\ B_y &= \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} = 0, \\ B_z &= \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} = B_0. \end{aligned} \tag{14.7}$$

By inspection, we see that one *possible* solution of these equations is

$$A_y = xB_0, \qquad A_x = 0, \qquad A_z = 0.$$

Or we could equally well take

$$A_x = -yB_0, \qquad A_y = 0, \qquad A_z = 0.$$

Still another solution is a linear combination of the two:

$$A_x = -\tfrac{1}{2}yB_0, \qquad A_y = \tfrac{1}{2}xB_0, \qquad A_z = 0. \tag{14.8}$$

---

* 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 $\boldsymbol{B}$, the vector potential $\boldsymbol{A}$ 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 $+x$, $\boldsymbol{A}$ must be at right angles to the vector from the z-axis, which we will call $\boldsymbol{r}'$ (the "prime" is to remind us that it is *not* the vector displacement from the origin). Also, the magnitude of $\boldsymbol{A}$ is proportional to $\sqrt{x^2 + y^2}$ and, hence, to r'. So $\boldsymbol{A}$ can be simply written (for our uniform field) as

$$\boldsymbol{A} = \tfrac{1}{2}\boldsymbol{B} \times \boldsymbol{r}'. \tag{14.9}$$

The vector potential $\boldsymbol{A}$ has the magnitude $Br'/2$ and rotates about the z-axis as shown in Fig. 14–1. If, for example, the $\boldsymbol{B}$ 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 $\boldsymbol{A}$ on any closed loop Γ can be related to the surface integral of $\nabla \times \boldsymbol{A}$ by Stokes' theorem, Eq. (3.38):

$$\oint_\Gamma \boldsymbol{A} \cdot d\boldsymbol{s} = \int_{\text{inside } \Gamma} (\nabla \times \boldsymbol{A}) \cdot \boldsymbol{n}\, da. \tag{14.10}$$

But the integral on the right is equal to the flux of $\boldsymbol{B}$ through the loop, so

$$\oint_\Gamma \boldsymbol{A} \cdot d\boldsymbol{s} = \int_{\text{inside } \Gamma} \boldsymbol{B} \cdot \boldsymbol{n}\, da. \tag{14.11}$$

So the circulation of $\boldsymbol{A}$ around *any* loop is equal to the flux of $\boldsymbol{B}$ through the loop. If we take a circular loop, of radius r' in a plane perpendicular to a uniform field $\boldsymbol{B}$, the flux is just

$$\pi r'^2 B.$$

If we choose our origin on an axis of symmetry, so that we can take $\boldsymbol{A}$ as circumferential and a function only of r', the circulation will be

$$\oint \boldsymbol{A} \cdot d\boldsymbol{s} = 2\pi r' A = \pi r'^2 B.$$

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.

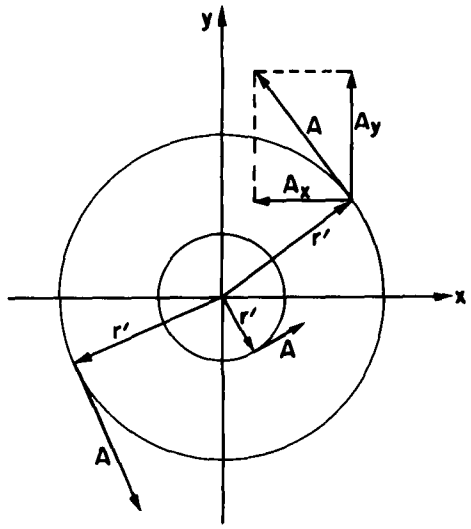


Fig. 14–1. A uniform magnetic field $\boldsymbol{B}$ in the z-direction corresponds to a vector potential $\boldsymbol{A}$ that rotates about the z-axis, with the magnitude $A = Br'/2$ (r' is the displacement from the z-axis).

14–2 The vector potential of known currents

Since $\boldsymbol{B}$ is determined by currents, so also is $\boldsymbol{A}$. We want now to find $\boldsymbol{A}$ in terms of the currents. We start with our basic equation (14.2):

$$c^2 \nabla \times \boldsymbol{B} = \frac{\boldsymbol{j}}{\epsilon_0},$$

which means, of course, that

$$c^2 \nabla \times (\nabla \times \boldsymbol{A}) = \frac{\boldsymbol{j}}{\epsilon_0}. \tag{14.12}$$

This equation is for magnetostatics what the equation

$$\nabla \cdot \nabla \phi = -\frac{\rho}{\epsilon_0} \tag{14.13}$$

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):

$$\nabla \times (\nabla \times A) = \nabla(\nabla \cdot A) - \nabla^2 A. \tag{14.14}$$

Since we have chosen to make $\nabla \cdot A = 0$ (and now you see why), Eq. (14.12) becomes

$$\nabla^2 A = -\frac{j}{\epsilon_0 c^2}. \tag{14.15}$$

This vector equation means, of course, three equations:

$$\nabla^2 A_x = -\frac{j_x}{\epsilon_0 c^2}, \qquad \nabla^2 A_y = -\frac{j_y}{\epsilon_0 c^2}, \qquad \nabla^2 A_z = -\frac{j_z}{\epsilon_0 c^2}. \tag{14.16}$$

And each of these equations is *mathematically identical* to

$$\nabla^2 \phi = -\frac{\rho}{\epsilon_0}. \tag{14.17}$$

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!

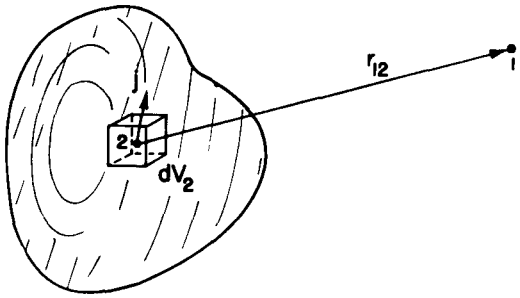


Fig. 14–2. The vector potential A at point 1 is given by an integral over the current elements $j\,dV$ at all points 2.

We have seen in Chapter 4 that a general solution for the electrostatic equation (14.17) is

$$\phi(1) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(2)\,dV_2}{r_{12}}.$$

So we know immediately that a general solution for A_x is

$$A_x(1) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{j_x(2)\,dV_2}{r_{12}}, \tag{14.18}$$

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

$$A(1) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{j(2)\,dV_2}{r_{12}}. \tag{14.19}$$

(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 of 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 $\rho_1 = j_x/c^2$, $\rho_2 = j_y/c^2$, and $\rho_3 = j_z/c^2$. 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

coordinates as shown in Fig. 14–3, the current density vector j has only a z-component. Its magnitude is

$$j_z = \frac{I}{\pi a^2} \tag{14.20}$$

inside the wire, and zero outside.

Since j_x and j_y are both zero, we have immediately

$$A_x = 0, \qquad A_y = 0.$$

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

$$\phi = -\frac{\lambda}{2\pi\epsilon_0} \ln r',$$

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 uniform current. Since $\pi a^2 j_z = I$, we can also write

$$A_z = -\frac{I}{2\pi\epsilon_0 c^2} \ln r'. \tag{14.21}$$

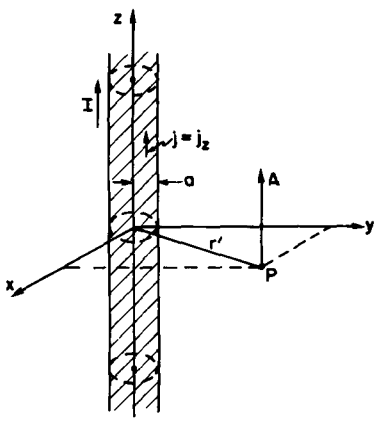


Fig. 14–3. A long cylindrical wire along the z-axis with a uniform current density j.

Now we can find $\boldsymbol{B}$ from (14.4). There are only two of the six derivatives that are not zero. We get

$$B_x = -\frac{I}{2\pi\epsilon_0 c^2} \frac{\partial}{\partial y} \ln r' = -\frac{I}{2\pi\epsilon_0 c^2} \frac{y}{r'^2}, \tag{14.22}$$

$$B_y = \frac{I}{2\pi\epsilon_0 c^2} \frac{\partial}{\partial x} \ln r' = \frac{I}{2\pi\epsilon_0 c^2} \frac{x}{r'^2}, \tag{14.23}$$

$$B_z = 0.$$

We get the same result as before: $\boldsymbol{B}$ circles around the wire, and has the magnitude

$$B = \frac{1}{4\pi\epsilon_0 c^2} \frac{2I}{r'}. \tag{14.24}$$

14–4 A long solenoid

Next, we consider again the infinitely long solenoid with a circumferential current on the surface of nI per unit length. (We imagine there are n turns of wire per unit length, 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" $\boldsymbol{J}$ 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 $\boldsymbol{J}$ is, here, nI. This surface current (see Fig. 14–4) has the components.

$$J_x = -J \sin\phi, \qquad J_y = J \cos\phi, \qquad J_z = 0.$$

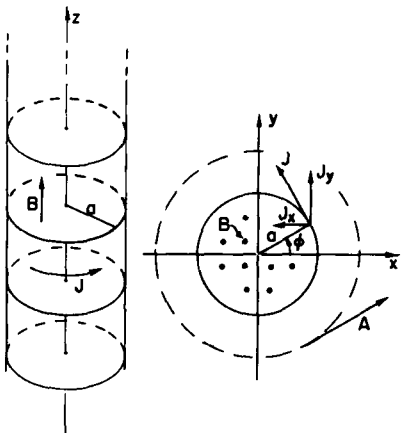


Fig. 14–4. A long solenoid with a surface current density J.

Now we must find A for such a current distribution.

First, we wish to find A_x for points outside the solenoid. The result is the same as the electrostatic potential outside a cylinder with a surface 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 *solid* cylinders of charge, one positive and one negative, with a slight relative displacement of their

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

$$\phi \propto \frac{\partial \ln r'}{\partial y} = \frac{y}{r'^2}.$$

So we know that

$$A_x = -K\frac{y}{r'^2}, \tag{14.25}$$

where K is some constant. Following the same argument, we would find

$$A_y = K\frac{x}{r'^2}. \tag{14.26}$$

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 its curl zero?

Clearly, B_x and B_y are zero, and

$$B_z = \frac{\partial}{\partial x}\left(K\frac{x}{r'^2}\right) - \frac{\partial}{\partial y}\left(-K\frac{y}{r'^2}\right)$$
$$= K\left(\frac{1}{r'^2} - \frac{2x^2}{r'^4} + \frac{1}{r'^2} - \frac{2y^2}{r'^4}\right) = 0.$$

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 B inside the coil (Eq. 14.11). The circulation is $A \cdot 2\pi r'$ or, since $A = K/r'$, the circulation is $2\pi K$. Notice that it is independent of r'. That is just as it should be if there is no B outside, because the flux is just the magnitude of 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 $nI/\epsilon_0 c^2$, so we can determine the constant K:

$$2\pi K = \pi a^2 \frac{nI}{\epsilon_0 c^2},$$

or

$$K = \frac{nIa^2}{2\epsilon_0 c^2}.$$

So the vector potential *outside* has the magnitude

$$A = \frac{nIa^2}{2\epsilon_0 c^2}\frac{1}{r'}, \tag{14.27}$$

and is always perpendicular to the vector r'.

We have been thinking of a solenoidal coil of wire, but we would produce the same fields if we rotated a long cylinder with an electrostatic charge on the 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 $v = a\omega$ is the velocity of the surface charge. There will then be a magnetic field $B = \sigma a\omega/\epsilon_0 c^2$ 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 $v \times B$ forces will cause the ends of the wire to be charged (they will charge up until the E-field from the charges just balances the $v \times B$ 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

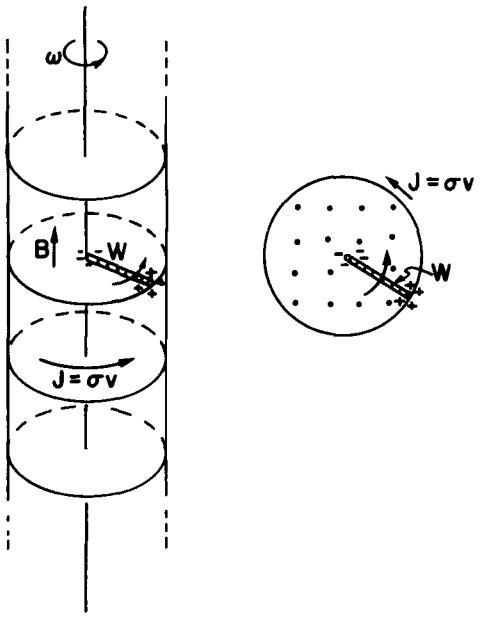


Fig. 14–5. A rotating charged cylinder produces a magnetic field inside. A short radial wire rotating with the cylinder has charges induced on its ends.

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.

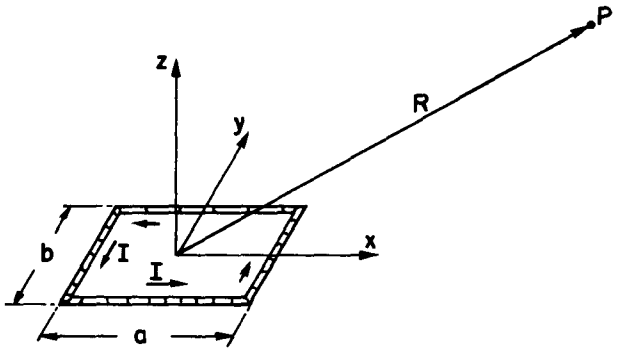


Fig. 14–6. A rectangular loop of wire with the current I. What is the magnetic field at P? ($R \gg a$, or b.)

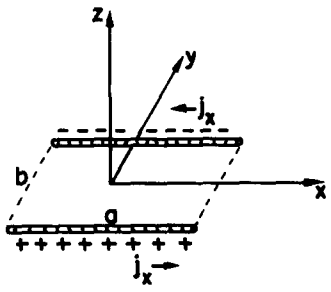


Fig. 14–7. The distribution of j_x in the current loop of Fig. 14–6.

We take first a rectangular loop, and choose our coordinates as shown in Fig. 14–6. There are no currents 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 (and current) is uniform. So the solution for A_x is just like the electrostatic potential from two charged rods (see Fig. 14–7). 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

$$\phi = \frac{1}{4\pi\epsilon_0}\frac{\boldsymbol{p}\cdot\boldsymbol{e}_R}{R^2}, \tag{14.28}$$

where $\boldsymbol{p}$ 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:

$$p = \lambda ab. \tag{14.29}$$

The dipole moment points in the negative y-direction, so the cosine of the angle between $\boldsymbol{R}$ and $\boldsymbol{p}$ is $-y/R$ (where y is the coordinate of P). So we have

$$\phi = -\frac{1}{4\pi\epsilon_0}\frac{\lambda ab}{R^2}\frac{y}{R}.$$

We get A_x simply by replacing λ by I/c^2:

$$A_x = -\frac{Iab}{4\pi\epsilon_0 c^2}\frac{y}{R^3}. \tag{14.30}$$

By the same reasoning,

$$A_y = \frac{Iab}{4\pi\epsilon_0 c^2}\frac{x}{R^3}. \tag{14.31}$$

Again, A_y is proportional to x and A_x 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:

$$\mu = Iab. \tag{14.32}$$

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

$$\mu = I \cdot (\text{area of loop}). \tag{14.33}$$

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

$$A = \frac{1}{4\pi\epsilon_0 c^2}\frac{\mu \times R}{R^3} = \frac{1}{4\pi\epsilon_0 c^2}\frac{\mu \times e_R}{R^2}. \tag{14.34}$$

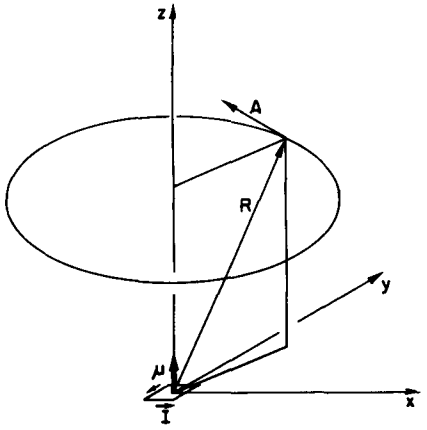


Fig. 14–8. The vector potential of a small current loop at the origin (in the xy-plane); a magnetic dipole field.

We have still to find $\boldsymbol{B}$. Using (14.33) and (14.34), together with (14.4), we get

$$B_x = -\frac{\partial}{\partial z}\frac{\mu}{4\pi\epsilon_0 c^2}\frac{x}{R^3} = \cdots\frac{3xz}{R^5} \tag{14.35}$$

(where by . . . we mean $\mu/4\pi\epsilon_0 c^2$),

$$\begin{aligned} B_y &= \frac{\partial}{\partial z}\left(-\cdots\frac{y}{R^3}\right) = \cdots\frac{3yz}{R^5}, \\ B_z &= \frac{\partial}{\partial x}\left(\cdots\frac{x}{R^3}\right) - \frac{\partial}{\partial y}\left(-\cdots\frac{y}{R^3}\right) \\ &= -\cdots\left(\frac{1}{r^3} - \frac{3z^2}{r^5}\right). \end{aligned} \tag{14.36}$$

The components of the $\boldsymbol{B}$-field behave exactly like those of the $\boldsymbol{E}$-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 $\nabla \times B = j/\epsilon_0 c^2$, 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 $\boldsymbol{E}$ and $\boldsymbol{B}$ 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 circuit

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 our volume element 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:

$$j\,dV = jS\,ds. \tag{14.37}$$

But jS is just what we call the current I in a wire, so our integral for the vector potential (14.19) becomes

$$A(1) = \frac{1}{4\pi\epsilon_0 c^2}\int \frac{I\,ds_2}{r_{12}} \tag{14.38}$$

(see Fig. 14–10). (We assume that I 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.

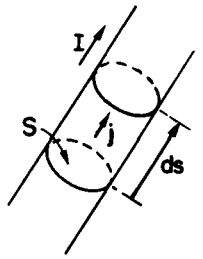


Fig. 14–9. For a fine wire $j\,dV$ is the same as $I\,ds$.

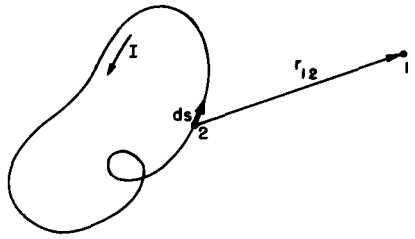


Fig. 14–10. The magnetic field of a wire can be obtained from an integral around the circuit.

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 an integral (Eq. 4–16):

$$E(1) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(2)e_{12}\,dV_2}{r_{12}^2}.$$

As we have seen, it is usually more work to evaluate this integral—there are really three integrals, one for each component—than 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. (14.19); we can get an integral for B by taking the curl of both sides:

$$B(1) = \nabla \times A(1) = \nabla \times \left[\frac{1}{4\pi\epsilon_0 c^2}\int \frac{j(2)\,dV_2}{r_{12}}\right]. \tag{14.39}$$

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, appear only in

$$r_{12} = [(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2]^{1/2}. \tag{14.40}$$

We have, for the x-component of B,

$$\begin{aligned} B_x &= \frac{\partial A_z}{\partial y_1} - \frac{\partial A_y}{\partial z_1} \\ &= \frac{1}{4\pi\epsilon_0 c^2}\int \left[j_z \frac{\partial}{\partial y_1}\left(\frac{1}{r_{12}}\right) - j_y \frac{\partial}{\partial z_1}\left(\frac{1}{r_{12}}\right)\right] dV_2 \\ &= -\frac{1}{4\pi\epsilon_0 c^2}\int \left[j_z \frac{y_1 - y_2}{r_{12}^3} - j_y \frac{z_1 - z_2}{r_{12}^3}\right] dV_2. \end{aligned} \tag{14.41}$$

The quantity in brackets is just the x-component of

$$\frac{j \times r_{12}}{r_{12}^3} = \frac{j \times e_{12}}{r_{12}^2}.$$

Corresponding results will be found for the other components, so we have

$$\boldsymbol{B}(1) = \frac{1}{4\pi\epsilon_0 c^2} \int \frac{\boldsymbol{j}(2) \times \boldsymbol{e}_{12}}{r_{12}^2}\, dV_2. \tag{14.42}$$

The integral gives $\boldsymbol{B}$ directly in terms of the known currents. The geometry involved is the same as that shown in Fig. 14–2.

If the currents exist only in circuits of small wires we can, as in the last section, immediately do the integral across the wire, replacing $\boldsymbol{j}\, dV$ by $I\, \boldsymbol{ds}$, where $\boldsymbol{ds}$ is an element of length of the wire. Then, using the symbols in Fig. 14–10,

$$\boldsymbol{B}(1) = -\frac{1}{4\pi\epsilon_0 c^2} \int \frac{I \boldsymbol{e}_{12} \times \boldsymbol{ds}_2}{r_{12}^2}. \tag{14.43}$$

(The minus sign appears because we have reversed the order of the cross product.) This equation for $\boldsymbol{B}$ 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 $\boldsymbol{B}$ directly with a vector integral? After all, $\boldsymbol{A}$ also involves three integrals!" Because of the cross product, the integrals for $\boldsymbol{B}$ are usually more complicated, as is evident from Eq. (14.41). Also, since the integrals for $\boldsymbol{A}$ are like those of electrostatics, we may already know them. Finally, we will see that in more advanced theoretical matters (in relativity, in advanced formulations 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.

15

The Vector Potential

15–1 The forces on a current loop; energy of a dipole

15–2 Mechanical and electrical energies

15–3 The energy of steady currents

15–4 B versus A

15–5 The vector potential and quantum mechanics

15–6 What is true for statics is false for dynamics

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

$$\mu = IA, \tag{15.1}$$

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

$$\boldsymbol{\mu} = IA\mathbf{n},$$

where $\mathbf{n}$ is the unit normal to the area A.

A current loop—or 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 z-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

$$F_1 = F_2 = IBb.$$

Their moment arm is

$$a \sin\theta,$$

so the torque is

$$\tau = Iab\,B \sin\theta,$$

or, since Iab is the magnetic moment of the loop,

$$\tau = \mu B \sin\theta.$$

The torque can be written in vector notation:

$$\boldsymbol{\tau} = \boldsymbol{\mu} \times \boldsymbol{B}. \tag{15.2}$$

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:

$$\boldsymbol{\tau} = \boldsymbol{p} \times \boldsymbol{E}.$$

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

$$dU = -\tau\, d\theta.$$

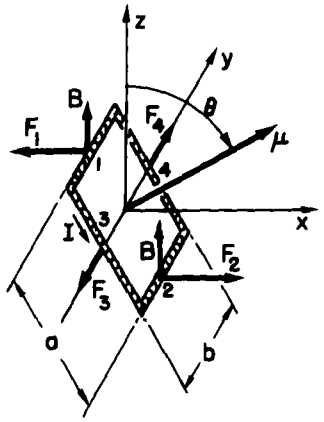


Fig. 15–1. A rectangular loop carrying the current I sits in a uniform field $\boldsymbol{B}$ (in the z-direction). The torque on the loop is $\boldsymbol{\tau} = \boldsymbol{\mu} \times \boldsymbol{B}$, where the magnetic moment $\mu = Iab$.

Setting $\tau = -\mu B \sin\theta$, and integrating, we can write for the energy

$$U = -\mu B \cos\theta + \text{a constant.} \tag{15.3}$$

(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_{mech}, 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:

$$U_{\text{mech}} = -\boldsymbol{\mu} \cdot \boldsymbol{B}. \tag{15.4}$$

Again, this corresponds to our result for an electric dipole:

$$U = -\boldsymbol{p} \cdot \boldsymbol{E}. \tag{15.5}$$

Now the electrostatic energy U in Eq. (15.5) is the true energy, but U_{mech} 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_{mech} 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 nonuniform 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.

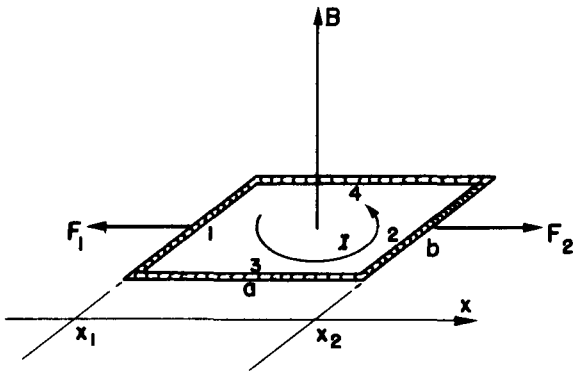


Fig. 15–2. A loop is carried along the x-direction through the field B, at right angles to x.

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 $IbB(x)$ 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:

$$W_2 = -\int_{-\infty}^{x_2} F_2\, dx = -Ib\int_{-\infty}^{x_2} B(x)\, dx. \tag{15.6}$$

Similarly, the work done against the forces on side 1 is

$$W_1 = -\int_{-\infty}^{x_1} F_1\, dx = Ib\int_{-\infty}^{x_1} B(x)\, dx. \tag{15.7}$$

To find each integral, we need to know how $B(x)$ depends on x. But 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

$$W = -Ib\int_{x_1}^{x_2} B(x)\,dx. \tag{15.8}$$

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 put in is

$$U_{\text{mech}} = W = -Iab\,B = -\mu B. \tag{15.9}$$

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 x-direction is

$$F_x = Ib(B_2 - B_1).$$

If the loop is "small," that is, if B_2 and B_1 are not too different, we can write

$$B_2 = B_1 + \frac{\partial B}{\partial x}\Delta x = B_1 + \frac{\partial B}{\partial x}a.$$

So the force is

$$F_x = Iab\,\frac{\partial B}{\partial x}\cdot \tag{15.10}$$

The total work done on the loop by *external* forces is

$$-\int_{-\infty}^{x} F_x\,dx = -Iab\int\frac{\partial B}{\partial x}\,dx = -IabB,$$

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

$$F_x\,\Delta x = -\Delta U_{\text{mech}} = -\Delta(-\boldsymbol{\mu}\cdot\boldsymbol{B}). \tag{15.11}$$

Our result, then, is that even though $U_{\text{mech}} = -\boldsymbol{\mu}\cdot\boldsymbol{B}$ 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 computing forces from the principle of virtual work, *provided* that the current in the loop (and all *other* currents) do not change. Let's see why all this works.

Imagine that the loop in Fig. 15–2 is moving in the $+x$-direction and take the z-axis in the direction of $\boldsymbol{B}$. The conduction electrons in 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, having work done on it 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 nonuniform 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 U_{mech} 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 goes 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 $\boldsymbol{B}$ 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 $q_e v_{\text{wire}} B$. So the rate at which electrical work is being done is $Fv_{\text{drift}} = (q_e v_{\text{wire}} B) v_{\text{drift}}$. If there are N conduction electrons in the unit length of the wire, the total rate at which electrical work is being done is

$$\frac{dU_{\text{elect}}}{dt} = N q_e v_{\text{wire}} B v_{\text{drift}}.$$

But $N q_e v_{\text{drift}} = I$, the current in the wire, so

$$\frac{dU_{\text{elect}}}{dt} = I v_{\text{wire}} B.$$

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 into 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 on 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

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}).$$

The rate at which work is done is

$$\boldsymbol{v} \cdot \boldsymbol{F} = q[\boldsymbol{v} \cdot \boldsymbol{E} + \boldsymbol{v} \cdot (\boldsymbol{v} \times \boldsymbol{B})]. \tag{15.12}$$

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 $\boldsymbol{B}_1$ produced by the current I_2 in a coil. Now the current I_1 in the loop will also be producing some magnetic field $\boldsymbol{B}_2$ at the coil. If the loop is moving, the field $\boldsymbol{B}_2$ will be changing. As we shall see in the next chapter, a changing magnetic field generates an $\boldsymbol{E}$-field; and this $\boldsymbol{E}$-field will do work on the charges in the coil. This energy must also be included in our balance sheet of the total energy.

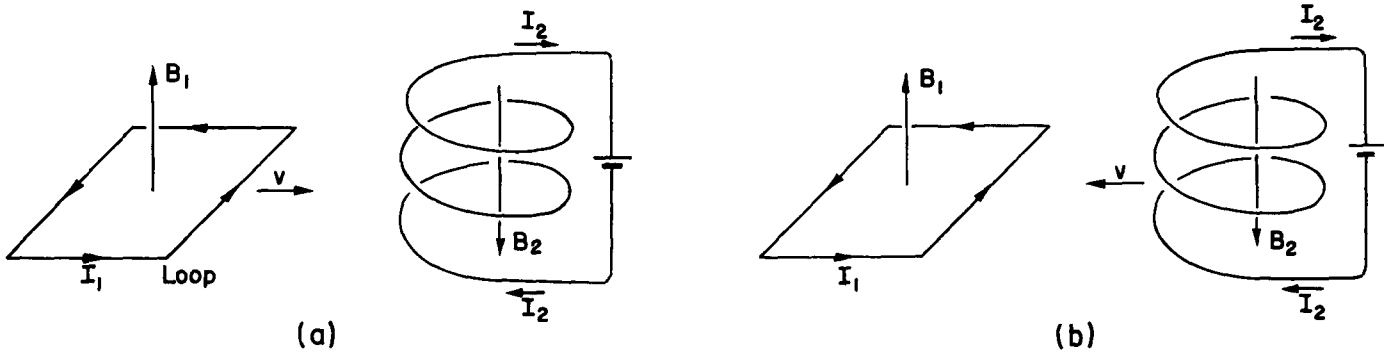


Fig. 15–3. Finding the energy of a small loop in a magnetic field.

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

$$U_{\text{mech}} + U_{\text{elect}}(\text{loop}) = 0.$$

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

$$U_{\text{mech}} + U_{\text{elect}}(\text{coil}) = 0.$$

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

$$2U_{\text{mech}} + U_{\text{elect}}(\text{loop}) + U_{\text{elect}}(\text{coil}) = 0.$$

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

$$U_{\text{total}} = U_{\text{elect}}(\text{loop}) + U_{\text{elect}}(\text{coil}) + U_{\text{mech}} = -U_{\text{mech}}. \qquad (15.13)$$

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

$$U_{\text{total}} = +\boldsymbol{\mu} \cdot \boldsymbol{B}.$$

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 have seen an analogous situation in electrostatics. We showed 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,

$$\Delta U = \frac{Q^2}{2}\Delta\left(\frac{1}{C}\right) = -\frac{Q^2}{2}\frac{\Delta C}{C^2}. \qquad (15.14)$$

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 force 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

$$\Delta U_{\text{mech}} = \Delta\left(-\frac{CV^2}{2}\right) = -\frac{V^2}{2}\Delta C, \tag{15.15}$$

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 voltages 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 $U_{\text{total}} = -U_{\text{mech}}$ 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_{total} just U, we write

$$U = \boldsymbol{\mu} \cdot \boldsymbol{B}. \tag{15.16}$$

Although we calculated this energy for a plane rectangular loop, the same result holds for a small plane loop of any shape.

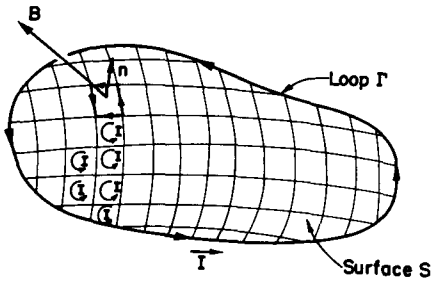


Fig. 15–4. The energy of a large loop in a magnetic field can be considered as the sum of energies of smaller loops.

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 fill 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 let the current I circulate around *each* of the little loops, the net result will be the same as a current around Γ, since the currents will cancel on all lines internal to Γ. Physically, the system of little currents is indistinguishable from the original circuit. The energy must 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 $I\,\Delta a B_n$, where B_n is the component normal to Δa. The total energy is

$$U = \sum IB_n\,\Delta a.$$

Going to the limit of infinitesimal loops, the sum becomes an integral, and

$$U = I\int B_n\,da = I\int \boldsymbol{B}\cdot\boldsymbol{n}\,da, \tag{15.17}$$

where $\boldsymbol{n}$ is the unit normal to da.

If we set $\boldsymbol{B} = \nabla \times \boldsymbol{A}$, we can connect the surface integral to a line integral, using Stokes' theorem,

$$I\int_S (\nabla \times \boldsymbol{A})\cdot\boldsymbol{n}\,da = I\oint_\Gamma \boldsymbol{A}\cdot d\boldsymbol{s}, \tag{15.18}$$

where $d\boldsymbol{s}$ is the line element along Γ. So we have the energy for a circuit of any shape:

$$U = I\oint_{\text{circuit}} \boldsymbol{A}\cdot d\boldsymbol{s}. \tag{15.19}$$

In this expression $\boldsymbol{A}$ refers, of course, to the vector potential due to those currents (other than the I in the wire) which produce the field $\boldsymbol{B}$ 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 $\boldsymbol{A}$ 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 counting the energy twice (we saw a similar effect in electrostatics), so the total energy can be written

$$U = \tfrac{1}{2}\int \boldsymbol{j}\cdot\boldsymbol{A}\,dV. \tag{15.20}$$

This formula corresponds to the result we found for the electrostatic energy:

$$U = \tfrac{1}{2}\int \rho\phi \, dV. \tag{15.21}$$

So we may if we wish think of $\boldsymbol{A}$ 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 $\boldsymbol{B}$ versus $\boldsymbol{A}$

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. One 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 came 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 a point* 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 not, however, anything 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 $\boldsymbol{E}$ and $\boldsymbol{A}$). But we are not worried about what happens if the field *can* be 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 $\boldsymbol{A}$. 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

$$\phi(1) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(2)}{r_{12}}\, dV_2. \tag{15.22}$$

From this ϕ, we get the three components of $\boldsymbol{E}$ by three differential operations. This procedure is usually easier to handle than evaluating the three integrals in the vector formula

$$\boldsymbol{E}(1) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(2)\boldsymbol{e}_{12}}{r_{12}^2}\, dV_2. \tag{15.23}$$

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 $\boldsymbol{A}$ is already a vector integral:

$$A(1) = \frac{1}{4\pi\epsilon_0 c^2}\int \frac{j(2)\,dV_2}{r_{12}}, \tag{15.24}$$

which is, of course, three integrals. Also, when we take the curl of $\boldsymbol{A}$ to get $\boldsymbol{B}$, 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 $\boldsymbol{B}$ directly from

$$\boldsymbol{B}(1) = \frac{1}{4\pi\epsilon_0 c^2}\int \frac{\boldsymbol{j}(2)\times \boldsymbol{e}_{12}}{r_{12}^2}\,dV_2. \tag{15.25}$$

Using the vector potential is often more difficult for simple problems for the following reason. Suppose we are interested only in the magnetic field $\boldsymbol{B}$ at one point, and that the problem has some nice 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 $\boldsymbol{B}$ by doing the integral of Eq. (15.25). If, however, we were to find $\boldsymbol{A}$ first, we would have to compute $\boldsymbol{B}$ from *derivatives* of $\boldsymbol{A}$, so we must know what $\boldsymbol{A}$ is at all points in the *neighborhood* of the point of interest. And most of these points are off the axis of symmetry, so the integral for $\boldsymbol{A}$ gets complicated. In the ring problem, for example, we would need to use elliptic integrals. In such problems, $\boldsymbol{A}$ is clearly not very useful. It is true that in many complex problems it is easier to work with $\boldsymbol{A}$, but it would be hard to argue that this ease of technique would justify making you learn about one more vector field.

We have introduced $\boldsymbol{A}$ 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

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v}\times\boldsymbol{B}), \tag{15.26}$$

so that, given the forces, everything about the motion is determined. In any region where $\boldsymbol{B} = 0$ even if $\boldsymbol{A}$ is not zero, such as outside a solenoid, there is no discernible effect of $\boldsymbol{A}$. Therefore for a long time it was believed that $\boldsymbol{A}$ was not a "real" field. It turns out, however, that there are phenomena involving quantum mechanics which show that the field $\boldsymbol{A}$ 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. I. 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 wavelengths related to momenta, and frequencies 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 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 differentiates 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 $\boldsymbol{A}$ and ϕ provide the most direct way of introducing electromagnetic effects into quantum descriptions.

We must review a little how quantum mechanics works. We will consider again the imaginary experiment described in Chapter 37 of Vol. I, in which elec-

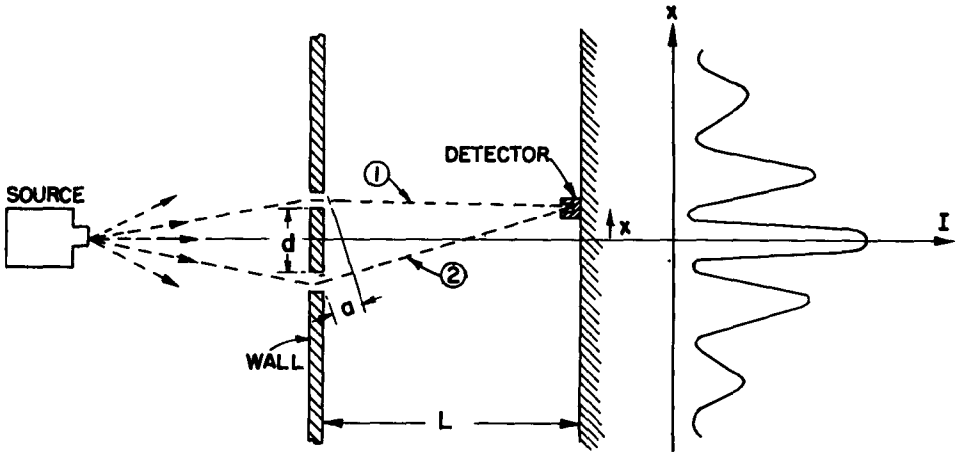


Fig. 15–5. An interference experiment with electrons (see also Chapter 37 of Vol. I).

trons 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 $C_1e^{i\Phi_1}$ and $C_2e^{i\Phi_2}$, the phase difference $\delta = \Phi_1 - \Phi_2$ determines their interference pattern [see Eq. (29.12) in Vol. I]. 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

$$\delta = \frac{a}{\lambda\!\!\!^{-}}. \tag{15.27}$$

As usual, we let $\lambda\!\!\!^{-} = \lambda/2\pi$, 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

$$\delta = \frac{x}{L}\frac{d}{\lambda\!\!\!^{-}}. \tag{15.28}$$

When x is zero, δ is zero; the waves are in phase, and the probability has a maximum. When δ is π, 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 $\boldsymbol{F} = q\boldsymbol{v} \times \boldsymbol{B}$. 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,

$$\text{Magnetic change in phase} = \frac{q}{\hbar}\int_{\text{trajectory}} \boldsymbol{A}\cdot d\boldsymbol{s}. \tag{15.29}$$

If there were no magnetic field there would be a certain phase of arrival. If there is a magnetic field anywhere, 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 mention 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:

$$\text{Electric change in phase} = -\frac{q}{\hbar}\int \phi\, dt.$$

These two expressions are correct not only for static fields, but together give the correct result for *any* electromagnetic field, static or dynamic. This is the law that replaces $\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$. 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

$$\Phi_1 = \Phi_1(B = 0) + \frac{q}{\hbar}\int_{(1)} \boldsymbol{A}\cdot d\boldsymbol{s}. \tag{15.30}$$

Similarly, the phase for trajectory (2) is

$$\Phi_2 = \Phi_2(B = 0) + \frac{q}{\hbar}\int_{(2)} \boldsymbol{A}\cdot d\boldsymbol{s}. \tag{15.31}$$

The interference of the waves at the detector depends on the phase difference

$$\delta = \Phi_1(B = 0) - \Phi_2(B = 0) + \frac{q}{\hbar}\int_{(1)} \boldsymbol{A}\cdot d\boldsymbol{s} - \frac{q}{\hbar}\int_{(2)} \boldsymbol{A}\cdot d\boldsymbol{s}. \tag{15.32}$$

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). So we have

$$\delta = \delta(B = 0) + \frac{q}{\hbar}\oint_{(1-2)} \boldsymbol{A}\cdot d\boldsymbol{s}. \tag{15.33}$$

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 $\boldsymbol{A}$ and $\boldsymbol{A}'$ whose difference is the gradient of some scalar function $\nabla\psi$, both represent the same magnetic field, since the curl of a gradient is zero. They give, therefore, the same classical force $q\boldsymbol{v} \times \boldsymbol{B}$. If in quantum mechanics the effects depend on the vector potential, *which* of the many possible $\boldsymbol{A}$-functions is correct?

The answer is that the same arbitrariness in $\boldsymbol{A}$ continues to exist for quantum mechanics. If in Eq. (15.33) we change $\boldsymbol{A}$ to $\boldsymbol{A}' = \boldsymbol{A} + \nabla\psi$, the integral on $\boldsymbol{A}$ becomes

$$\oint_{(1-2)} \boldsymbol{A}'\cdot d\boldsymbol{s} = \oint_{(1-2)} \boldsymbol{A}\cdot d\boldsymbol{s} + \oint_{(1-2)} \nabla\psi\cdot d\boldsymbol{s}.$$

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 $\boldsymbol{A}$ and $\boldsymbol{A}'$ give the same phase differences and the same quantum-mechanical interference effects. In both classical and quantum theory it is only the curl of $\boldsymbol{A}$ that matters; any choice of the function of $\boldsymbol{A}$ 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 $\boldsymbol{A}$ around a closed path is the flux of $\boldsymbol{B}$ through the path, which here is the flux between paths (1) and (2). Equation (15.33) can, if we wish, be written as

$$\delta = \delta(B = 0) + \frac{q}{\hbar}[\text{flux of } \boldsymbol{B} \text{ between (1) and (2)}], \qquad (15.34)$$

where by the flux of $\boldsymbol{B}$ we mean, as usual, the surface integral of the normal component of $\boldsymbol{B}$. The result depends only on $\boldsymbol{B}$, and therefore only on the curl of $\boldsymbol{A}$.

Now because we can write the result in terms of $\boldsymbol{B}$ as well as in terms of $\boldsymbol{A}$, you might be inclined to think that the $\boldsymbol{B}$ holds its own as a "real" field and that the $\boldsymbol{A}$ 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 $\boldsymbol{B}$ is zero—or at least arbitrarily small—at any place where there is some chance to find the particles, 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 $\boldsymbol{B}$-field inside but none outside, while there is lots of $\boldsymbol{A}$ circulating around outside, as shown in Fig. 15–6. If we arrange a situation in which electrons are to be found only *outside* of the solenoid—only where there is $\boldsymbol{A}$—there will still be an influence on the motion, according to Eq. (15.33). Classically, that is impossible. Classically, the force depends only on $\boldsymbol{B}$; in order to know that the solenoid is carrying current, the particle must go through it. But quantum-mechanically you can find out that there is a magnetic field inside the solenoid by going *around* it—without ever going close to it!

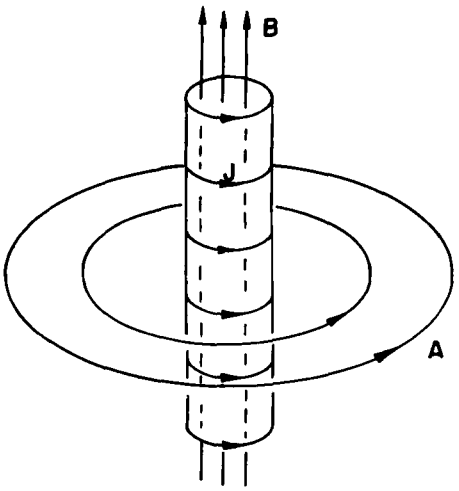


Fig. 15–6. The magnetic field and vector potential of a long solenoid.

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 diffraction 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?

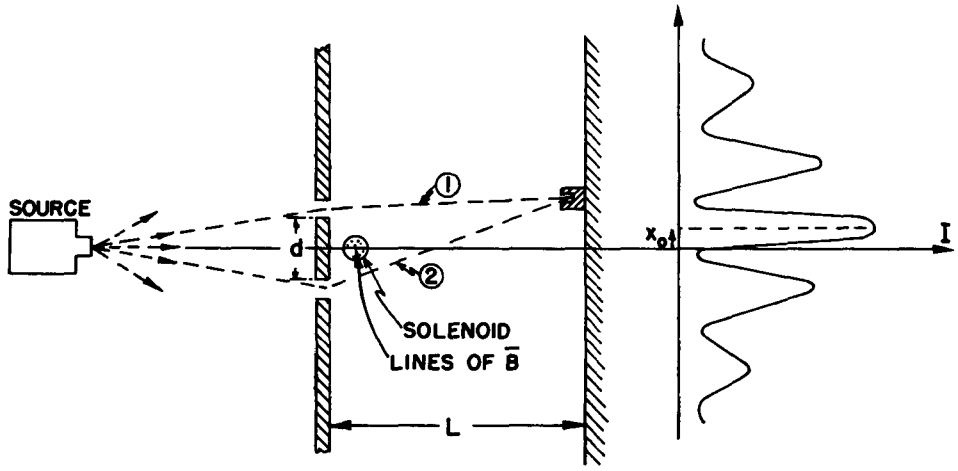


Fig. 15–7. A magnetic field can influence the motion of electrons even though it exists only in regions where there is an arbitrarily small probability of finding the electrons.

We compare the situation with and without a current through the solenoid. If we have no current, we have no $\boldsymbol{B}$ or $\boldsymbol{A}$ 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 $\boldsymbol{B}$ inside, then there is an $\boldsymbol{A}$ outside. There is a shift in the phase difference proportional to the circulation of $\boldsymbol{A}$ 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 $\boldsymbol{B}$ inside is a constant for any pair of paths, so also is the circulation of $\boldsymbol{A}$. For every arrival point there is the same phase change; this corresponds

to shifting the entire pattern in x by a constant 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 $\delta(B = 0)$, we have

$$x_0 = -\frac{L}{d}\,\lambda\!\!\!^{-}\,\frac{q}{\hbar}\oint_{(1-2)} \boldsymbol{A}\cdot d\boldsymbol{s}, \tag{15.35}$$

or

$$x_0 = -\frac{L}{d}\,\lambda\!\!\!^{-}\,\frac{q}{\hbar}\,[\text{flux of } \boldsymbol{B} \text{ between (1) and (2)}]. \tag{15.36}$$

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 $\boldsymbol{A}$-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 $\boldsymbol{B}$-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 man 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 $\boldsymbol{A}$ 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" even in quantum mechanics. It seems strange in retrospect that no one thought of discussing this experiment until 1956, when Bohm and Aharanov 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 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 $q\boldsymbol{v} \times$ the curl of $\boldsymbol{A}$. 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 fields, 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 $\boldsymbol{B}$ comes out of the plane of the figure, the flux as we have defined it is negative and x_0 is positive.

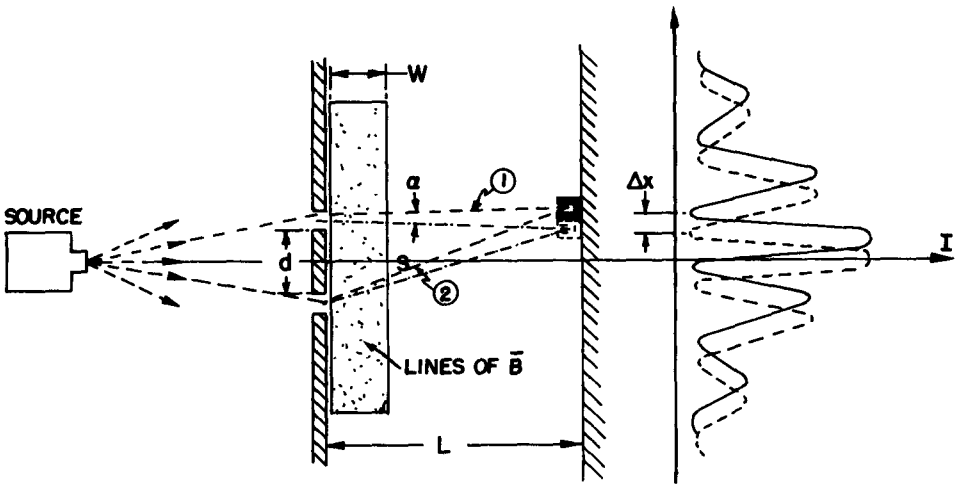


Fig. 15–8. The shift of the interference pattern due to a strip of magnetic field.

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. (That can easily be arranged; the backstop can be put as far out as we want.) In order to calculate the shift in phase, we must take the two integrals of $\boldsymbol{A}$ along the two trajectories (1) and (2). They differ, as we have seen, merely by the flux of $\boldsymbol{B}$ between the paths. To our approximation, the flux is Bwd. The phase difference for the two paths is then

$$\delta = \delta(B = 0) + \frac{q}{\hbar} Bwd. \tag{15.37}$$

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),

$$\Delta x = \frac{L\lambda}{d} \Delta\delta = \frac{L\lambda}{d} [\delta - \delta(B = 0)].$$

Using (15.37) for $\delta - \delta(B = 0)$,

$$\Delta x = L\lambda \frac{q}{\hbar} Bw. \tag{15.38}$$

Such a shift is equivalent to deflecting all the trajectories by the small angle α (see Fig. 15–8), where

$$\alpha = \frac{\Delta x}{L} = \frac{\lambda}{\hbar} qBw. \tag{15.39}$$

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 $q\boldsymbol{v} \times \boldsymbol{B}$ which lasts for a time w/v. The change in their transverse momentum is just equal to this impulse, so

$$\Delta p_x = qwB. \tag{15.40}$$

The angular deflection [Fig. 15–9(b)] is equal to the ratio of this transverse momentum to the total momentum p. We get that

$$\alpha' = \frac{\Delta p_x}{p} = \frac{qwB}{p}. \tag{15.41}$$

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-

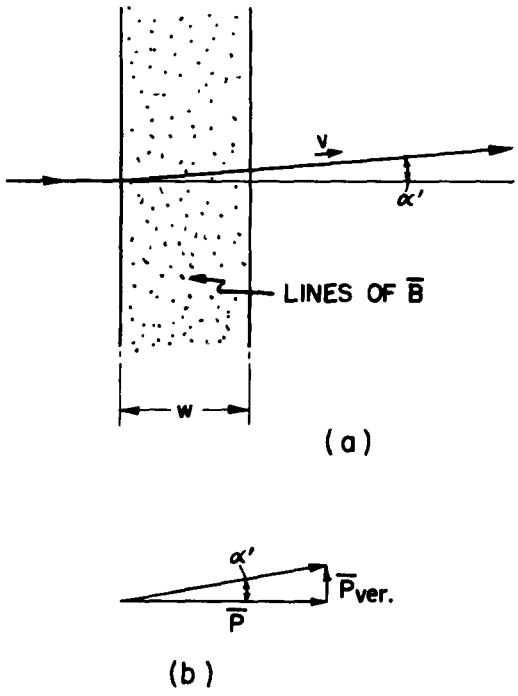


Fig. 15–9. Deflection of a particle due to passage through a strip of magnetic field.

tum amplitude varying with the wavelength $\lambda = \hbar/p$. With this equality, α and α' are identical; the classical 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) appears 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 mysterious, 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 the danger in this process that before we get to see the complete story, the incomplete truths learned on the way 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 (often called the *Lorentz force*) $F = q(E + v \times B)$ is *true*. It is only Coulomb's law that is false, to be used 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, $\nabla \cdot E = \rho/\epsilon_0$, remains, but the curl of E is *not* zero in general. So E cannot always be equated to the gradient of a scalar—the electrostatic potential. We will see that a scalar potential still remains, 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 cur-

Table 15–1

FALSE IN GENERAL (true only for statics)	TRUE ALWAYS
$F = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r^2}$ (Coulomb's law)	$F = q(E + v \times B)$ (Lorentz force) $\rightarrow \nabla \cdot E = \frac{\rho}{\epsilon_0}$ (Gauss' law)
$\nabla \times E = 0$ $E = -\nabla\phi$ $E(1) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(2)e_{12}}{r_{12}^2}\,dV_2$ For conductors, $E = 0$, ϕ = constant. $Q = CV$	$\rightarrow \nabla \times E = -\frac{\partial B}{\partial t}$ (Faraday's law) $E = -\nabla\phi - \frac{\partial A}{\partial t}$ In a conductor, E makes currents.
	$\rightarrow \nabla \cdot B = 0$ (No magnetic charges) $B = \nabla \times A$
$c^2\nabla \times B = \frac{j}{\epsilon_0}$ (Ampere's law) $B(1) = \frac{1}{4\pi\epsilon_0 c^2}\int \frac{j(2) \times e_{12}}{r_{12}^2}\,dV_2$	$\rightarrow c^2\nabla \times B = \frac{j}{\epsilon_0} + \frac{\partial E}{\partial t}$
$\nabla^2\phi = -\frac{\rho}{\epsilon_0}$ (Poisson's equation) $\nabla^2 A = -\frac{j}{\epsilon_0 c^2}$ with $\nabla \cdot A = 0$	$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = -\frac{\rho}{\epsilon_0}$ and $\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{j}{\epsilon_0 c^2}$ with $c^2\nabla \cdot A + \frac{\partial\phi}{\partial t} = 0$
$\phi(1) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(2)}{r_{12}}\,dV_2$ $A(1) = \frac{1}{4\pi\epsilon_0 c^2}\int \frac{j(2)}{r_{12}}\,dV_2$	$\phi(1,t) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(2,t')}{r_{12}}\,dV_2$ and $A(1,t) = \frac{1}{4\pi\epsilon_0 c^2}\int \frac{j(2,t')}{r_{12}}\,dV_2$ with $t' = t - \frac{r_{12}}{c}$
$U = \frac{1}{2}\int \rho\phi\,dV + \frac{1}{2}\int j \cdot A\,dV$	$U = \int \left(\frac{\epsilon_0}{2}E \cdot E + \frac{\epsilon_0 c^2}{2}B \cdot B\right)dV$

The equations marked by an arrow (→) are Maxwell's equations.

rents. So in varying fields a conductor is *not* an equipotential. It also follows that the idea of a capacitance is no longer precise.

Since there are no magnetic charges, the divergence of $\boldsymbol{B}$ is *always* zero. So $\boldsymbol{B}$ can always be equated to $\nabla \times \boldsymbol{A}$. (Everything doesn't change!) But the generation of $\boldsymbol{B}$ is not only from currents; $\nabla \times \boldsymbol{B}$ is proportional to the current density *plus* a new term $\partial \boldsymbol{E}/\partial t$. This means that $\boldsymbol{A}$ is related to currents by a new equation. It is also related to ϕ. If we make use of our freedom to choose $\nabla \cdot \boldsymbol{A}$ for our own convenience, the equations for $\boldsymbol{A}$ or ϕ can be arranged to take on a simple and elegant form. We therefore make the condition that $c^2 \nabla \cdot \boldsymbol{A} = -\partial\phi/\partial t$, and the differential equations for $\boldsymbol{A}$ or ϕ appear as shown in the table.

The potentials $\boldsymbol{A}$ 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 $\boldsymbol{j}$ 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 $\boldsymbol{A}$ and ϕ, we get $\boldsymbol{B}$ from $\nabla \times \boldsymbol{A}$, as before, and $\boldsymbol{E}$ from $-\nabla\phi - \partial \boldsymbol{A}/\partial t$.

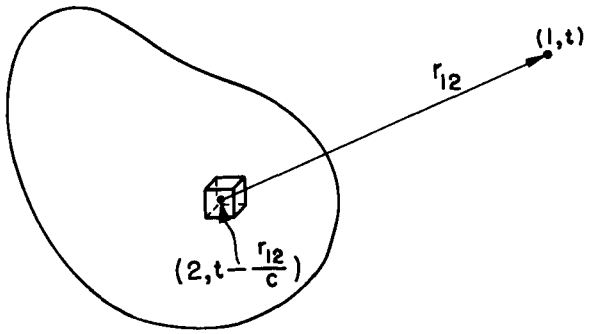


Fig. 15–10. The potentials at point (1) and at the time t are given by summing the contributions from each element of the source at the roving point (2), using the currents and charges which were present at the earlier time $t - r_{12}/c$.

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 contrary example is the expression for the electrostatic energy in terms of a volume integral of $\rho\phi$. 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.

16

Induced Currents

16–1 Motors and generators

16–2 Transformers and inductances

16–3 Forces on induced currents

16–4 Electrical technology

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 as 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 excitements whenever there is a mechanical 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 electromagnetic motor is shown in bare outline in Fig. 16–1. A permanent 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 on a shaft so that 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. Thus 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 make the coil turn, 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.

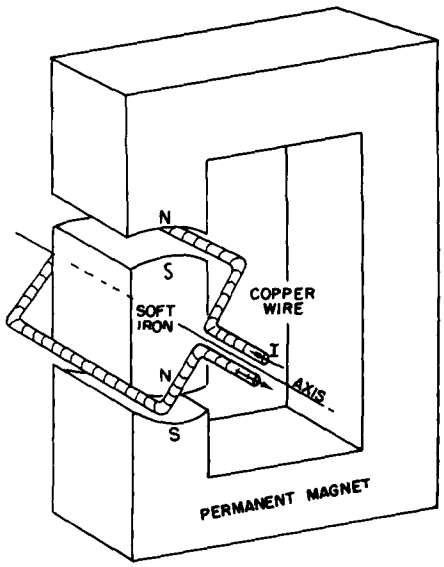


Fig. 16–1. Schematic outline of a simple electromagnetic motor.

With the realization that electric currents make magnetic fields, people immediately suggested that, somehow or other, magnets might 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 finding 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 such law as "likes prefer 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 1840 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. We say that currents are *induced*. This was the induction effect discovered

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 first 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 $\boldsymbol{v} \times \boldsymbol{B}$ 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—at right angles to the field and to the motion. The force 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 push—by electric repulsion—the electrons a little farther down the wire; they, in turn, repel the electrons a little farther on, and so on for a long distance. An amazing thing.

It was so amazing 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 way wires carry 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 an 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 $\boldsymbol{v} \times \boldsymbol{B}$ force on the electrons in the wire. This is the new effect that Faraday found. Today, we might hope to understand it from a relativity 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.

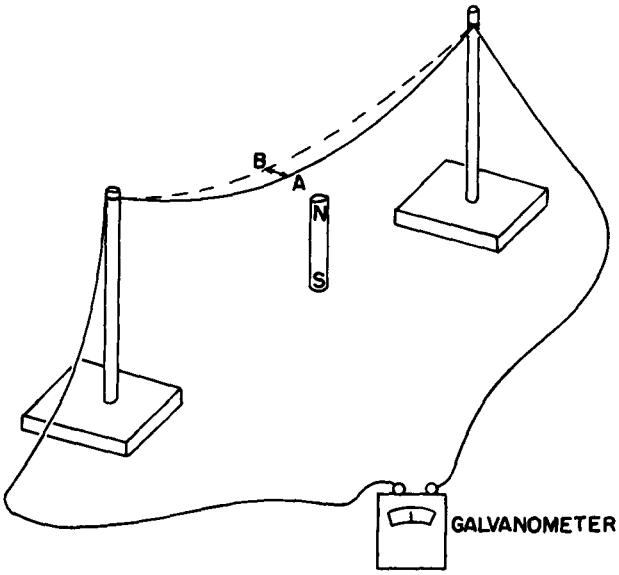


Fig. 16–2. Moving a wire through a magnetic field produces a current, as shown by the galvanometer.

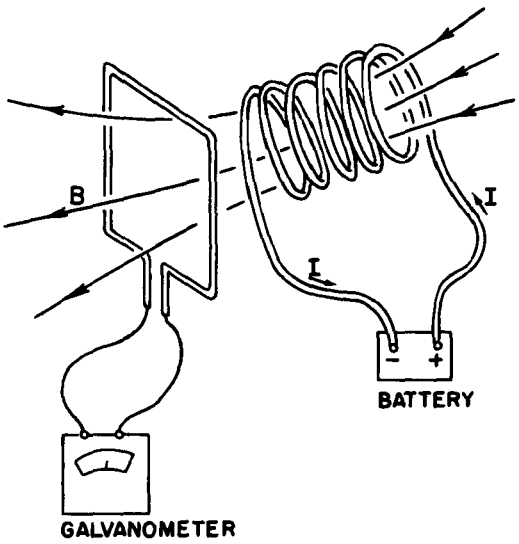


Fig. 16–3. A coil with current produces a current in a second coil if the first coil is moved or if its current is changed.

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 (this flux is the normal component of $\boldsymbol{B}$ 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 flux 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-rings—(just so the wires won't get twisted) we have an alternating-current generator.

Or we can also arrange, 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, an "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.

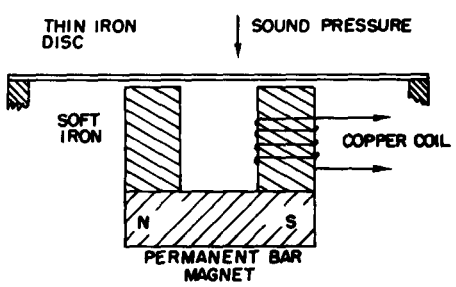


Fig. 16–4. A telephone transmitter or receiver.

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 gadget, 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-button 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 discoveries is not that an emf exists in a moving coil—which we can understand in terms of the magnetic force $q\boldsymbol{v} \times \boldsymbol{B}$—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 magnetic 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 in Fig. 16–5. Now we connect one of the coils—coil (a)—to an alternating-current generator. The continually changing current produces a continuously varying magnetic field. This varying field generates an alternating emf in the second coil—coil (b). This emf can, for example, produce enough power to light an electric bulb.

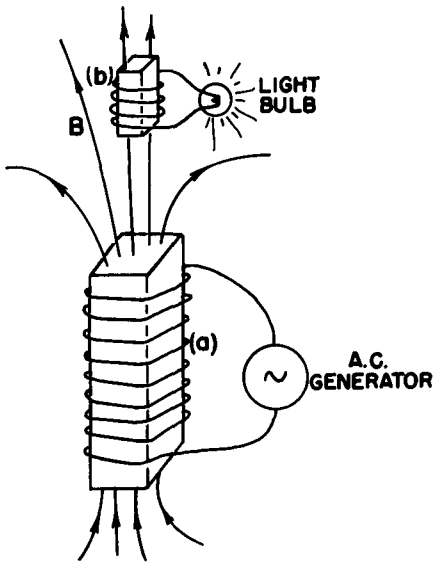


Fig. 16–5. Two coils, wrapped around bundles of iron sheets, allow a generator to light a bulb with no direct connection.

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 (b) can be larger or smaller than the current in coil (a). 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 made much larger than that in the generator by winding coil (b) with many turns, since 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 emf's 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. 16–5 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 itself 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 its 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 $\boldsymbol{B}$ that opposes the change in $\boldsymbol{B}$ 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 winding of Fig. 16–3.

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 to the current when the current is increasing, and it is in the direction of the current

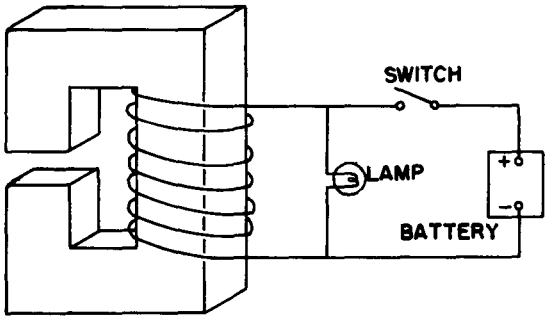


Fig. 16–6. Circuit connections for an electromagnet. The lamp allows the passage of current when the switch is opened, preventing the appearance of excessive emf's.

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 emf from 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 gadget 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 coil is connected to an alternating-current generator by closing the switch, the ring flies into the air. The force comes, 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 making a north pole at its top, the induced current in the ring is making a downward-point north pole. The ring and the coil are repelled just like 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.

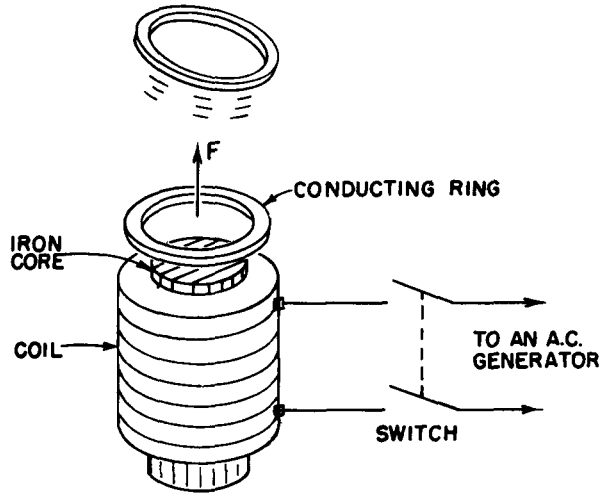


Fig. 16–7. A conducting ring is strongly repelled by an electromagnet with a varying current.

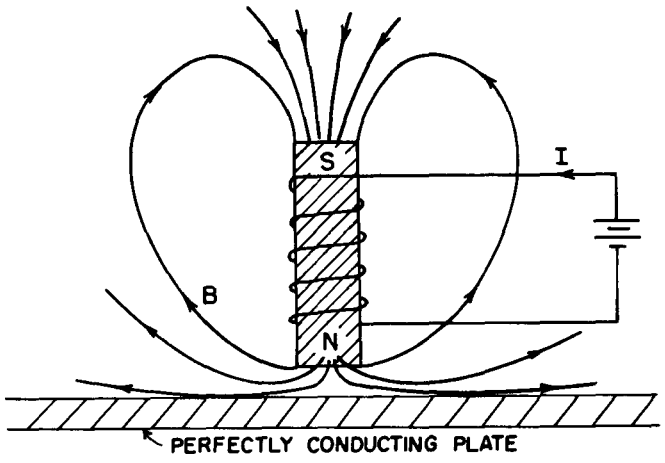


Fig. 16–8. An electromagnet near a perfectly conducting plate.

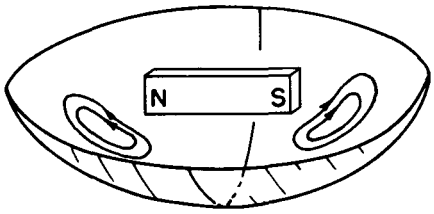


Fig. 16–9. A bar magnet is suspended above a superconducting bowl, by the repulsion of eddy currents.

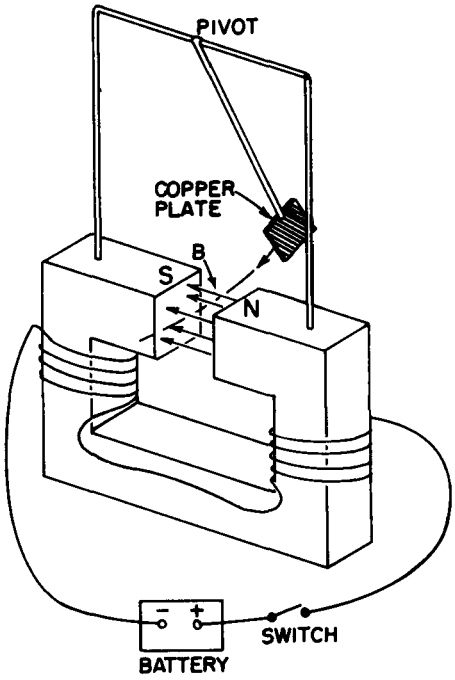


Fig. 16–10. The braking of the pendulum shows the forces due to eddy currents.

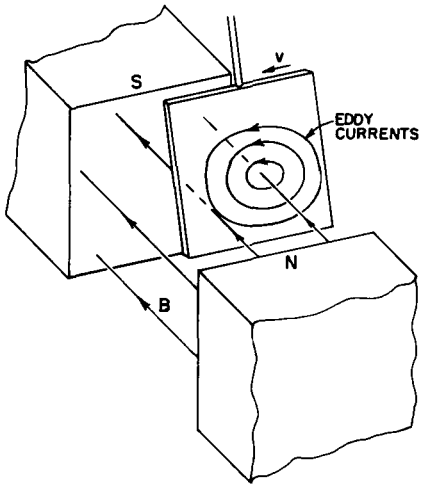


Fig. 16–11. The eddy currents in the copper pendulum.

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 *slightest* emf would generate an arbitrarily large current—which really means that there can be no emf's at all. Any attempt to make a magnetic flux go through such a sheet generates currents that create opposite $\boldsymbol{B}$ fields—all with infinitesimal emf's, 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. There are no perfect conductors at ordinary temperatures, but some materials become perfect conductors at low enough temperatures. For instance, below 3.8°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 an emf to keep them going, and to have an emf the flux must keep changing. The flux of the magnetic field gradually penetrates the conductor.

In a normal conductor, there are not only repulsive forces from eddy currents, but there can also be sidewise 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 changing of 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 enters the gap 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 at first bring 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 slots 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 honey.

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—end over end—near a conducting plate or ring, the ring is dragged around; currents in the ring will create a torque that tends to rotate the ring with the magnet.

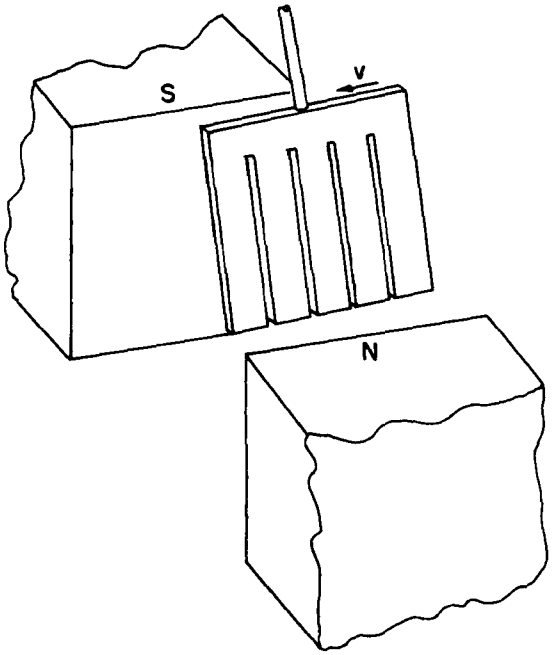


Fig. 16–12. Eddy-current effects are drastically reduced by cutting slots in the plate.

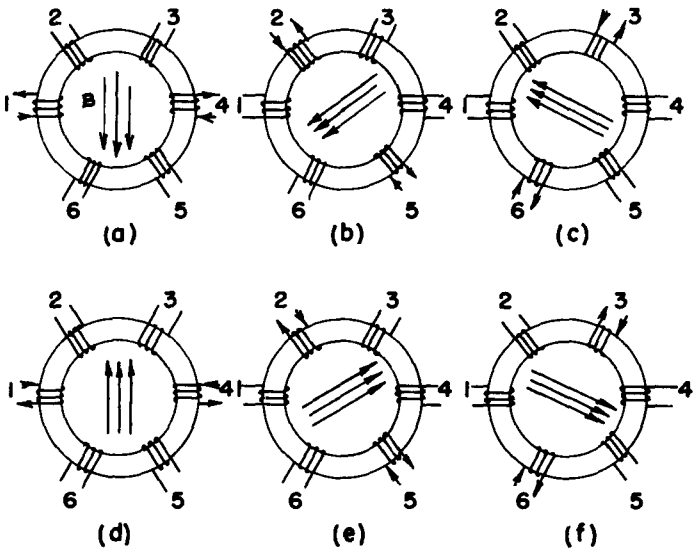


Fig. 16–13. Making a rotating magnetic field.

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 coils 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.

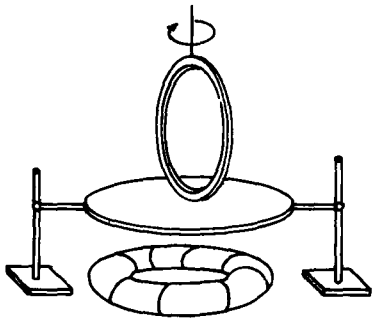


Fig. 16–14. The rotating field of Fig. 16–13 can be used to provide torque on a conducting ring.

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 $\boldsymbol{B}$ 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 change of flux through it, so the magnetic field above the plate always lags the field above that half of the pole which is not covered. This so-called "shaded-pole" effect produces a field which in the "shaded" region varies

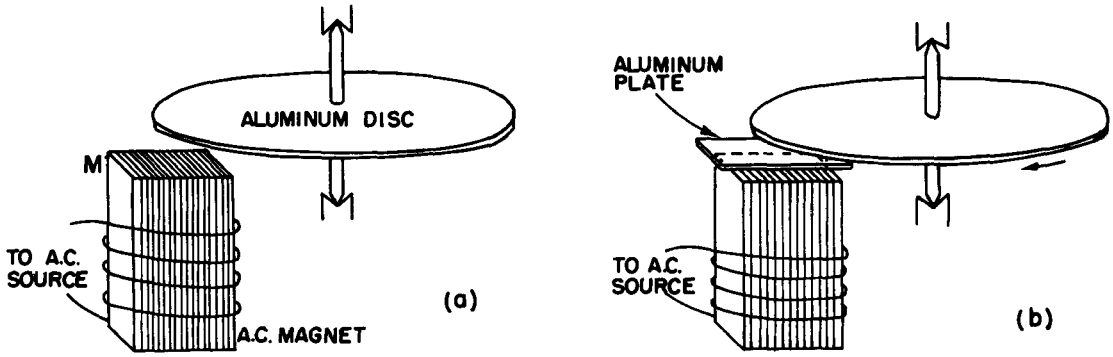


Fig. 16–15. A simple example of a shaded-pole induction motor.

much like that in the "unshaded" region except that it is delayed a constant amount in time. The whole effect 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 toys but examples chosen in most 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 interwoven. With two parts—one that turns and one that doesn't. All a complex intermixture of a few materials, mostly iron and copper but also some paper and shellac for insulation. A revolving monster thing. A generator. Somewhere out of the mess of copper and iron come 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 curls 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 left in the transformer, that energy would need to be taken out somehow. 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 recreate the current in another form . . . sometimes to engines turning for big industrial purposes . . . to more transformers . . . then more splitting and spreading . . . until finally the river is spread throughout the whole city—turning motors, making heat, making light, working gadgetry. The miracle of hot lights from 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 river are running the dentist's drill, come again into the building of extremely fine instruments . . . 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 created magnetic fields . . . blocks of rotating 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 uncovers Boulder Dam, we may guess that he would admire the beauty of its curves. But also the explorers from some great future civilizations 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 forces 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 motors can be taken from a shelf. 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 returning 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 created. 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 life 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 that apply, no matter what. (We have not yet completed the base, because we have yet to consider in detail the properties of iron and 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 baby developed into a prodigy and changed the face of the earth in ways its proud father could never have imagined.

17

The Laws of Induction

17–1 The physics of induction

17–2 Exceptions to the "flux rule"

17–3 Particle acceleration by an induced electric field; the betatron

17–4 A paradox

17–5 Alternating-current generator

17–6 Mutual inductance

17–7 Self-inductance

17–8 Inductance and magnetic energy

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 wLB, so the "flux rule" would give for the emf—which we write as $\mathcal{E}$—

$$\mathcal{E} = wB\frac{dL}{dt} = wBv,$$

where v is the speed of translation of the crossbar.

Now we should be able to understand this result from the magnetic $\boldsymbol{v} \times \boldsymbol{B}$ forces on the charges in the moving crossbar. These charges will feel a force, tangential to the wire, equal to vB per unit charge. It is constant along the length w of the crossbar and zero elsewhere, so the integral is

$$\mathcal{E} = wvB,$$

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 $\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$; there are no new special "forces due to changing magnetic fields." Any forces on charges at rest in a stationary wire come from the $\boldsymbol{E}$ 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

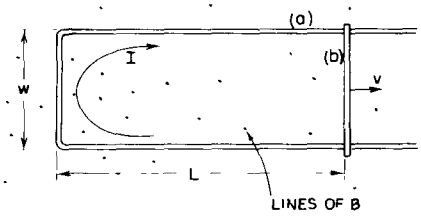


Fig. 17–1. An emf is induced in a loop if the flux is changed by varying the area of the circuit.

field which drives the electrons around the wire—and so is responsible for the emf in a stationary circuit when there is a changing magnetic flux.

The general law for the electric field associated with a changing magnetic field is

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}. \tag{17.1}$$

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

$$\oint_\Gamma \boldsymbol{E} \cdot d\boldsymbol{s} = \int_S (\nabla \times \boldsymbol{E}) \cdot \boldsymbol{n}\, da = -\int_S \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}\, da, \tag{17.2}$$

where, as usual, Γ is any closed curve and S is any surface bounded by it. Here, remember, Γ is a *mathematical* curve fixed in space, and S is a fixed surface. Then the time derivative can be taken outside the integral and we have

$$\begin{aligned} \oint_\Gamma \boldsymbol{E} \cdot d\boldsymbol{s} &= -\frac{\partial}{\partial t} \int_S \boldsymbol{B} \cdot \boldsymbol{n}\, da \\ &= -\frac{\partial}{\partial t} (\text{flux through } S). \end{aligned} \tag{17.3}$$

Applying this relation to a curve Γ 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. 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 (or both). The two possibilities—"circuit moves" or "field changes"—are not distinguished in the statement of the rule. Yet in our explanation of the rule we have used two completely distinct laws for the two cases—$\boldsymbol{v} \times \boldsymbol{B}$ for "circuit moves" and $\nabla \times \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t$ for "field changes."

We know of no other place in physics where such a simple and accurate general principle requires for its real understanding an analysis in terms of *two different phenomena*. Usually such a beautiful generalization is found to stem from a single deep underlying principle. Nevertheless, in this case there does not appear to be any such profound implication. We have to understand the "rule" as the combined effects of two quite separate phenomena.

We must look at the "flux rule" in the following way. In general, the force per unit charge is $\boldsymbol{F}/q = \boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}$. In moving wires there is the force from the second term. Also, there is an $\boldsymbol{E}$-field if there is somewhere a changing magnetic field. They are independent effects, but the emf around the loop of wire is always equal to the rate of change of 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 emf's. 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 part of the emf that comes from the $\boldsymbol{E}$-field does not depend on the existence of a physical wire (as does the $\boldsymbol{v} \times \boldsymbol{B}$ part). The $\boldsymbol{E}$-field can exist in free space, and its line integral around any imaginary line fixed in space is the rate of change of the flux of $\boldsymbol{B}$ through that line. (Note that this is quite unlike the $\boldsymbol{E}$-field produced by static charges, for in that case the line integral of $\boldsymbol{E}$ around a closed loop is always zero.)

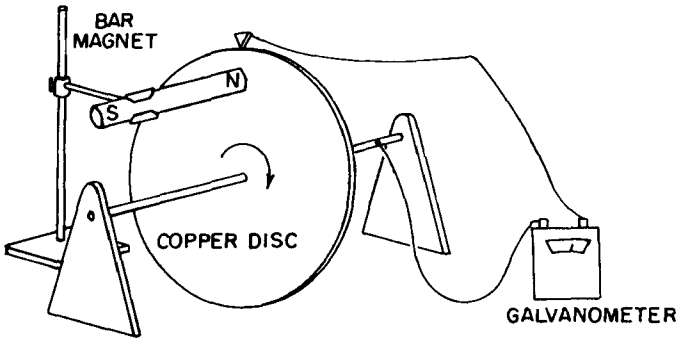


Fig. 17–2. When the disc rotates there is an emf from $\boldsymbol{v} \times \boldsymbol{B}$, but with no change in the linked flux.

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 rubs on the outer periphery of the disc. A circuit is completed through a galvanometer. As the disc rotates, the "circuit," in the sense of the place in space where the currents are, is always the same. But the part of the "circuit" in the disc is in material which is moving. Although the flux 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 the $\boldsymbol{v} \times \boldsymbol{B}$ 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" (again 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 imagine the "circuit" to be completed through 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 rocking can be done with small motions, so that $\boldsymbol{v} \times \boldsymbol{B}$ 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 laws. The *correct* physics is always given by the two basic laws

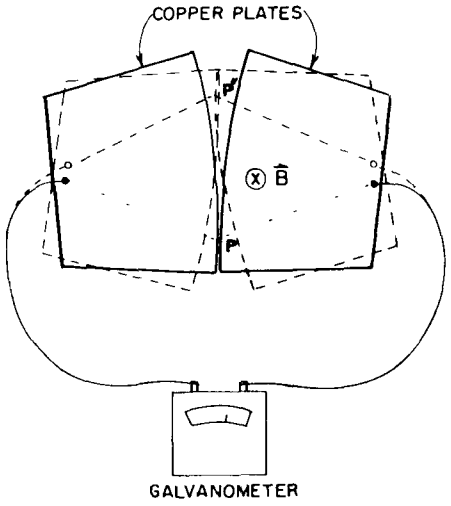


Fig. 17–3. When the plates are rocked in a uniform magnetic field, there can be a large change in the flux linkage without the generation of an emf.

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}),$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}.$$

17–3 Particle acceleration by an induced electric field; the betatron

We have said that the electromotive force generated by a changing magnetic field can exist even without conductors; that is, there can be magnetic induction without wires. We may still imagine an electromotive force around an arbitrary mathematical curve in space. It is defined as the tangential component of $\boldsymbol{E}$ integrated around the curve. Faraday's law says that this 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 induced electric field, we want now to consider the motion of an electron in a changing magnetic field. We imagine a magnetic field which, everywhere on a plane, points in a vertical direction, as shown in Fig. 17–4. The magnetic field is produced by an electromagnet, but we will not worry about the details For our example we will imagine that the magnetic field is symmetric about some axis, i.e., that the strength of the magnetic field will depend only on the distance from the axis. The magnetic field is also varying with time We now imagine an electron that is moving in this field on a path that is a circle of constant radius with its center at the axis of the field. (We will see later

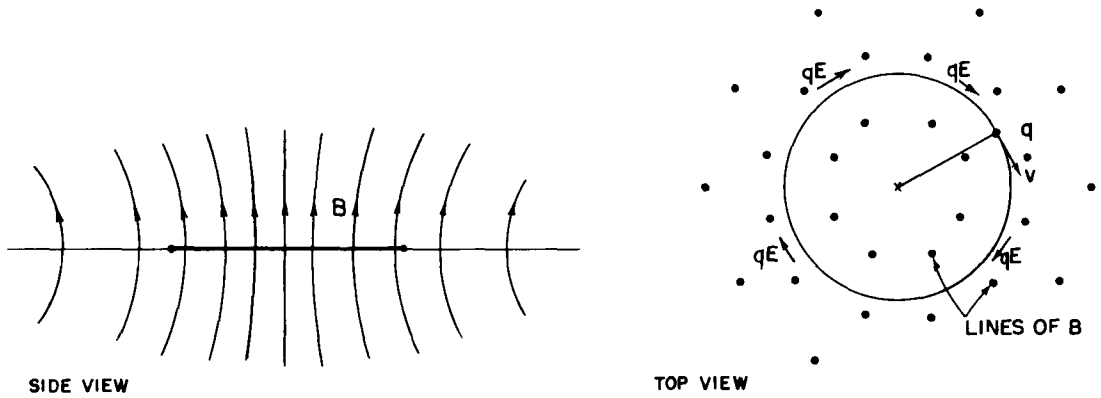


Fig. 17–4. An electron accelerating in an axially symmetric, time-varying magnetic field.

how this motion can be arranged.) Because of the changing magnetic field, there will be an electric field $\boldsymbol{E}$ tangential to the electron's orbit which will drive it around 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 $\boldsymbol{E}$ around the orbit is equal to the rate of change of the magnetic flux through the circle. The line integral of $\boldsymbol{E}$ is just its magnitude times the circumference of the circle, $2\pi r$. 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

$$2\pi rE = \frac{\partial}{\partial t}(B_{\text{av}} \cdot \pi r^2).$$

Since we are assuming r is constant, E is proportional to the time derivative of the average field:

$$E = \frac{r}{2}\frac{dB_{\text{av}}}{dt}. \tag{17.4}$$

The electron will feel the electric force $q\boldsymbol{E}$ 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

$$qE = \frac{dp}{dt}. \tag{17.5}$$

For the circular orbit we have assumed, 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. (17.5). Combining Eqs. (17.5) and (17.4), we may relate the rate of change of momentum to the change of the average magnetic field:

$$\frac{dp}{dt} = \frac{qr}{2}\frac{dB_{\text{av}}}{dt}. \tag{17.6}$$

Integrating with respect to t, we find for the electron's momentum

$$p = p_0 + \frac{qr}{2}\Delta B_{\text{av}}, \tag{17.7}$$

where p_0 is the momentum with which the electrons start out, and ΔB_{av} is the subsequent change in B_{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 11 of Vol. I the principle involved. If we arrange that there is a magnetic field $\boldsymbol{B}$ at the orbit of the electron, there will be a transverse force $q\boldsymbol{v} \times \boldsymbol{B}$ which, for a suit-

ably chosen $\boldsymbol{B}$, can cause the electron to keep moving on its assumed orbit. In the betatron this transverse 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 (see Fig 17–4), $\boldsymbol{B}$ is at right angles to $\boldsymbol{v}$, so the transverse force is qvB. Thus the force is equal to the rate of change of the transverse component p_t of the momentum:

$$qvB = \frac{dp_t}{dt}. \tag{17.8}$$

When a particle is moving in a *circle*, the rate of change of its transverse momentum is equal to the magnitude of the total momentum times ω, the angular velocity of rotation (following the arguments of Chapter 11, Vol. I):

$$\frac{dp_t}{dt} = \omega p, \tag{17.9}$$

where, since the motion is circular,

$$\omega = \frac{v}{r}. \tag{17.10}$$

Setting the magnetic force equal to the transverse acceleration, we have

$$qvB_{\text{orbit}} = p\frac{v}{r}, \tag{17.11}$$

where B_{orbit} is the field at the radius r.

As the betatron operates, the momentum of the electron grows in proportion to B_{av}, according to Eq. (17.7), and if the electron is to continue to move in its proper circle, Eq. (17.11) must continue to hold as the momentum of the electron increases. The value of B_{orbit} must increase in proportion to the momentum p. Comparing Eq. (17.11) with Eq. (17.7), which determines p, we see that the following relation must hold between B_{av}, the average magnetic field *inside* the orbit at the radius r, and the magnetic field B_{orbit} at the orbit:

$$\Delta B_{\text{av}} = 2\,\Delta B_{\text{orbit}}. \tag{17.12}$$

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 attaining 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—to 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 should happen. Of course, in physics there are never any real paradoxes because there is only one correct answer; at least we believe that nature will

act in only one way (and that is the *right way*, naturally). So in physics a paradox is only a confusion in our own understanding. Here is our paradox.

Imagine that we construct a device like that shown in Fig. 17–5. There is a thin, circular plastic disc supported on a concentric shaft with excellent bearings, so that it is quite free to rotate. On the disc is a coil of wire in the form of a short solenoid concentric with the axis of rotation. This solenoid carries a steady current I provided by a small battery, also mounted on the disc. Near the edge of the disc and spaced uniformly around its circumference are a number of small metal spheres insulated from each other and from the solenoid by the plastic material of the disc. Each of these small conducting spheres is charged with the same electrostatic charge Q. Everything is quite stationary, and the disc is at rest. Suppose now that by some accident—or by prearrangement—the current in the solenoid is interrupted, without, however, any intervention from the outside. So long as the current continued, there was a magnetic flux through the solenoid more or less parallel to the axis of the disc. When the current is interrupted, this flux must go to zero. There will, therefore, be an electric field induced which will circulate around in circles centered at the axis. The charged spheres on 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 for all 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 disappears, the disc would begin to rotate. If we knew the moment of inertia of the disc, the current in the solenoid, and the charges on the small spheres, we could compute the resulting angular velocity.

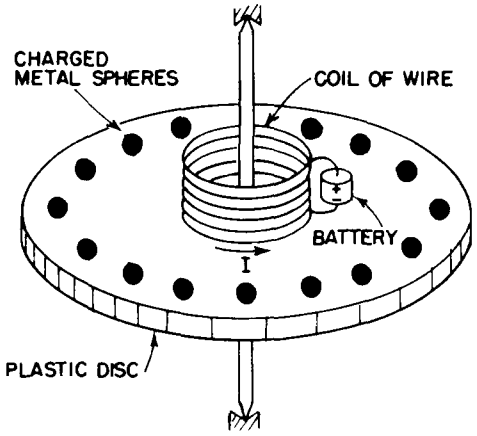


Fig. 17–5. Will the disc rotate if the current I is stopped?

But we could also make a different argument. Using the principle of the conservation of angular momentum, we could say that the angular momentum of the disc with all its equipment is initially zero, and so the angular momentum of the assembly should remain zero. There should be no rotation when the current is stopped. Which argument is correct? Will the disc rotate 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 nonessential feature, such as the asymmetric position of a battery, for example. In fact, you can imagine an ideal situation such as the following: The solenoid is made of superconducting wire through which there is a current. After the disc has been carefully placed at rest, the temperature of the solenoid is allowed to rise slowly. When the temperature of the wire reaches the transition temperature between superconductivity and normal conductivity, the current in the solenoid will be brought to zero by the resistance of the wire. The flux will, as before, fall to zero, and there will be an electric field around the axis. We should also warn you that the solution is not easy, nor is it a trick. When you figure it out, you will have discovered an important principle of electromagnetism.

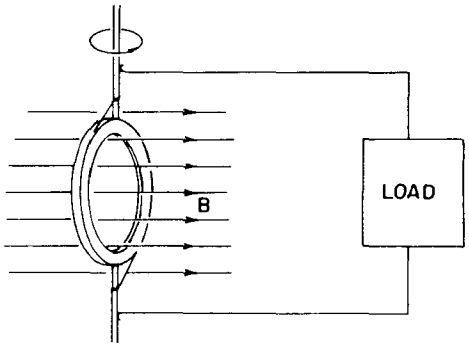


Fig. 17–6. A coil of wire rotating in a uniform magnetic field—the basic idea of the ac generator.

17–5 Alternating-current generator

In the remainder of this chapter we apply the principles of Section 17–1 to analyze a number of the phenomena discussed in Chapter 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 result can also be achieved by a fixed coil in a magnetic field whose direction rotates in the manner described in the last chapter. We will consider only the former case. Suppose we have a circular coil of wire which can be turned on an axis along one of its diameters. Let this coil be located in a uniform magnetic field perpendicular to the axis of rotation, as in Fig. 17–6. We also imagine that the two ends of the coil are brought to external connections through some kind of sliding contacts.

Due to the rotation of the coil, the magnetic flux through it will be changing. The circuit of the coil will therefore have an emf in it. Let S be the area of the coil and θ the angle between the magnetic field and the normal to the plane of the coil.*

* Now that we are using the letter A for the vector potential, we prefer to let S stand for a Surface area.

The flux through the coil is then

$$BS \cos \theta. \tag{17.13}$$

If the coil is rotating at the uniform angular velocity ω, θ varies with time as $\theta = \omega t$. The emf $\mathcal{E}$ in the coil is then

$$\mathcal{E} = -\frac{d}{dt}(\text{flux}) = -\frac{d}{dt}(BS \cos \omega t),$$

or

$$\mathcal{E} = BS\omega \sin \omega t. \tag{17.14}$$

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 $\boldsymbol{E}$ 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,

$$V = BS\omega \sin \omega t = V_0 \sin \omega t.$$

The potential difference between the wires varies as $\sin \omega t$. 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 pushed some excess charges out to the wire until the electric field from them is strong enough to exactly counterbalance the induction force. Seen from outside the generator, the two wires appear as though they had been electrostatically charged to the potential difference V, and as though the charge was being changed with time to give an alternating potential difference. There is also another difference from an electrostatic situation. If we connect the generator to an external circuit that permits passage of a current, we find that the emf does not permit the wires to be discharged but continues 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. Since the emf has a sinusoidal time variation, so also does the current. There is an alternating current

$$I = \frac{\mathcal{E}}{R} = \frac{V_0}{R} \sin \omega t.$$

The schematic diagram of such a circuit is shown in Fig. 17–7.

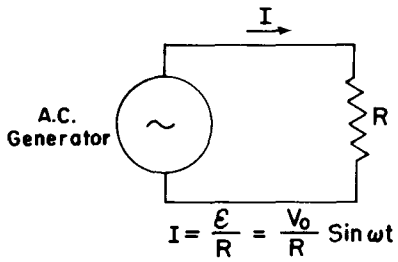


Fig. 17–7. A circuit with an ac generator and a resistance.

We can also see that the emf determines how much energy is supplied by the generator. Each charge in the wire is receiving energy at the rate $\boldsymbol{F} \cdot \boldsymbol{v}$, where $\boldsymbol{F}$ is the force on the charge and $\boldsymbol{v}$ is its velocity. Now let the number of moving charges per unit length of the wire be n; then the power being delivered into any element $d\boldsymbol{s}$ of the wire is

$$\boldsymbol{F} \cdot \boldsymbol{v} n \, d\boldsymbol{s}.$$

For a wire, $\boldsymbol{v}$ is always along $d\boldsymbol{s}$, so we can rewrite the power as

$$nv\boldsymbol{F} \cdot d\boldsymbol{s}.$$

The total power being delivered to the complete circuit is the integral of this expression around the complete loop:

$$\text{Power} = \oint nv\boldsymbol{F} \, d\boldsymbol{s}. \tag{17.15}$$

Now remember that qnv is the current I, and that the emf is defined as the integral of F/q around the circuit. We get the result

$$\text{Power from a generator} = \mathcal{E}I. \tag{17.16}$$

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

$$\tau = ISB \sin \theta. \tag{17.17}$$

The rate at which mechanical work must be done to keep the coil rotating is the angular velocity ω times the torque:

$$\frac{dW}{dt} = \omega\tau = \omega ISB \sin \theta. \tag{17.18}$$

Comparing this equation with Eq. (17.14), we see that the rate of mechanical work required to rotate the coil against the magnetic forces is just equal to $\mathcal{E}I$, 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 another example of the currents and forces due to an induced emf, let's analyze what happens in the setup 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 U (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

$$\mathcal{E} = vBw. \tag{17.19}$$

The current in the circuit is proportional to this emf and inversely proportional to the resistance of the circuit:

$$I = \frac{\mathcal{E}}{R} = \frac{vBw}{R}. \tag{17.20}$$

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

$$F = BIw. \tag{17.21}$$

Taking I from Eq. (17.20), we have for the force

$$F = \frac{B^2w^2}{R} v. \tag{17.22}$$

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-proportional" 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 conductor, even though such situations, in 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 provide 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 count is an indication of the total energy consumption, i.e., the number of watthours used.

We may also point out that Eq. (17.22) shows that the force from induced currents—that is, any eddy-current force—is inversely proportional to the resistance. The force will be larger, the better the conductivity of the material. The reason, of course, is that an emf produces more current if the resistance is low, and the stronger currents represent 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 $\mathcal{E}I$. The rate at which work is done in moving the conducting crossbar is the force on the bar times its velocity. Using Eq. (17.21) for the force, the rate of doing work is

$$\frac{dW}{dt} = \frac{v^2B^2w^2}{R}.$$

We see that this is indeed equal to the product $\mathcal{E}I$ 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 but 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 at each instant 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 conducting wire wound in the form of a long solenoid. Around this coil—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.

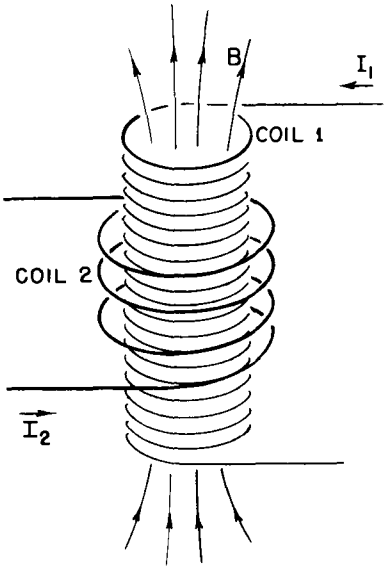


Fig. 17–8. A current in coil 1 produces a magnetic field through coil 2.

We have seen in Section 13–5 that the magnetic field inside a long solenoid is uniform and has the magnitude

$$B = \frac{1}{\epsilon_0 c^2}\frac{N_1 I_1}{l}, \tag{17.23}$$

where N_1 is the number of turns in coil 1, I_1 is the current through it, and l is its length. Let's say that the cross-sectional area of coil 1 is S; then the flux of $\boldsymbol{B}$ is its magnitude times S. If coil 2 has N_2 turns, this flux links the coil N_2 times. Therefore the emf in coil 2 is given by

$$\mathcal{E}_2 = -N_2 S\frac{dB}{dt}. \tag{17.24}$$

The only quantity in Eq. (17.23) which varies with time is I_1. The emf is therefore given by

$$\mathcal{E}_2 = -\frac{N_1 N_2 S}{\epsilon_0 c^2 l}\frac{dI_1}{dt}. \tag{17.25}$$

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 geometric factor of the two coils, is called the *mutual inductance*, and is usually designated $\mathfrak{M}_{21}$. Equation (17.25) is then written

$$\mathcal{E}_2 = \mathfrak{M}_{21}\frac{dI_1}{dt}. \tag{17.26}$$

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

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 dI_2/dt: We can write

$$\mathcal{E}_1 = \mathfrak{M}_{12} \frac{dI_2}{dt}. \tag{17.27}$$

The computation of $\mathfrak{M}_{12}$ would be more difficult than the computation we have just done for $\mathfrak{M}_{21}$. We will not carry through that computation now, because we will show later in this chapter that $\mathfrak{M}_{12}$ is necessarily equal to $\mathfrak{M}_{21}$.

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 $\mathfrak{M}_{21}$ and $\mathfrak{M}_{12}$ would be different. Their values would depend on the shapes of the coils and their relative positions.

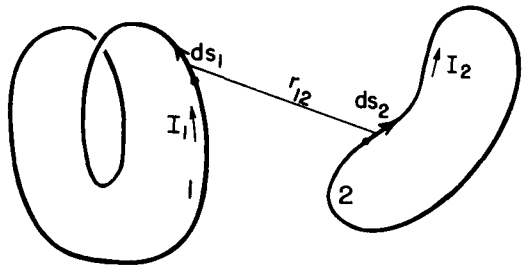


Fig. 17–9. Any two coils have a mutual inductance $\mathfrak{M}$ proportional to the integral of $d\boldsymbol{s}_1 \cdot d\boldsymbol{s}_2/r_{12}$.

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

$$\mathcal{E}_1 = -\frac{d}{dt}\int_{(1)} \boldsymbol{B}\cdot\boldsymbol{n}\,da,$$

where $\boldsymbol{B}$ is the magnetic field and the integral is to be taken over a surface bounded by circuit 1. We have seen in Section 14–1 that such a surface integral of $\boldsymbol{B}$ can be related to a line integral of the vector potential. In particular,

$$\int_{(1)} \boldsymbol{B}\cdot\boldsymbol{n}\,da = \oint_{(1)} \boldsymbol{A}\cdot d\boldsymbol{s}_1,$$

where $\boldsymbol{A}$ represents the vector potential and $d\boldsymbol{s}_1$ is an element of circuit 1. The line integral is to be taken around circuit 1. The emf in coil 1 can therefore be written as

$$\mathcal{E}_1 = -\frac{d}{dt}\oint_{(1)} \boldsymbol{A}\cdot d\boldsymbol{s}_1. \tag{17.28}$$

Now let's assume that the vector potential at circuit 1 comes from currents in circuit 2. Then it can be written as a line integral around circuit 2:

$$\boldsymbol{A} = \frac{1}{4\pi\epsilon_0 c^2}\oint_{(2)} \frac{I_2\,d\boldsymbol{s}_2}{r_{12}}, \tag{17.29}$$

where I_2 is the current in circuit 2, and r_{12} is the distance from the element of the circuit $d\boldsymbol{s}_2$ to the point on circuit 1 at which we are evaluating the vector potential. (See Fig. 17–9.) Combining Eqs. (17.28) and (17.29), we can express the emf in circuit 1 as a double line integral:

$$\mathcal{E}_1 = -\frac{1}{4\pi\epsilon_0 c^2}\frac{d}{dt}\oint_{(1)}\oint_{(2)} \frac{I_2\,d\boldsymbol{s}_2}{r_{12}}\cdot d\boldsymbol{s}_1.$$

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

integration. We may therefore take it out of the integrals. The emf can then be written as

$$\mathcal{E}_1 = \mathfrak{M}_{12} \frac{dI_2}{dt},$$

where the coefficient $\mathfrak{M}_{12}$ is

$$\mathfrak{M}_{12} = -\frac{1}{4\pi\epsilon_0 c^2} \oint_{(1)} \oint_{(2)} \frac{d\mathbf{s}_2 \cdot d\mathbf{s}_1}{r_{12}}. \tag{17.30}$$

We see from this integral that $\mathfrak{M}_{12}$ depends only on the circuit geometry. It depends on a kind of average separation of the two circuits, with the average weighted most for 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 $\mathfrak{M}_{12}$ is identical to the integral for $\mathfrak{M}_{21}$. We have therefore shown that the two coefficients are identical. For a system with only two coils, the coefficients $\mathfrak{M}_{12}$ and $\mathfrak{M}_{21}$ are often represented by the symbol $\mathfrak{M}$ without subscripts, called simply the *mutual inductance*:

$$\mathfrak{M}_{12} = \mathfrak{M}_{21} = \mathfrak{M}.$$

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 magnetic flux linking either coil will be the sum of the two fluxes which would exist separately, because the law of superposition applies for magnetic fields. The emf in either coil 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*

$$\mathcal{E}_2 = \mathfrak{M}_{21} \frac{dI_1}{dt} + \mathfrak{M}_{22} \frac{dI_2}{dt}. \tag{17.31}$$

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:

$$\mathcal{E}_1 = \mathfrak{M}_{12} \frac{dI_2}{dt} + \mathfrak{M}_{11} \frac{dI_1}{dt}. \tag{17.32}$$

The coefficients $\mathfrak{M}_{22}$ and $\mathfrak{M}_{11}$ are always negative numbers. It is usual to write

$$\mathfrak{M}_{11} = -\mathfrak{L}_1, \qquad \mathfrak{M}_{22} = -\mathfrak{L}_2, \tag{17.33}$$

where $\mathfrak{L}_1$ and $\mathfrak{L}_2$ 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 $\mathfrak{L}$. 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

$$\mathcal{E} = -\mathfrak{L} \frac{dI}{dt}. \tag{17.34}$$

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 current in the coil has a kind of inertia. In fact, if we wish to change the current in

---

* The sign of $\mathfrak{M}_{12}$ and $\mathfrak{M}_{21}$ in Eqs. (17.31) and (17.32) depends on the arbitrary choices for the sense of a positive current in the two coils.

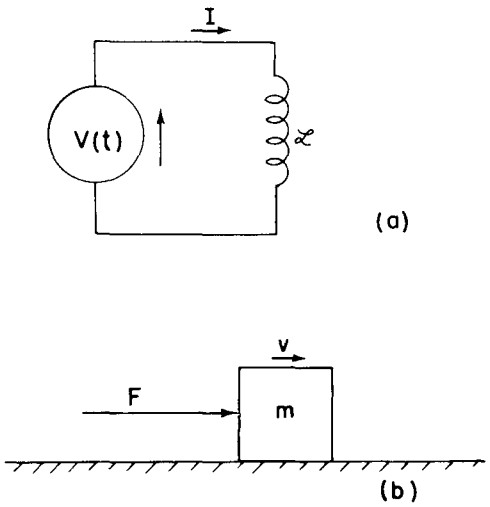


Fig. 17–10 (a) A circuit with a voltage source and an inductance. (b) An analogous mechanical system.

a coil we must overcome this inertia by 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 $\mathcal{V}$ according to the relation

$$\mathcal{V} = \mathcal{L}\frac{dI}{dt}. \tag{17.35}$$

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 equations have the same solutions." Thus, if we make the externally applied voltage $\mathcal{V}$ correspond to an externally applied force F, and the current I in a coil correspond to the velocity v of a particle, the inductance $\mathcal{L}$ 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.

Particle	*Coil*
F (force)	$\mathcal{V}$ (potential difference)
v (velocity)	I (current)
x (displacement)	q (charge)
$F = m\dfrac{dv}{dt}$	$\mathcal{V} = \mathcal{L}\dfrac{dI}{dt}$
mv (momentum)	$\mathcal{L}I$
$\frac{1}{2}mv^2$ (kinetic energy)	$\frac{1}{2}\mathcal{L}I^2$ (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 should be an analogous quantity equal to $\mathcal{L}I$, whose rate of change is $\mathcal{V}$. We have no right, of course, to say that $\mathcal{L}I$ is the real momentum of the circuit; in fact, it isn't. The whole circuit may be standing still and have no momentum. It is only that $\mathcal{L}I$ is analogous to the momentum mv in the sense of satisfying corresponding equations. In the same way, to the kinetic energy $\frac{1}{2}mv^2$, there corresponds an analogous quantity $\frac{1}{2}\mathcal{L}I^2$. But there we have a surprise. This $\frac{1}{2}\mathcal{L}I^2$ is really the energy in the electrical case also. This is because the rate of doing work on the inductance is $\mathcal{V}I$, and in the mechanical system it is Fv, the corresponding quantity. Therefore, in the case of the energy, the quantities not only correspond mathematically, but also have the same physical meaning as well.

We may see this in more detail as follows. As we found in Eq. (17.16), the rate of electrical work by induced forces is the product of the electromotive force and the current:

$$\frac{dW}{dt} = \mathcal{E}I.$$

Replacing $\mathcal{E}$ by its expression in terms of the current from Eq. (17.34), we have

$$\frac{dW}{dt} = -\mathcal{L}I\frac{dI}{dt}. \tag{17.36}$$

Integrating this equation, we find that the energy required from an external source to overcome the emf in the self-inductance while building up the current† (which must equal the energy stored, U) is

$$-W = U = \tfrac{1}{2}\mathcal{L}I^2 \tag{17.37}$$

Therefore the energy stored in an inductance is $\frac{1}{2}\mathcal{L}I^2$.

* This is, incidentally, *not* the *only* way a correspondence can be set up between mechanical and electrical quantities.

† 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 but do not change the energy which goes into the inductance.

Applying the same arguments to a pair of coils such as those in Figs. 17–8 or 17–9, we can show that the total electrical energy of the system is given by

$$U = \tfrac{1}{2}\mathcal{L}_1 I_1^2 + \tfrac{1}{2}\mathcal{L}_2 I_2^2 + \mathfrak{M} I_1 I_2. \tag{17.38}$$

For, starting with $I = 0$ in both coils, we could first turn on the current I_1 in coil 1, with $I_2 = 0$. The work done is just $\frac{1}{2}\mathcal{L}_1 I_1^2$. But now, on turning up I_2, we not only do the work $\frac{1}{2}\mathcal{L}_2 I_2^2$ against the emf in circuit 2, but also an additional amount $\mathfrak{M} I_1 I_2$, which is the integral of the emf $[\mathfrak{M}(dI_2/dt)]$ in circuit 1 times the now *constant* current I_1 in that circuit.

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 change in the energy of Eq. (17.38). We must remember, of course, that as we change the relative positions of the coils the only quantity which varies is the mutual inductance $\mathfrak{M}$. We might then write the equation of virtual work as

$$-F\,\Delta x = \Delta U = I_1 I_2\,\Delta\mathfrak{M} \text{ (wrong)}.$$

But this equation is wrong because, as we have seen earlier, it includes 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 emf's 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_{mech}, the "mechanical energy." We can therefore write for the force

$$-F\,\Delta x = \Delta U_{\text{mech}} = -\Delta U. \tag{17.39}$$

The force between two coils is then given by

$$F\,\Delta x = I_1 I_2\,\Delta\mathfrak{M}.$$

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 $\mathfrak{M}$ and the self-inductances $\mathcal{L}_1$ and $\mathcal{L}_2$ 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 some 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 to happen! Now our energy equation, Eq. (17.38), can equally well be written in the following form:

$$U = \frac{1}{2}\,\mathcal{L}_1\left(I_1 + \frac{\mathfrak{M}}{\mathcal{L}_1} I_2\right)^2 + \frac{1}{2}\left(\mathcal{L}_2 - \frac{\mathfrak{M}^2}{\mathcal{L}_1}\right) I_2^2. \tag{17.40}$$

That is just an algebraic transformation. 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

$$I_2 = -\frac{\mathcal{L}_1}{\mathfrak{M}} I_1 \tag{17.41}$$

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 have the requirement that

$$\mathcal{L}_1\mathcal{L}_2 > \mathfrak{M}^2.$$

We have thus proved the general result that the magnitude of the mutual inductance $\mathfrak{M}$ of any two coils is necessarily less than or equal to the geometric mean of the two self-inductances. ($\mathfrak{M}$ itself may be positive or negative, depending on the sign

conventions for the currents I_1 and I_2.)

$$|\mathfrak{M}| < \sqrt{\mathfrak{L}_1\mathfrak{L}_2}. \tag{17.42}$$

The relation between $\mathfrak{M}$ and the self-inductances is usually written as

$$\mathfrak{M} = k\sqrt{\mathfrak{L}_1\mathfrak{L}_2}. \tag{17.43}$$

The constant k 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 "tightly 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 have given in Eq. (17.30) 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 doesn't 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 size 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 15–3, an expression for the magnetic energy of a distribution of stationary currents:

$$U = \tfrac{1}{2}\int \boldsymbol{j}\cdot\boldsymbol{A}\,dV. \tag{17.44}$$

If we know the distribution of current density $\boldsymbol{j}$, we can compute the vector potential $\boldsymbol{A}$ and then evaluate the integral of Eq. (17.44) to get the energy. This energy is equal to the magnetic energy of the self-inductance, $\frac{1}{2}\mathfrak{L}I^2$. Equating the two gives us a formula for the inductance:

$$\mathfrak{L} = \frac{1}{I^2}\int \boldsymbol{j}\cdot\boldsymbol{A}\,dV. \tag{17.45}$$

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. (17.45) 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 $\boldsymbol{j}$ and again through the vector potential $\boldsymbol{A}$. The integral divided by I^2 will depend on the geometry of the circuit but not on the current I.

Equation (17.44) 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 is a form that is important because it is more generally valid. In the energy equation, Eq. (17.44), both $\boldsymbol{A}$ and $\boldsymbol{j}$ can be related to $\boldsymbol{B}$, so we can hope to express the energy in terms of the magnetic field—just as we were able to relate the electrostatic energy to the electric field. We begin by replacing $\boldsymbol{j}$ by $\epsilon_0 c^2\nabla\times\boldsymbol{B}$. We cannot replace $\boldsymbol{A}$ so easily, since $\boldsymbol{B} = \nabla\times\boldsymbol{A}$ cannot be reversed to give $\boldsymbol{A}$ in terms of $\boldsymbol{B}$. Anyway, we can write

$$U = \frac{\epsilon_0 c^2}{2}\int (\nabla\times\boldsymbol{B})\cdot\boldsymbol{A}\,dV. \tag{17.46}$$

The interesting thing is that—with some restrictions—this integral can be written as

$$U = \frac{\epsilon_0 c^2}{2} \int \boldsymbol{B} \cdot (\boldsymbol{\nabla} \times \boldsymbol{A})\, dV. \tag{17.47}$$

To see this, we write out in detail a typical term. Suppose that we take the term $(\boldsymbol{\nabla} \times \boldsymbol{B})_z A_z$ which occurs in the integral of Eq. (17.46). Writing out the components, we get

$$\int \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}\right) A_z\, dx\, dy\, dz.$$

(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

$$\int \frac{\partial B_y}{\partial x} A_z\, dx = B_y A_z - \int B_y \frac{\partial A_z}{\partial x}\, dx.$$

Now suppose that our system—meaning 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 $B_y A_z$ at the limits will give zero. We have left only the term with $B_y(\partial A_z/\partial x)$, which is evidently one part of $B_y(\boldsymbol{\nabla} \times \boldsymbol{A})_y$ and, therefore, of $\boldsymbol{B} \cdot (\boldsymbol{\nabla} \times \boldsymbol{A})$. If you work out the other five terms, you will see that Eq. (17.47) is indeed equivalent to Eq. (17.46).

But now we can replace $(\boldsymbol{\nabla} \times \boldsymbol{A})$ by $\boldsymbol{B}$, to get

$$U = \frac{\epsilon_0 c^2}{2} \int \boldsymbol{B} \cdot \boldsymbol{B}\, dV. \tag{17.48}$$

We have expressed the energy of a magnetostatic situation in terms of the magnetic field only. The expression corresponds closely to the formula we found for the electrostatic energy:

$$U = \frac{\epsilon_0}{2} \int \boldsymbol{E} \cdot \boldsymbol{E}\, dV. \tag{17.49}$$

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 $\boldsymbol{E}$ and $\boldsymbol{B}$ 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 $\boldsymbol{B}$ of a single coil, we can find the self-inductance by equating the energy expression (17.48) to $\frac{1}{2}\mathcal{L}I^2$. 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 $\boldsymbol{B}$ outside is zero. The magnitude of the field inside is $B = nI/\epsilon_0 c^2$, where n is the number of turns per unit 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

$$U = \frac{\epsilon_0 c^2}{2} B^2 \cdot (\text{Vol}) = \frac{n^2 I^2}{2\epsilon_0 c^2} \pi r^2 L,$$

which is equal to $\frac{1}{2}\mathcal{L}I^2$. Or,

$$\mathcal{L} = \frac{\pi r^2 n^2}{\epsilon_0 c^2} L. \tag{17.50}$$

18

The Maxwell Equations

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

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 electromagnetic 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 fixed charges. Although we have been very careful to point out the restrictions whenever we wrote an equation, it is easy to forget all of the qualifications and to learn too 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 $\boldsymbol{E}$ 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 $\boldsymbol{E}$ 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 $\boldsymbol{B}$ through any closed surface is always zero. The second equation, that the curl of $\boldsymbol{E}$ is $-\partial\boldsymbol{B}/\partial t$, is Faraday's law and was discussed in the last two chapters. It also is generally true. The last equation has 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 $\boldsymbol{B}$ is $\boldsymbol{j}/\epsilon_0 c^2$, but the correct 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

$$\nabla \times \boldsymbol{B} = \frac{\boldsymbol{j}}{\epsilon_0 c^2}. \tag{18.1}$$

Maxwell began by considering 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 $\boldsymbol{j}$ also be zero. But if the divergence of $\boldsymbol{j}$ is zero, then the total flux of current out of any closed surface is also zero.

The flux of current from a closed surface is the decrease of the charge inside the surface. This certainly cannot in general be zero because we know that the charges can be moved from one place to another. The equation

$$\nabla \cdot \boldsymbol{j} = -\frac{\partial \rho}{\partial t} \tag{18.2}$$

has, in fact, been almost our definition of $\boldsymbol{j}$. This equation expresses the very funda-

Table 18–1 Classical Physics

Maxwell's equations

I. $\nabla \cdot \boldsymbol{E} = \dfrac{\rho}{\epsilon_0}$ (Flux of $\boldsymbol{E}$ through a closed surface) = (Charge inside)/ϵ_0

II. $\nabla \times \boldsymbol{E} = -\dfrac{\partial \boldsymbol{B}}{\partial t}$ (Line integral of $\boldsymbol{E}$ around a loop) = $-\dfrac{d}{dt}$ (Flux of $\boldsymbol{B}$ through the loop)

III. $\nabla \cdot \boldsymbol{B} = 0$ (Flux of $\boldsymbol{B}$ through a closed surface) = 0

IV. $c^2 \nabla \times \boldsymbol{B} = \dfrac{\boldsymbol{j}}{\epsilon_0} + \dfrac{\partial \boldsymbol{E}}{\partial t}$ c^2(Integral of $\boldsymbol{B}$ around a loop) = (Current through the loop)/ϵ_0 $+\dfrac{\partial}{\partial t}$ (Flux of $\boldsymbol{E}$ through the loop)

[**Conservation of charge**

$\nabla \cdot \boldsymbol{j} = -\dfrac{\partial \rho}{\partial t}$ (Flux of current through a closed surface) = $-\dfrac{\partial}{\partial t}$ (Charge inside)]

Force law

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$$

Law of motion

$\dfrac{d}{dt}(\boldsymbol{p}) = \boldsymbol{F}$, where $\boldsymbol{p} = \dfrac{m\boldsymbol{v}}{\sqrt{1 - v^2/c^2}}$ (Newton's law, with Einstein's modification)

Gravitation

$$\boldsymbol{F} = -G\frac{m_1 m_2}{r^2}\boldsymbol{e}_r$$

mental law that electric charge is conserved—any flow of charge must come from some supply. Maxwell appreciated this difficulty and proposed that it could be avoided by adding the term $\partial \boldsymbol{E}/\partial t$ to the right-hand side of Eq. (18.1); he then got the fourth equation in Table 18–1:

$$\text{IV.} \qquad c^2 \nabla \times \boldsymbol{B} = \frac{\boldsymbol{j}}{\epsilon_0} + \frac{\partial \boldsymbol{E}}{\partial t}.$$

It was not 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 solid. He also tried to explain the meaning of his new equation 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 counts are 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 untold 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 required to straighten out the difficulty 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:

$$\nabla \cdot \frac{\boldsymbol{j}}{\epsilon_0} + \nabla \cdot \frac{\partial \boldsymbol{E}}{\partial t} = 0. \tag{18.3}$$

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

$$\nabla \cdot \boldsymbol{j} + \epsilon_0 \frac{\partial}{\partial t} \nabla \cdot \boldsymbol{E} = 0. \tag{18.4}$$

But the first of Maxwell's equations says that the divergence of $\boldsymbol{E}$ 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 $\boldsymbol{E}$, we found that a whole new class of phenomena was described. We shall see that Maxwell's little addition to the equation for $\nabla \times \boldsymbol{B}$ 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. Suppose we imagine a little sphere with radioactive material on it. This radioactive material is squirting 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

$$\frac{\partial Q(r)}{\partial t} = -4\pi r^2 j(r). \tag{18.5}$$

We now ask about the magnetic field produced by the currents in this situation. Suppose we draw some loop Γ 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.

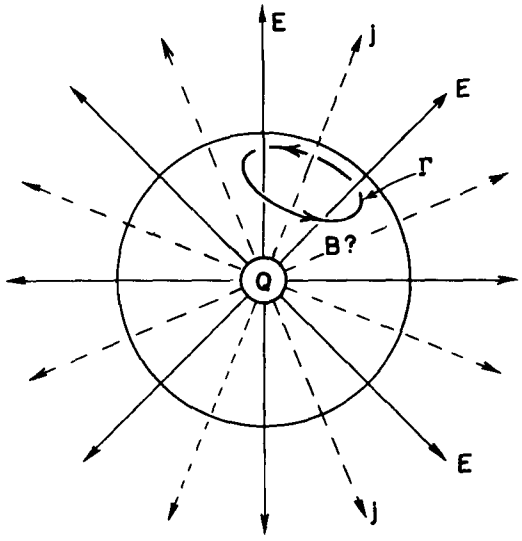


Fig. 18–1. What is the magnetic field of a spherically symmetric current?

But we are already in difficulty. How can the $\boldsymbol{B}$ have any particular direction on the sphere? A different choice of Γ would allow us to conclude that its direction is exactly opposite to that shown. So how *can* there be any circulation of $\boldsymbol{B}$ around the currents?

We are saved by Maxwell's equation. The circulation of $\boldsymbol{B}$ depends not only on the total *current* through Γ but also on the rate of change with time of the *electric 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 symmetrically distributed, as we assume. It is radial, and its rate of change is then

$$\frac{\partial E}{\partial t} = \frac{1}{4\pi\epsilon_0 r^2} \frac{\partial Q}{\partial t}. \tag{18.6}$$

Comparing this with Eq. (18.5), we see that at any radius

$$\frac{\partial E}{\partial t} = -\frac{j}{\epsilon_0}. \tag{18.7}$$

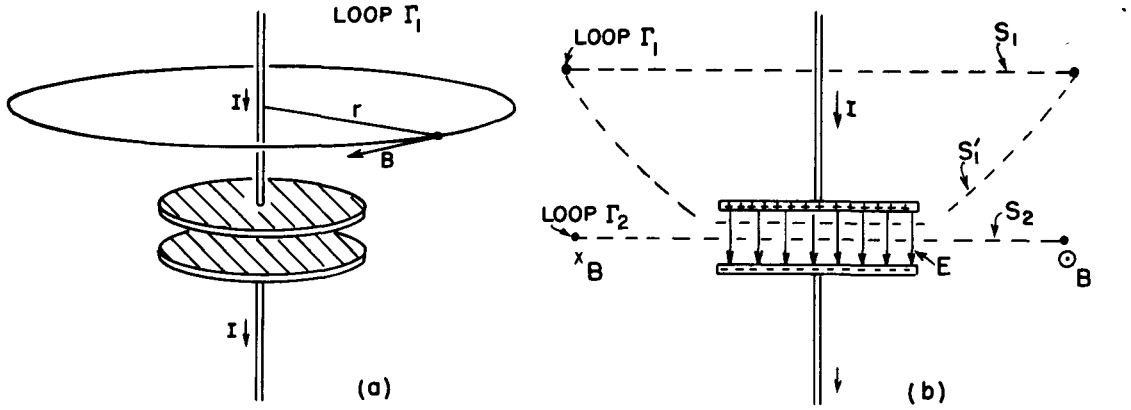


Fig. 18–2. The magnetic field near a charging capacitor.

In Eq. IV the two source terms cancel and the curl of $\boldsymbol{B}$ 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 $\epsilon_0 c^2$. We have

$$2\pi rB = \frac{I}{\epsilon_0 c^2}. \tag{18.8}$$

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 inside the circle, there are no electric fields on it (assuming the wire to be a very good conductor). The surface integral of $\partial \boldsymbol{E}/\partial t$ 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 $\boldsymbol{B}$ around Γ_2 is $2\pi rB$. This must equal the time derivative of the flux of $\boldsymbol{E}$ through the plane circular surface S_2. This flux of $\boldsymbol{E}$, we know from Gauss' law, must be equal to $1/\epsilon_0$ times the charge Q on one of the condenser plates. We have

$$c^2\, 2\pi rB = \frac{d}{dt}\left(\frac{Q}{\epsilon_0}\right). \tag{18.9}$$

That is very convenient. It is the same result we found in Eq. (18.8). 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_1' 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_1' there is no current, but an electric flux changing at the rate I/ϵ_0. The same $\boldsymbol{B}$ is obtained if we use Eq. IV 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. IV *by itself*, nothing particularly new comes out. The words "*by itself*" are, however, all-important. Maxwell's small change in Eq. IV, 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 what they do to charges. Knowing $\boldsymbol{E}$ and $\boldsymbol{B}$, 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—so 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. 18–3. We have a sheet of charge located on the yz-plane. The sheet is first at rest, then instantaneously given a velocity u in the y-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 u very quickly. So we have suddenly a surface current J (J is the current per unit

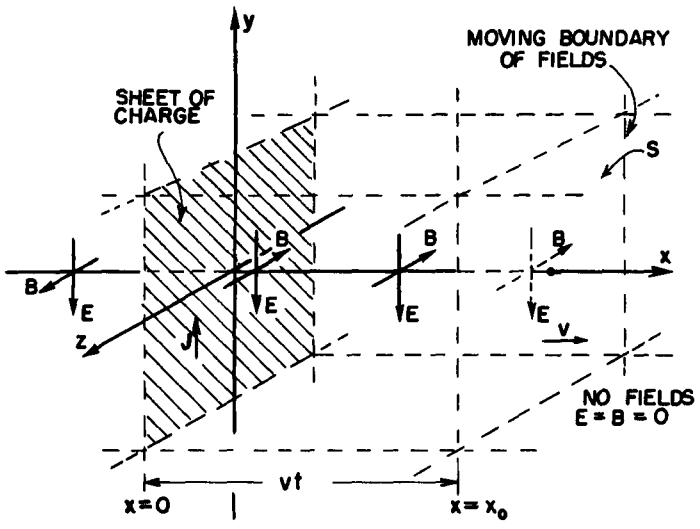


Fig. 18–3. An infinite sheet of charge is suddenly set into motion parallel to itself. There are magnetic and electric fields that propagate out from the sheet at a constant speed.

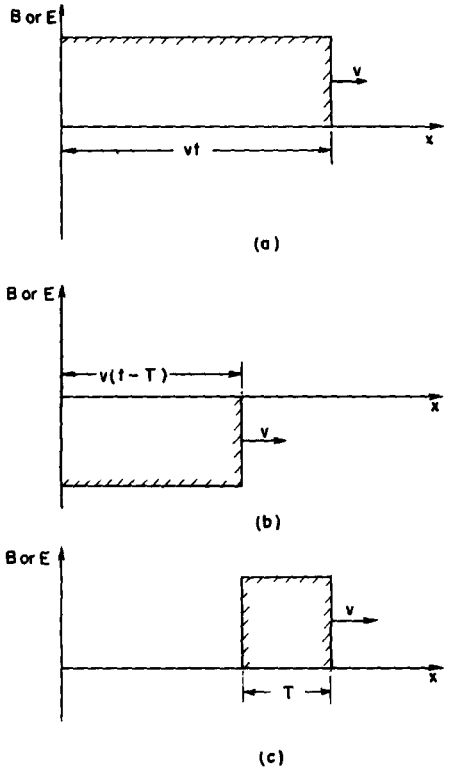


Fig. 18–4. (a) The magnitude of B (or E) as a function of x at the time t after the charge sheet is set in motion. (b) The fields for a charge sheet set in motion, toward negative y at $t = T$. (c) The sum of (a) and (b).

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 on the yz-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 $\pm y$ and $\pm z$. 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 $\boldsymbol{B}$ by using the fact that the line integral of the magnetic field will be equal to the current over $\epsilon_0 c^2$. We would get that $B = J/2\epsilon_0 c^2$ (since the current I in a strip of width w is Jw and the line integral of $\boldsymbol{B}$ is $2Bw$).

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 farther 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 moved 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 $\partial \boldsymbol{E}/\partial t$ 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 intermixing, 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 $\boldsymbol{B}$ 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 room for any electric flux to go through it. But the field $\boldsymbol{B}$ out farther—for larger 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 $\boldsymbol{B}$; then the turning on of $\boldsymbol{B}$ 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 plus and minus x-directions.

The $\boldsymbol{E}$-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 $\boldsymbol{E}$ and $\boldsymbol{B}$ are uniform out to the distance $x = vt$, 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 $\boldsymbol{E}$ or $\boldsymbol{B}$ 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 $x = \pm vt$ 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 $\boldsymbol{E}$ and $\boldsymbol{B}$, 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 quantitatively 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 z-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 $x = \pm vt$.

Let's see if these fields are consistent with Maxwell's equations. Let's first draw one of those loops that we use to calculate a line integral, say the rectangle Γ_2 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 magnetic flux through this loop. If it is changing, there should be an emf around it. If the wavefront is moving, we will have a changing magnetic

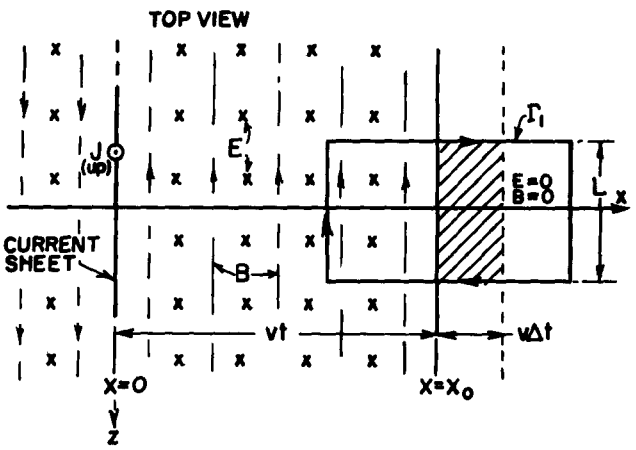


Fig. 18–5. Top view of Fig. 18–3.

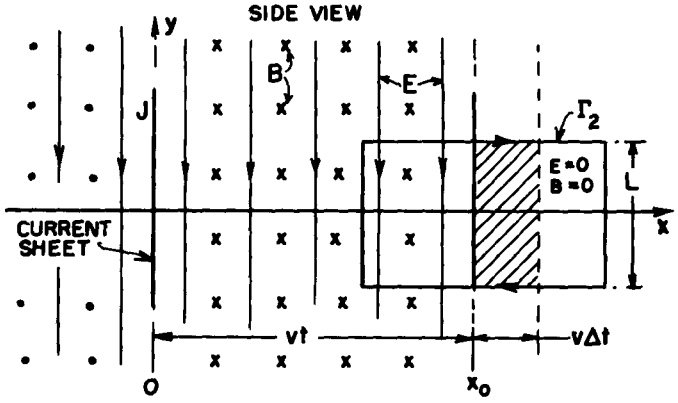


Fig. 18–6. Side view of Fig. 18–3.

flux, because the area in which $\boldsymbol{B}$ exists is progressively increasing at the velocity v. The flux inside Γ_2 is B times the part of the area inside Γ_2 which has a magnetic field. The rate of change of the flux, since the magnitude of $\boldsymbol{B}$ 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 Γ_2 is L, the area in which $\boldsymbol{B}$ exists changes by $Lv\,\Delta t$ in the time Δt. (See Fig. 18–6.) The rate of change of flux is then BLv. According to Faraday's law, this should equal the line integral of $\boldsymbol{E}$ around Γ_2, which is just EL. We have the equation

$$E = vB. \tag{18.10}$$

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 $\boldsymbol{E}$ and $\boldsymbol{B}$:

$$c^2\nabla \times \boldsymbol{B} = \frac{\boldsymbol{j}}{\epsilon_0} + \frac{\partial \boldsymbol{E}}{\partial t}. \tag{18.11}$$

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 $\boldsymbol{B}$ next to the current sheet. Also, for any loop drawn outside the sheet but behind the wavefront, there is no curl of $\boldsymbol{B}$ nor any $\boldsymbol{j}$ or changing $\boldsymbol{E}$, so the equation is correct there. Now let's look at what happens for the curve Γ_1 that intersects the wavefront, as shown in Fig. 18–5. Here there are no currents, so Eq. (18.11) can be written—in integral form—as

$$c^2 \oint_{\Gamma_1} \boldsymbol{B}\cdot d\boldsymbol{s} = \frac{d}{dt} \int_{\text{inside } \Gamma_1} \boldsymbol{E}\cdot \boldsymbol{n}\, da. \tag{18.12}$$

The line integral of $\boldsymbol{B}$ is just B times L. The rate of change of the flux of $\boldsymbol{E}$ is due only to the advancing wavefront. The area inside Γ_1, where $\boldsymbol{E}$ is not zero, is increasing at the rate vL. The right-hand side of Eq. (18.12) is then vLE. That equation becomes

$$c^2B = Ev. \tag{18.13}$$

We have a solution in which we have a constant $\boldsymbol{B}$ and a constant $\boldsymbol{E}$ 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),

$$E = vB, \qquad \text{and} \qquad E = \frac{c^2}{v}B.$$

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 an example in which the electrical influence from a current propagates at a certain finite velocity c.

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 know 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 wavefront again travels out at the velocity c. At the time t it has reached the distance $x = \pm c\,(t - T)$, as shown in Fig. 18–4(b). So we have two "blocks" of field marching 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 $x = c(t - T)$ 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 sheet 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 maintain itself? The answer is: by the combined effects of the Faraday law, $\nabla \times \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t$, and the new term of Maxwell, $c^2 \nabla \times \boldsymbol{B} = \partial \boldsymbol{E}/\partial t$. 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 tries to go away, the changing electric field would create a magnetic field back again. So by a perpetual interplay—by the swishing back and forth from one field to the other—they must go on forever. It is impossible for them to disappear.* They maintain themselves in a kind of a dance—one making the other, the second making the first—propagating onward through space.

18–5 The speed of light

We have a wave which leaves the material 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:

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0} \tag{18.14}$$

and

$$\nabla \times \boldsymbol{B} = \frac{\boldsymbol{j}}{\epsilon_0 c^2}. \tag{18.15}$$

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 $\epsilon_0 c^2$ 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 31, 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 $\epsilon_0 c^2/\epsilon_0$ 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 undulations 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 this 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 $\boldsymbol{E}$ and $\boldsymbol{B}$ 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 $\boldsymbol{B}$ is perpendicular to $\boldsymbol{E}$, and that $E = cB$—are generally true for any electromagnetic wave. Our 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 come out simpler. Although by this time you are thoroughly used to each of the Maxwell equations, there are many pieces that must all be put together. That's what we want to do.

We begin with $\nabla \cdot \boldsymbol{B} = 0$—the simplest of the equations. We know that it implies that $\boldsymbol{B}$ is the curl of something. So, if we write

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}, \tag{18.16}$$

we have already solved one of Maxwell's equations. (Incidentally, you appreciate that it remains true that another vector $\boldsymbol{A}'$ would be just as good if $\boldsymbol{A}' = \boldsymbol{A} + \nabla\psi$ —where ψ is any scalar field—because the curl of $\nabla\psi$ is zero, and $\boldsymbol{B}$ is still the same. We have talked about that before.)

We take next the Faraday law, $\nabla \times \boldsymbol{E} = -\partial\boldsymbol{B}/\partial t$, because it doesn't involve any currents or charges. If we write $\boldsymbol{B}$ as $\nabla \times \boldsymbol{A}$ and differentiate with respect to t, we can write Faraday's law in the form

$$\nabla \times \boldsymbol{E} = -\frac{\partial}{\partial t} \nabla \times \boldsymbol{A}.$$

Since we can differentiate either with respect to time or to space first, we can also write this equation as

$$\nabla \times \left(\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t}\right) = 0. \tag{18.17}$$

We see that $\boldsymbol{E} + \partial\boldsymbol{A}/\partial t$ is a vector whose curl is equal to zero. Therefore that vector is the gradient of something. When we worked on electrostatics, we had $\nabla \times \boldsymbol{E} = 0$, and then we decided that $\boldsymbol{E}$ 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 $\boldsymbol{E} + \partial\boldsymbol{A}/\partial t$; we set

$$\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} = -\nabla\phi. \tag{18.18}$$

We use the same symbol ϕ so that, in the electrostatic case where nothing changes with time and the $\partial\boldsymbol{A}/\partial t$ term disappears, $\boldsymbol{E}$ will be our old $-\nabla\phi$. So Faraday's equation can be put in the form

$$\boldsymbol{E} = -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{18.19}$$

We have solved two of Maxwell's equations already, and we have found that to describe the electromagnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$, we need four potential functions: a scalar potential ϕ and a vector potential $\boldsymbol{A}$, which is, of course, three functions.

Now that $\boldsymbol{A}$ determines part of $\boldsymbol{E}$, as well as $\boldsymbol{B}$, what happens when we change $\boldsymbol{A}$ to $\boldsymbol{A}' = \boldsymbol{A} + \nabla\psi$? In general, $\boldsymbol{E}$ would change if we didn't take some special precaution. We can, however, still allow $\boldsymbol{A}$ to be changed in this way without affecting the fields $\boldsymbol{E}$ and $\boldsymbol{B}$—that is, without changing the physics—if we always change $\boldsymbol{A}$ and ϕ *together* by the rules

$$\boldsymbol{A}' = \boldsymbol{A} + \nabla\psi, \qquad \phi' = \phi - \frac{\partial\psi}{\partial t}. \tag{18.20}$$

Then neither $\boldsymbol{B}$ nor $\boldsymbol{E}$, obtained from Eq. (18.19), is changed.

Previously, we chose to make $\nabla \cdot \boldsymbol{A} = 0$, 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 $\boldsymbol{j}$. Once we can determine $\boldsymbol{A}$ and ϕ from the currents and charges, we can always get $\boldsymbol{E}$ and $\boldsymbol{B}$ 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 $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$; we get

$$\nabla \cdot \left(-\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t}\right) = \frac{\rho}{\epsilon_0},$$

which we can write also as

$$-\nabla^2\phi - \frac{\partial}{\partial t}\nabla \cdot \boldsymbol{A} = \frac{\rho}{\epsilon_0}. \tag{18.21}$$

This is one equation relating ϕ and $\boldsymbol{A}$ to the sources.

Our final equation will be the most complicated. We start by rewriting the fourth Maxwell equation as

$$c^2\nabla \times \boldsymbol{B} - \frac{\partial \boldsymbol{E}}{\partial t} = \frac{\boldsymbol{j}}{\epsilon_0},$$

and then substitute for $\boldsymbol{B}$ and $\boldsymbol{E}$ in terms of the potentials, using Eqs. (18.16) and (18.19):

$$c^2\nabla \times (\nabla \times \boldsymbol{A}) - \frac{\partial}{\partial t}\left(-\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t}\right) = \frac{\boldsymbol{j}}{\epsilon_0}.$$

The first term can be rewritten using the algebraic identity: $\nabla \times (\nabla \times \boldsymbol{A}) = \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2\boldsymbol{A}$; we get

$$-c^2\nabla^2\boldsymbol{A} + c^2\nabla(\nabla \cdot \boldsymbol{A}) + \frac{\partial}{\partial t}\nabla\phi + \frac{\partial^2 \boldsymbol{A}}{\partial t^2} = \frac{\boldsymbol{j}}{\epsilon_0}. \tag{18.22}$$

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*

$$\nabla \cdot A = -\frac{1}{c^2}\frac{\partial \phi}{\partial t}. \tag{18.23}$$

When we do that, the two middle terms in A and ϕ in Eq. (18.22) cancel, and that equation becomes much simpler:

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{j}{\epsilon_0 c^2}. \tag{18.24}$$

And our equation for ϕ—Eq. (18.21)—takes on the same form:

$$\nabla^2 \phi - \frac{1}{c^2}\frac{\partial^2 \phi}{\partial t^2} = -\frac{\rho}{\epsilon_0}. \tag{18.25}$$

What a beautiful set of equations! They are beautiful, first, because they are nicely separated—with the charge density, goes ϕ; with the current, goes 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

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2 \phi}{\partial t^2} = -\frac{\rho}{\epsilon_0}. \tag{18.26}$$

It has a nice symmetry in x, y, z, t—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 47 of Vol. I, we had an equation of the form

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 \phi}{\partial t^2},$$

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 onward through free space, as in our example at the beginning of the chapter.

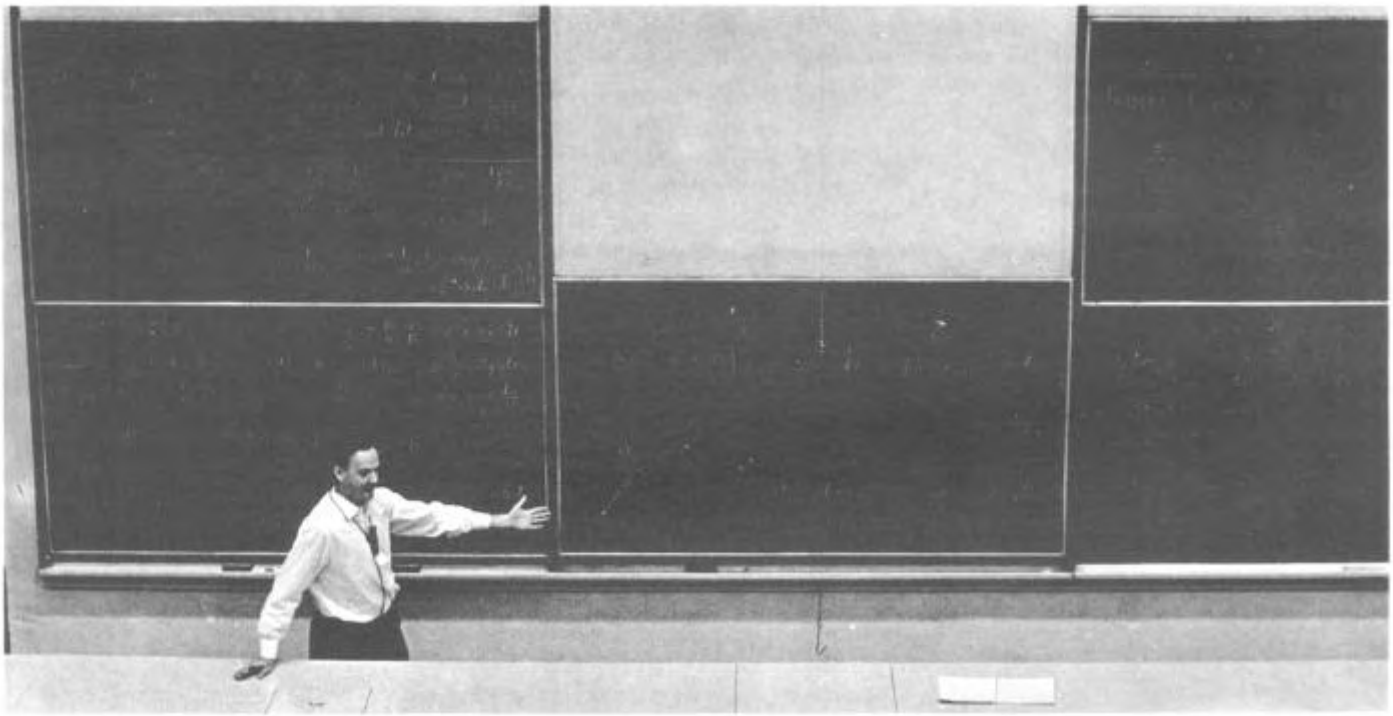
With Maxwell's new term in Eq. IV, we have been able to write the field equations in terms of A and ϕ in a form that is simple and that makes immediately apparent that there are electromagnetic waves. For many practical purposes, it will still be convenient to use the original equations in terms of E 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 $\nabla\psi$ is called a "gauge transformation." Equation (18.23) is called "the Lorentz gauge."

19

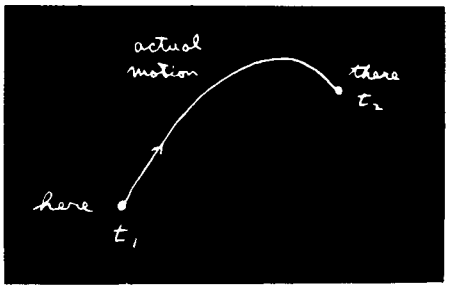
The Principle of Least Action



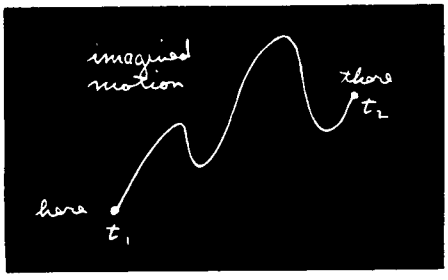
A special lecture—almost verbatim*

"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 throw 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 motion. Suppose that to get from here to there, it went like this

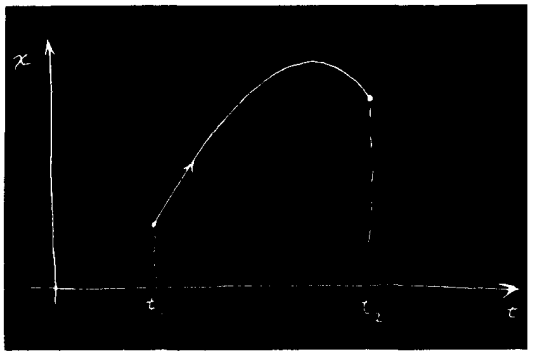


but got there in just the same amount of time. Then he said this: If you calculate the kinetic energy at every moment on the path, take away the potential energy, and integrate it over the time during the whole path, you'll find that the number you'll get is *bigger* than that for the actual motion.

* Later chapters do not depend 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 $F = ma$ 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.

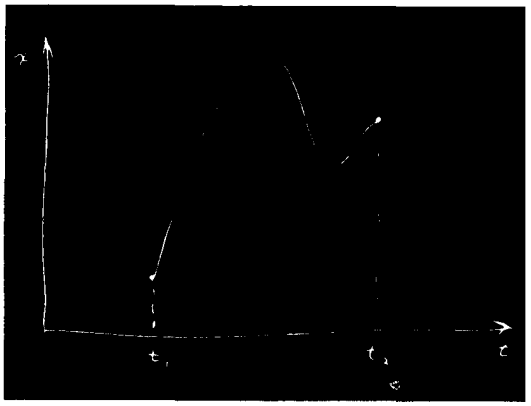
"Let me illustrate a little bit better what it means. If you take the case of the gravitational field, then if the particle has the path $x(t)$ (let's just take one dimension for a moment; we take a trajectory that goes up and down and not sideways), where x is the height above the ground, the kinetic energy is $\frac{1}{2}m\,(dx/dt)^2$, and the potential energy at any time is mgx. 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_1 we started at some height and at the end of the time t_2 we are definitely ending at some other place.



"Then the integral is

$$\int_{t_1}^{t_2}\left[\frac{1}{2}\,m\left(\frac{dx}{dt}\right)^2 - mgx\right]dt.$$

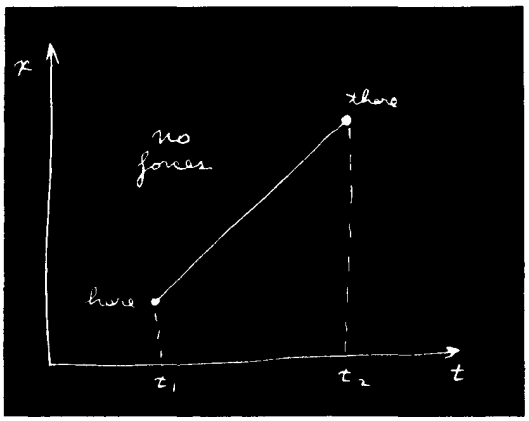
The actual motion is some kind of a curve—it's a parabola if we plot against the time—and gives a certain value for the integral. But we could *imagine* some other motion that went very high and came up and down in some peculiar way.



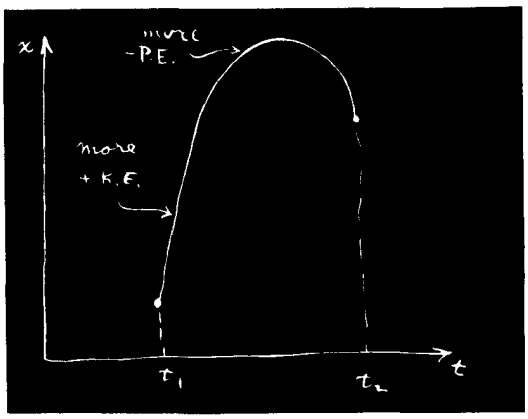
We can calculate the kinetic energy minus the potential energy and integrate for such a path . . . or for any other path we want. The miracle is that the true path is the one for which that integral is least.

"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 a uniform speed. (We know that's the right answer—to go at 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 the car. You can do it several ways: You can accelerate like mad at 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 the 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 if you wobbled your velocity than if you went at a uniform velocity. So we see that the integral is a minimum if the velocity is a constant (when there are no forces). The correct path is like this.



"Now, an object thrown up in a gravitational field does rise faster first and then slow 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. Because the potential energy rises as we go up in space, we will get a lower *difference* if we can get as soon 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 too far, because you will then have too much kinetic energy involved—you have to go very fast to get way 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—trying to get the difference, kinetic 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.

$$\text{Action} = S = \int_{t_1}^{t_2} (\text{KE} - \text{PE})\, dt.$$

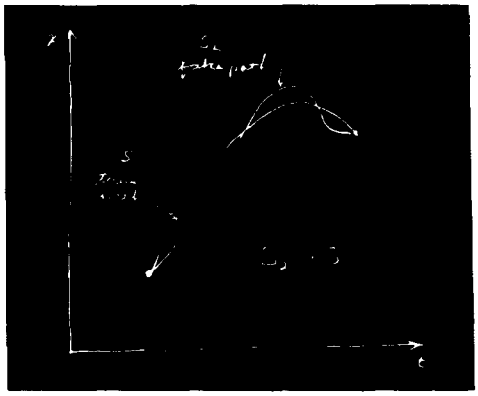
Remember that the PE and KE are both functions of time. For each different possible path you get a different number for this action. Our mathematical problem is to find out for what curve that number is the least.

"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 must 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.



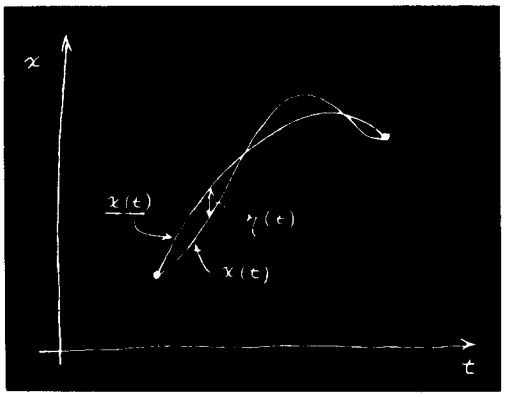
"Problem: Find the true path. Where is it? One way, of course, is to calculate the action for millions and millions of paths and look at which one is lowest. When you find the lowest one, that's the true path.

"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 deviation 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 it this way: We call $\underline{x(t)}$ (with an underline) the true path—the one we are trying to find. We take some trial path $x(t)$ 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 $\underline{S}$ and 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 must 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 h to x and argue that the correction to $f(x)$ in the first order in h must be zero at the minimum. You would substitute $x + h$ for x and expand out to the first order in h . . . just as we are going to do with η.

"The idea is then that we substitute $x(t) = \underline{x(t)} + \eta(t)$ 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 I get this expression:

$$S = \int_{t_1}^{t_2} \left[\frac{m}{2}\left(\frac{d\underline{x}}{dt} + \frac{d\eta}{dt}\right)^2 - V(\underline{x} + \eta)\right] dt.$$

"Now I must write this out in more detail. For the squared term I get

$$\left(\frac{d\underline{x}}{dt}\right)^2 + 2\,\frac{d\underline{x}}{dt}\,\frac{d\eta}{dt} + \left(\frac{d\eta}{dt}\right)^2.$$

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

$$\frac{m}{2}\left(\frac{d\underline{x}}{dt}\right)^2 + m\,\frac{d\underline{x}}{dt}\,\frac{d\eta}{dt} + \text{(second and higher order)}.$$

"Now we need the potential V at $\underline{x} + \eta$. I consider η small, so I can write $V(x)$ as a Taylor series. It is approximately $V(\underline{x})$; 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:

$$V(\underline{x} + \eta) = V(\underline{x}) + \eta V'(\underline{x}) + \frac{\eta^2}{2} V''(\underline{x}) + \cdots$$

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,

$$S = \int_{t_1}^{t_2} \left[\frac{m}{2}\left(\frac{d\underline{x}}{dt}\right)^2 - V(\underline{x}) + m\,\frac{d\underline{x}}{dt}\,\frac{d\eta}{dt} - \eta V'(\underline{x}) + \text{(second and higher order)}\right] dt\,.$$

Now if we look carefully at the thing, we see that the first two terms which I have arranged here correspond to the action $\underline{S}$ that I would have calculated with the true path $\underline{x}$. The thing I want to concentrate on is the change in S—the difference between the S and the $\underline{S}$ 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

$$\delta S = \int_{t_1}^{t_2} \left[m \frac{d\underline{x}}{dt} \frac{d\eta}{dt} - \eta V'(\underline{x}) \right] dt.$$

"Now the problem is this: Here is a certain integral. I don't know what the $\underline{x}$ 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 $d\eta/dt$? Well, after all, if η can be anything at all, its derivative is anything also, so you conclude that the coefficient of $d\eta/dt$ 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 $\eta(t)$ must be zero at both t_1 and t_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 out 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 ηf:

$$\frac{d}{dt}(\eta f) = \eta \frac{df}{dt} + f \frac{d\eta}{dt}.$$

The integral you want is over the last term, so

$$\int f \frac{d\eta}{dt}\, dt = \eta f - \int \eta \frac{df}{dt}\, dt.$$

"In our formula for δS, the function f is m times $d\underline{x}/dt$; therefore, I have the following formula for δS.

$$\delta S = m \frac{d\underline{x}}{dt} \eta(t) \Big|_{t_1}^{t_2} - \int_{t_1}^{t_2} \frac{d}{dt}\left(m \frac{d\underline{x}}{dt} \right) \eta(t)\, dt - \int_{t_1}^{t_2} V'(\underline{x})\, \eta(t)\, dt.$$

The first term must be evaluated at the two limits t_1 and t_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

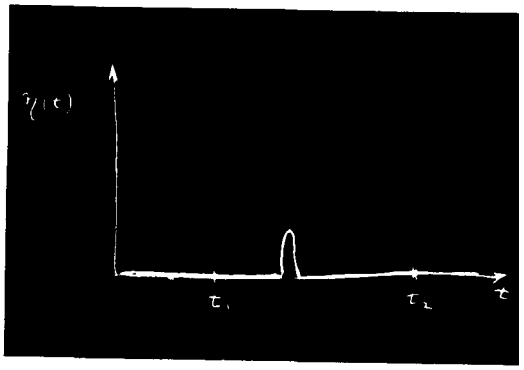
$\eta(t_1) = 0$, and $\eta(t_2) = 0$. So the integrated term is zero. We collect the other terms together and obtain this:

$$\delta S = \int_{t_1}^{t_2} \left[-m \frac{d^2\underline{x}}{dt^2} - V'(\underline{x}) \right] \eta(t)\, dt.$$

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 $F(t)$ 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:

$$\left[-m \frac{d^2\underline{x}}{dt^2} - V'(\underline{x}) \right] = 0.$$

It's not really so complicated; you have seen it before. It is 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 more complicated. For three-dimensional motion, you have to use the complete kinetic energy—$(m/2)$ times the whole velocity squared. That is,

$$\text{KE} = \frac{m}{2}\left[\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 \right].$$

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 paths 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 η only in the x-direction and

say that coefficient must 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. Incidentally, you could use any coordinate system you want, polar or otherwise, and get Newton's laws appropriate to that system right off by seeing what happens if you have the shift η 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 ma is equal to any 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 on 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 equation 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:

$$S = -m_0c^2 \int_{t_1}^{t_2} \sqrt{1 - v^2/c^2}\, dt - q \int_{t_1}^{t_2} [\phi(x, y, z, t) - \boldsymbol{v} \cdot \boldsymbol{A}(x, y, z, t)]\, dt.$$

The first part of the action integral is the rest mass m_0 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 $\boldsymbol{v}$ times the vector potential $\boldsymbol{A}$. Of course, we are then including only electromagnetic forces. All electric and magnetic fields are given in terms of ϕ and $\boldsymbol{A}$. This action function gives the complete theory of relativistic motion of a single particle in an electromagnetic field.

"Of course, wherever I have written $\boldsymbol{v}$, 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 point 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 those replacements for the $\boldsymbol{v}$'s that you have the formula for the action for a relativistic particle. I will leave to the more ingenious 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 $\boldsymbol{A}$, that is, for no magnetic field? Then you should get the components of the equation of motion, $d\boldsymbol{p}/dt = -q\,\nabla\phi$, where, you remember, $\boldsymbol{p} = m\boldsymbol{v}/\sqrt{1 - v^2/c^2}$.

"It is much more difficult to include also the case with a vector potential. The variations get much more complicated. But in the end, the force term does come out equal to $q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$, 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 integrand no longer has the form of the kinetic energy

minus the potential energy. That's only true in the nonrelativistic approximation. For example, the term $m_0c^2\sqrt{1 - v^2/c^2}$ 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*, $\mathcal{L}$, which is a function only of the velocities and positions of particles. So the principle of least action is also written

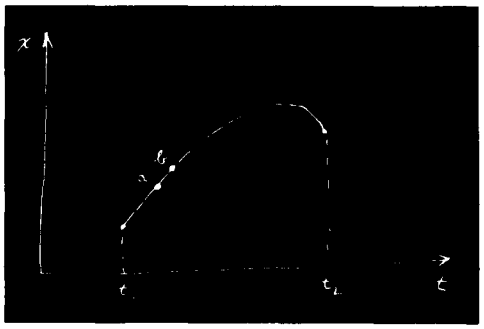
$$S = \int_{t_1}^{t_2} \mathcal{L}(x_i, v_i)\, dt,$$

where by x_i and v_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

$$\mathcal{L} = -m_0c^2\sqrt{1 - v^2/c^2} - q(\phi + \mathbf{v} \cdot \mathbf{A}).$$

"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 hate 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—and of a law which says that as you go along, there is a force that makes it accelerate. The second way tells how you inch your way along the path, and the other is a grand 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 actual 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 part from a to b is a little bit more. Otherwise you could just fiddle 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 also 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 on cause 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 things 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 $x(t)$ that we could have—for every possible imaginary 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 when a path and a nearby path all 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, Bader, 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 matrix mathematics (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 $\underline{\phi}$, plus a small deviation f, then in the first order, the change in U^* is zero. So we write

$$\phi = \underline{\phi} + f.$$

The $\underline{\phi}$ 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

$$(\nabla\phi)^2 = (\nabla\underline{\phi})^2 + 2\,\nabla\underline{\phi}\cdot\nabla f + (\nabla f)^2.$$

The only first-order term that will vary is

$$2\,\nabla\underline{\phi}\cdot\nabla f.$$

In the second term of the quantity U^*, the integrand is

$$\rho\phi = \rho\underline{\phi} + \rho f,$$

whose variable part is ρf. So, keeping only the variable parts, we need the integral

$$\Delta U^* = \int(\epsilon_0\nabla\underline{\phi}\cdot\nabla f - \rho f)\,dV.$$

"Now, following the old general rule, we have to get the darn thing all clear of derivatives of f. Let's look at what the derivatives are. The dot product is

$$\frac{\partial\underline{\phi}}{\partial x}\frac{\partial f}{\partial x} + \frac{\partial\underline{\phi}}{\partial y}\frac{\partial f}{\partial y} + \frac{\partial\underline{\phi}}{\partial z}\frac{\partial f}{\partial z},$$

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 $\underline{\phi}$. It's the same general idea we used to get rid of derivatives with respect to t. We use the equality

$$\int\frac{\partial\underline{\phi}}{\partial x}\frac{\partial f}{\partial x}\,dx = f\frac{\partial\phi}{\partial x} - \int f\frac{\partial^2\underline{\phi}}{\partial x^2}\,dx.$$

The integrated term is zero, since we have to make f zero at infinity. (That corresponds to making η zero at t_1 and t_2. So our principle should be more accurately stated: U^* is less 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

$$\Delta U^* = \int(-\epsilon_0\nabla^2\underline{\phi} - \rho)f\,dV.$$

In order for this variation to be zero for any f, no matter what, the coefficient of f must be zero and, therefore,

$$\nabla^2\underline{\phi} = -\rho/\epsilon_0.$$

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:

$$\nabla \cdot (f\,\nabla\underline{\phi}) = \nabla f \cdot \nabla\underline{\phi} + f\nabla^2\underline{\phi}.$$

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 $-\nabla\underline{\phi} \cdot \nabla f$ by $f\nabla^2\underline{\phi} - \nabla \cdot (f\nabla\underline{\phi})$, which gets integrated over volume. The divergence term integrated over volume can be replaced by a surface integral:

$$\int \nabla \cdot (f\,\nabla\underline{\phi})\,dV = \int f\,\nabla\underline{\phi} \cdot \boldsymbol{n}\,da.$$

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 U^* only in the space outside of all conductors. Then, since we can't vary $\underline{\phi}$ on the conductor, f is zero on all those surfaces, and the surface integral

$$\int f\,\nabla\underline{\phi} \cdot \boldsymbol{n}\,da$$

is still zero. The remaining volume integral

$$\Delta U^* = \int (-\epsilon_0\,\nabla^2\underline{\phi} - \rho\underline{\phi})f\,dV$$

is only to be carried out in the spaces between conductors. Of course, we get Poisson's equation again,

$$\nabla^2\underline{\phi} = -\rho/\epsilon_0.$$

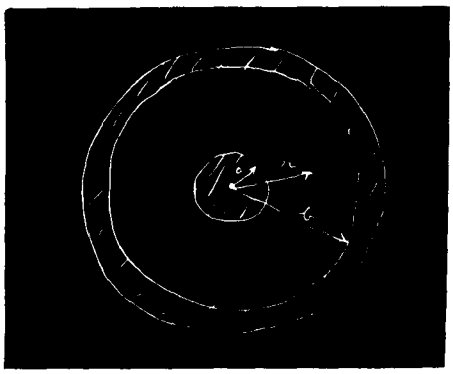
So we have shown that our original integral U^* is also a minimum if we evaluate it over the space outside of conductors all at fixed potentials (that is, such that any trial $\phi(x, y, z)$ must equal the given potential of the conductors when x, y, z is a point on the surface of a conductor).

"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 one, 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* $\underline{\phi}$, and calculate $\epsilon_0/2\int(\nabla\underline{\phi})^2\,dV$, 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 assumed 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 in 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 to find the integral U^* 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

$$\frac{1}{2}CV^2(\text{first try}) = \frac{\epsilon_0}{2}\int_a^b \frac{V^2}{(b - a)^2}\,2\pi r\,dr.$$

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 $C = 2\pi\epsilon_0/\ln(b/a)$, 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:

$\frac{b}{a}$	$\frac{C_{\text{true}}}{2\pi\epsilon_0}$	$\frac{C\text{ (first approx.)}}{2\pi\epsilon_0}$
2	1.4423	1.500
4	0.721	0.833
10	0.434	0.612
100	0.267	0.51
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 odd 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. What I get is

$$\frac{C}{2\pi\epsilon_0} = \frac{a}{b-a}\left[\frac{b}{a}\left(\frac{\alpha^2}{6} + \frac{2\alpha}{3} + 1\right) + \frac{1}{6}\alpha^2 + \frac{1}{3}\right].$$

It looks a little complicated, but it comes out of integrating the square of the field. Now I can pick my α. I know that the truth lies lower than 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 I can, 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(quadratic). Here is a table that compares C(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 isothermal) 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 entropy 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 out 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 felt by an electron moving through an ionic crystal like NaCl. [Feynman, Hellworth, Iddings, and Platzman, "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."