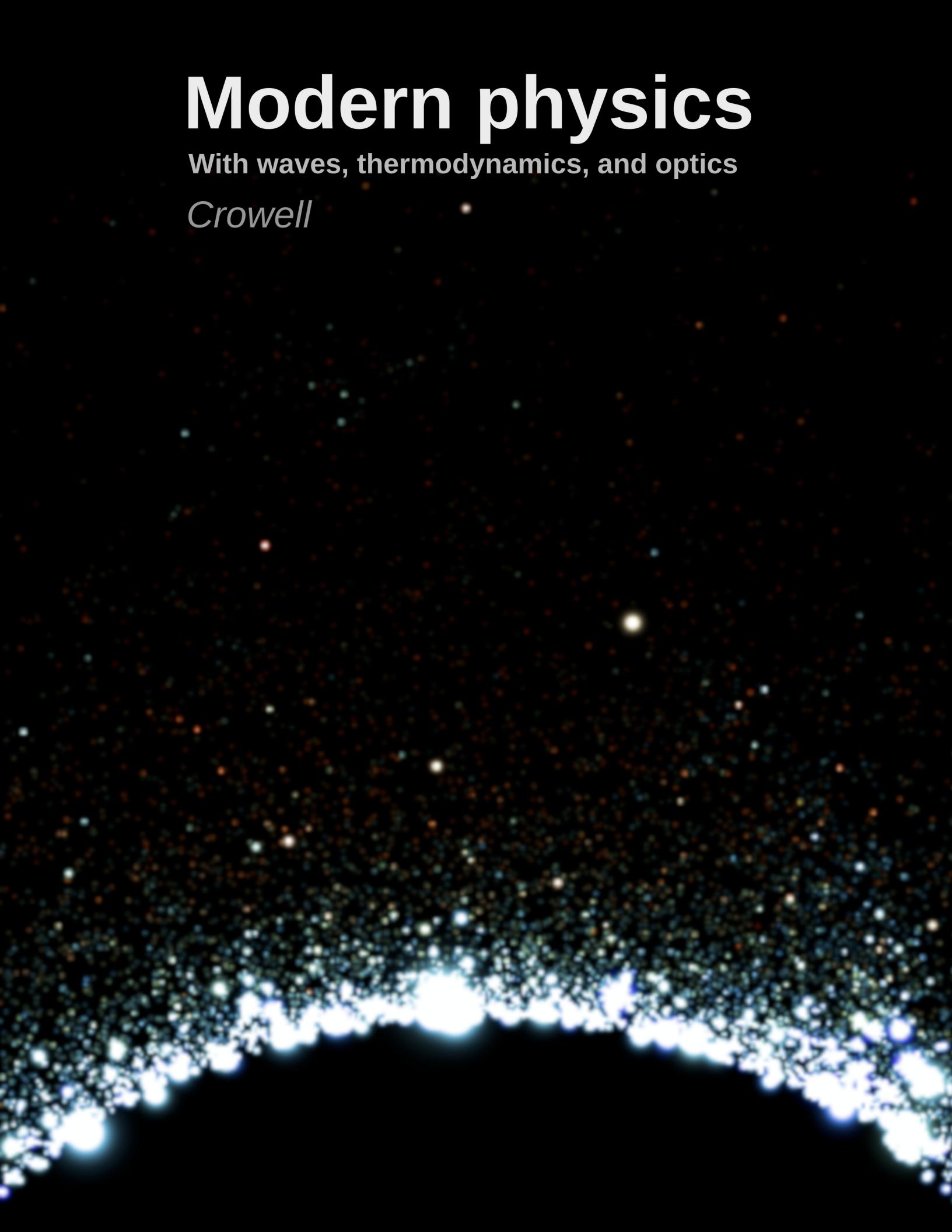


Modern physics

With waves, thermodynamics, and optics

Crowell



Modern Physics



Light and Matter

Fullerton, California

www.lightandmatter.com

copyright 2019 Benjamin Crowell

rev. October 7, 2021



This book is licensed under the Creative Commons Attribution-ShareAlike license, version 3.0, <http://creativecommons.org/licenses/by-sa/3.0/>, except for those photographs and drawings of which I am not the author, as listed in the photo credits. If you agree to the license, it grants you certain privileges that you would not otherwise have, such as the right to copy the book, or download the digital version free of charge from www.lightandmatter.com.

About the cover: The cover shows a simulated image of the sky as viewed by an observer who has fallen into a black hole. The constellations Orion and Canis Major are visible. Although the observer is already inside the event horizon, the bending of light rays by gravity makes it appear as though the black hole only occupies a relatively small part of the field of view. The bright ring consists of light from stars that are not actually very bright. Their light is amplified because they happen to lie near the event horizon. The image was rendered by the open-source library Karl, github.com/bcrowell/karl.

Brief Contents

<i>Waves and relativity</i>		
1	Time	13
2	Waves	31
3	Electromagnetic waves	63
4	The Lorentz transformation	85
5	Waves done medium well	111
6	Relativistic energy and momentum	147
<i>Thermodynamics</i>		
7	Statistics and the ideal gas	167
8	The macroscopic picture	183
9	Entropy	199
<i>Optics</i>		
10	Images, qualitatively	213
11	Images, quantitatively	233
12	Wave optics	261
<i>The microscopic description of matter and quantum physics</i>		
13	The atom and the nucleus	287
14	Probability distributions and a first glimpse of quantum physics	325
15	Light as a particle	343
16	Matter as a wave	365
17	The Schrödinger equation	393
18	Quantization of angular momentum	425

Contents

Waves and relativity

1 Time

1.1 Time-reversal symmetry	13
1.2 Relativity of time	14
1.3 Spacetime	15
1.4 The spacetime interval	17
1.5 The correspondence principle	19
1.6 A universal speed limit	20
Theoretical argument, 20.—Experimental evidence, 20.—Invariance of c , 21.—Speed of light, 22.	
1.7 Natural units	22
1.8 A speed limit for cause and effect	24
Problems	27
Exercise 1A: Spacetime	28
Exercise 1B: Signals and time	30

2 Waves

2.1 Wave motion	32
Superposition, 32.—The medium is not transported with the wave., 34.—A wave's velocity depends on the medium., 35.—Wave patterns and energy transport, 36.	
2.2 Waves on a string	38
Intuitive ideas, 38.—Approximate treatment, 38.—Treatment using calculus, 40.—Significance of the result, 41.	
2.3 Sound waves	42
2.4 Periodic waves	42
Period and frequency of a periodic wave, 42.—Graphs of waves as a function of position, 43.—Wavelength, 44.—Wave velocity related to frequency and wavelength, 44.—Sinusoidal waves, 45.—The wavenumber, 45.	
2.5 The Doppler effect	47
2.6 Reflection and standing waves	49
Reflection of waves, 49.—Standing waves, 49.	
2.7 Waves in two or three dimensions	51
The wave-vector, 51.—Degeneracy, 52.—Separability, 53.—The Laplacian, 53.	
Problems	57

3 Electromagnetic waves

3.1 Energy and momentum of electric and magnetic fields	64
A thought experiment, 64.—Expressions for the energy and momentum density, 65.—Examples of the momentum, 66.	
3.2 Geometry of a plane wave	68
\mathbf{E} and \mathbf{B} perpendicular to the direction of propagation, 68.— \mathbf{E} and \mathbf{B} equal in energy, 68.— \mathbf{E} and \mathbf{B} perpendicular to each other, 69.	
3.3 Propagation at a fixed velocity	70
3.4 The electromagnetic spectrum	71
3.5 Momentum and rate of energy flow	73
Momentum of a plane wave, 73.—Rate of energy flow, 73.	
3.6 Relativistic consequences	75
$E=mc^2$, 75.—Einstein's motorcycle, 78.	
Notes for chapter 3	80
Problems	82

4 The Lorentz transformation

4.1 Relativity of simultaneity	85
4.2 The Lorentz transformation	89
4.3 The light cone	91
4.4 The diagonal stretch factor and two of its applications	94
Definition of the stretch factor, 94.—Combination of velocities, 94.—The relativistic Doppler shift, 95.	
4.5 Length contraction	98
4.6 Magnetism as a relativistic effect	101
Notes for chapter 4	104
Problems	106
Exercise 4A: The Lorentz transformation	108

5 Waves done medium well

5.1 Measures of amplitude	112
5.2 Impedance	112
5.3 More about reflection and transmission	114
Why reflection happens, 114.—How much reflection?, 115.—Inverting and uninverting reflections, 116.—Total reflection, 116.	

5.4 Symmetric and asymmetric standing wave patterns	118
5.5 *Musical consonance and dissonance	119
5.6 Refraction and reflection in two dimensions	120
Snell's law, 120.—The index of refraction, 122.—Total internal reflection, 124.	
5.7 Review of resonance and complex numbers	125
Physical motivation for use of complex numbers: feedback systems, 125.—Complex numbers, 125.—Euler's formula, 129.—Simple harmonic motion and the LC circuit, 130.—Damped oscillations, 133.—Resonance, 134.—Dispersion, 136.	
Notes for chapter 5	139
Problems	140
6 Relativistic energy and momentum	
6.1 Mystery stuff	147
6.2 The energy-momentum vector	148
6.3 Four-vectors in general	149
6.4 Mass	150
6.5 Applications	151
6.6 A tiny bit of linear algebra	155
6.7 *Tachyons	157
Problems	159
Exercise 6A: Sports in slowlightland	161
Exercise 6B: Four-vectors and inner products	162
Thermodynamics and the microscopic description of matter	
7 Statistics, equilibrium, and energy sharing	
7.1 Basics of probability and statistics	168
Statistical independence, 168.—Addition of probabilities, 168.—Normalization, 169.—Averages, 169.	
7.2 A statistical argument for irreversible processes	170
7.3 Sizes of fluctuations	171
7.4 Equipartition	172
7.5 Heat capacities of gases	174
7.6 The ideal gas law	176
Notes for chapter 7	178
Problems	179
8 The macroscopic picture	
8.1 Pressure	184
Only pressure differences are normally significant., 185.—Variation of pressure with depth, 186.	
8.2 Temperature	188
Thermal equilibrium, 188.—Thermal expansion, 189.—Absolute zero and the kelvin scale, 190.	
8.3 Heat	191
8.4 Adiabatic expansion of a gas	192
Problems	197
9 Entropy	
9.1 Heat engines	199
9.2 Entropy	201
9.3 Entropy from a microscopic perspective	204
9.4 Phase space	205
9.5 Summary of the laws of thermodynamics	207
Notes for chapter 9	208
Problems	209

Optics

10 Images, qualitatively

10.1 Vision and the nature of light	213
10.2 The ray model of light	214
Models of light, 214.—Ray diagrams, 216.	
10.3 A virtual image	218
10.4 Curved mirrors.	221
10.5 A real image	222
10.6 Images of images	223
Problems	228
Exercise 10: Exploring images with a curved mirror	230

11 Images, quantitatively

11.1 A real image formed by a converging mirror	233
Location of the image, 233.— Magnification, 237.	
11.2 Other cases with curved mirrors	237
11.3 Images formed by lenses	242
Lenses, 242.—★The lensmaker's equation, 243.—Dispersion, 244.—★Microscopic description of refraction, 245.	
11.4 ★Aberrations	246
Problems	250
Exercise 11A: Object and image distances	259
Exercise 11B: How strong are your glasses?	260

12 Wave optics

12.1 Diffraction	261
12.2 Scaling of diffraction	264
12.3 The correspondence principle	264
12.4 Huygens' principle	265
12.5 Double-slit diffraction	266
12.6 Repetition	270
12.7 Single-slit diffraction	271
12.8 Coherence	273
Problems	276
Exercise 12A: Two-source interference	280
Exercise 12B: Single-slit interference	282

The microscopic description of matter and quantum physics

13 The atom and the nucleus

13.1 The electrical nature of matter and quantization of charge	287
13.2 The electron	288
Cathode rays, 288.—Were cathode rays a form of light, or of matter?, 289.—Thomson's experiments, 290.—The cathode ray as a subatomic particle: the electron, 291.	
13.3 The raisin cookie model of the atom	292
13.4 The nucleus	294
Radioactivity, 294.—The planetary model, 298.—Atomic number, 301.—The struc-	

ture of nuclei, 306.—The strong nuclear force, alpha decay and fission, 310.—The weak nuclear force; beta decay, 312.—Fusion, 314.—Nuclear energy and binding energies, 316.—Biological effects of ionizing radiation, 318.

Problems	323
Exercise 13: Nuclear decay	324

14 Probability distributions and a first glimpse of quantum physics

14.1 Probability distributions	325
14.2 The variance and standard deviation	327
14.3 Errors in random counts: Poisson statistics	329
14.4 Exponential decay	329
Half-life, 329.—Calculations for exponential decay, 331.	
14.5 A first glimpse of quantum physics	333
Notes for chapter 14	336
Problems	337
Exercise 14: Probability distributions	340

15 Light as a particle

15.1 Evidence for light as a particle	343
15.2 How much light is one photon?	346
The photoelectric effect, 346.—An unexpected dependence on frequency, 346.—Numerical relationship between energy and frequency, 348.	

15.3 Wave-particle duality	351
A wrong interpretation: photons interfering with each other, 352.—The probability interpretation, 352.	
15.4 Nonlocality and entanglement	354
Nonlocality, 354.—Entanglement, 355.	
15.5 Photons in three dimensions	359
Problems	361

16 Matter as a wave

16.1 Electrons as waves	366
Wavelength related to momentum, 366.—What kind of wave is it?, 369.—Quantum numbers and bra-ket notation, 372.—“Same state” versus “same wavefunction”, 372.	
16.2 Dispersive waves	373

16.3 Bound states	376	411.	
16.4 The uncertainty principle	379	17.8 Time evolution and unitarity	413
Eliminating randomness through measurement?, 379.—The Heisenberg uncertainty principle, 379.		The simplest cases of time evolution, 413.—The two-state system, 414.—The time-dependent Schrödinger equation, 414.—Unitarity, 415.	
16.5 Decoherence and quantum computing	381	Notes for chapter 17	420
Decoherence, 381.—Brains, classical computers, and quantum computers, 383.		Problems	421
16.6 A crude model of the hydrogen atom	384	18 Quantization of angular momentum	
Modeling, 384.—Estimation of the energy levels, 385.—Comparison with experiment, 386.		18.1 Quantization of angular momentum	425
Notes for chapter 16	388	18.2 Three dimensions	426
Problems	389	18.3 Quantum numbers	428
17 The Schrödinger equation		Completeness, 428.—Sets of compatible quantum numbers, 429.—Complete and compatible sets of quantum numbers, 430.	
17.1 Electrons in electric fields	393	18.4 The Stern-Gerlach experiment	430
Defining a wavelength when the wavelength is varying, 393.—Tunneling, 394.		18.5 Intrinsic spin	432
17.2 The Schrödinger equation	394	Experimental evidence, 432.—Odds and evens, and how they add up, 433.—Inner product, 435.—Classification of states in hydrogen, 435.	
17.3 Solutions when U is constant; tunneling	396	18.6 The Pauli exclusion principle	437
17.4 Three dimensions	400	Notes for chapter 18	439
17.5 Use of complex numbers	403	Problems	440
17.6 Linearity of the Schrödinger equation	408	Exercise 18: The quantum moat	442
17.7 The inner product and observables	409	Photo Credits	460
The inner product, 409.—Observables,			

Waves and relativity

A note on the notes

Since the book is free online, I've tried to format it so that it's easy to hop around in it conveniently on a laptop. The blue text in the table of contents is hyperlinked. Sometimes there are mathematical details or technical notes that are not likely to be of much interest to you on the first read through. These are relegated to the end of each chapter. In the main text, they're marked with blue hyperlinked symbols that look like this: ≥ 137 . On a computer, you can click through if you want to read the note, and then click on a similar-looking link to get back to the main text. The numbers are page numbers, so if you're using the book in print, you can also get back and forth efficiently.

Chapter 1

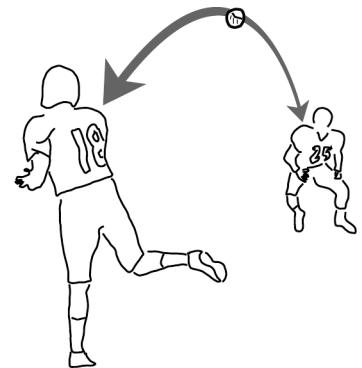
Time

This type of physics course can easily seem to the student like a random grab-bag of topics, consisting of everything that didn't fit in the earlier semesters on mechanics and electromagnetism. But there is a clear organizing principle for most of what we'll be studying. It has to do with two surprising facts about time. In particular, one of these facts leads us to the conclusion that light and matter can't really be made of particles, as envisioned by Isaac Newton's grand vision of the universe — they must be made of waves.

1.1 Time-reversal symmetry

The first of these facts about time is that the laws of physics we have studied so far never distinguish the past from the future. We can take a video of the football pass in figure a, reverse it in time, and then check the motion of the ball to see if it still complies with Newton's laws. It does. Mathematically, this is because Newton's laws predict accelerations, and since acceleration is a second derivative, $d^2 x / dt^2$, it doesn't change when we mirror-image the $x(t)$ graph, interchanging the positive and negative t axes. For example, a negative second derivative means that the graph is concave down, and it's still concave down after we mirror-image it.

This time-reversal symmetry also holds for electromagnetism. As a simple example, Gauss's law tells us, in the special case of an electric field in a vacuum, that the field lines can't begin or end at a point. (They can only begin or end on charges, which don't exist in a vacuum.) This law doesn't even refer to time explicitly, so clearly nothing goes wrong if we reverse time. Some of the other electromagnetic quantities, such as currents, reverse themselves in a time-reversed world, but this can all be handled in a consistent



a / The football could travel in either direction while obeying Newton's laws.

way, if we adopt certain rules for which things flip and which ones don't. For an example in the same spirit as the football pass, we can reverse the current through an electromagnet. The magnetic field reverses itself, but the time-reversed version of the experiment still satisfies the rules we've learned about electromagnetism.

The mystery, then, is why there are certain phenomena that, unlike the football pass or the electromagnet, seem impossible to act out in reverse. The concrete sidewalk in front of my house was wet when it was first poured. Then the concrete hardened. When it rains, the concrete doesn't reverse the chemical reaction and become soft again, nor will it do so even if I apply some heat, although it would be consistent with conservation of energy if it did so.

There are many other phenomena that are irreversible. Candles don't unburn. Eggs don't unscramble. We can't remember what the stock market will do tomorrow. Any one of these could be used to define an "arrow of time," e.g., the future is the direction in which there are more scrambled eggs and fewer unscrambled ones. All of these arrows of time seem to be consistent with each other, which suggests that they have some common underlying explanation. But this explanation does not seem to be in the laws that govern atoms and light waves.¹ We will see in our study of *thermodynamics*, later in this course, that the explanation comes from statistics and cosmology.

1.2 Relativity of time

The second of our two strange facts about time relates directly to our upcoming topic of waves, and it also forms the basis for our discussion of relativity, which will be a thread running through the whole course. A colorful experiment demonstrating this fact was done by Hafele and Keating in 1971 (figure b). The two physicists brought atomic clocks with them on round-the-world flights aboard commercial passenger jets, then compared the clocks with other clocks that had been left at home. When the clocks were reunited, they *disagreed* by ~ 100 ns. The results were consistent with Einstein's 1915 theory of relativity, and were interpreted as a combined effect from motion and gravity. Because it's difficult to move a clock very fast without putting it on an airplane, it wasn't until 2010 that Chou *et al.*² succeeded in carrying out a conceptually simpler tabletop experiment in which a clock was simply moved around (at speeds on the order of 10 m/s) without taking it to high elevation, thus isolating the effect of motion from the gravitational



b / The clock took up two seats, and two tickets were bought for it under the name of "Mr. Clock."

¹There is one obscure instance in nuclear physics in which the basic laws *do* violate time-reversal symmetry, but as far as we can tell this has nothing to do with all the other phenomena, which don't involve this very specific type of nuclear decay.

²Science 329 (2010) 1630

effect. It is the effect of motion that will be of interest to us here.

It would be natural to try to explain this effect of motion as arising from the clocks' sensitivity to noise, cabin pressure, or vibration. But exactly the same effect is observed with other, completely different types of clocks under completely different circumstances, even with processes such as the decay of elementary particles moving at high speeds — the radioactive half-life is prolonged if the particles are in motion.

The conclusion is that *time itself* is not absolute. If we synchronize two clocks side by side, then move them in different ways, and finally reunite them, we will in general see that they disagree. In experiments of this type, we find that time interval is greatest for a clock that moves *inertially*, i.e., without accelerating. Any other motion results in a smaller time. The direction of the effect is easily remembered by thinking of the Planet of the Apes movies, in which human astronauts return to earth and find themselves in a distant future, in which chimpanzees, gorillas, and orangutans have dominion over the planet. *Less* time passes for the astronauts, who undergo the violent accelerations of space travel.

Discussion question

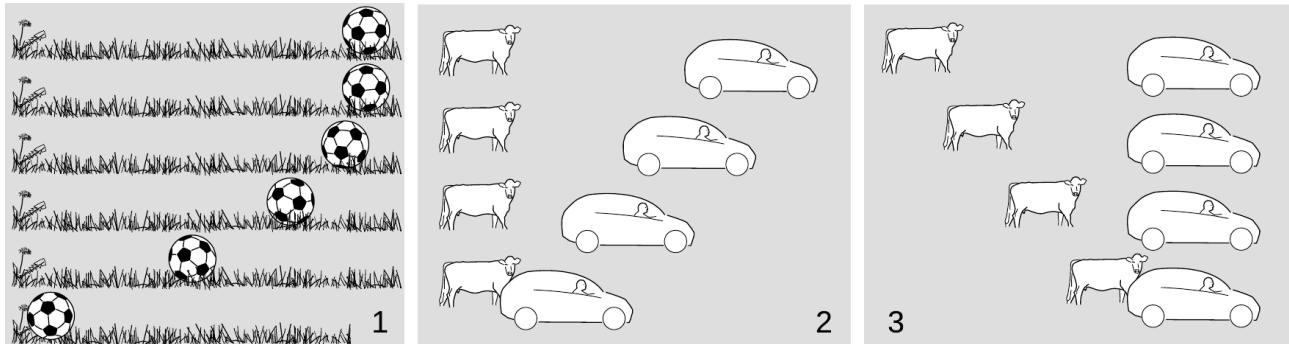
A Mechanical clocks can be affected by motion. For example, it was a significant technological achievement to build a clock that could sail aboard a ship and still keep accurate time, allowing longitude to be determined. How is this similar to or different from relativistic time dilation?

1.3 Spacetime

We can visualize what's going on using *spacetime diagrams* like the ones in figure c. In c/1, we follow the soccer ball's motion as we go up the page. It moves to the right, is decelerated by the frictional force from the grass, and comes to a stop. This is essentially a graph of the ball's position as a function of time, but tilted sideways. The graph is curved, which shows that the ball's motion is noninertial.

Aristotle would have accepted figure c/2, but not c/3, which is drawn in the driver's frame of reference, according to which the cow and the landscape are moving to the left. Aristotle believed in the absolute nature of notions such as distance and whether or not two events happen in the same place. Galileo taught us that these things are relative rather than absolute. Looking at the initial and final images of the car, we can't necessarily say how far apart in space they are from each other. According to c/2, they're several meters apart, but c/3 says they're in the same place.

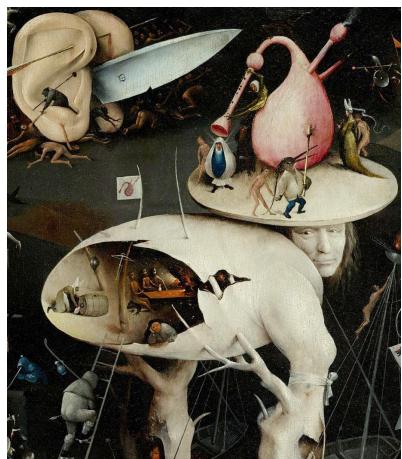
Einstein extended this relativity to time as well. We accept Einstein's view because of evidence such as the atomic clock exper-



c / 1. The ball's world-line is noninertial. 2. As time passes, the cow says that the car moves. 3. In the driver's frame, it's the cow that moves.

iment described on p. 14. We will see shortly that relativity makes different observers disagree not just on the lengths of time intervals, as shown in the atomic clock experiment, but also on whether or not two things happen simultaneously. What Galileo did to the notion of “same place,” Einstein did to “same time.”

One of the reasons that nineteenth-century Europeans found Marxism alarming was because it was atheistic, and they felt that without the framework of religion, there could be no basis for morality. For similar reasons, I was deeply disoriented when I first encountered relativity. The idea had been firmly inculcated that the universe was described by mathematical functions, and the natural habitat of those functions was graph paper. The graph paper provided what seemed like a necessary framework. For a position-time graph, the vertical lines meant “same time,” and the horizontal ones “same place.” Somehow it didn’t bother me much when Galileo erased the same-place lines (or at least relegated them to subjectivity), but without the same-time lines I felt lost, as if I were wandering in a landscape of Hieronymus Bosch’s hell or Dali’s melting watches.



d / Hell, according to Hieronymus Bosch (1450-1516).

One of the disorienting things about this vision of the universe is that it takes away the notion that we can have a literal “vision” of the universe. We no longer have the idea of a snapshot of the landscape at a certain moment frozen in time. The sense of vision is merely a type of optical measurement, in which we receive signals that have traveled to our eyes at some finite speed (the speed of light). What relativity substitutes for the Galilean instantaneous snapshot is the concept of *spacetime*, which is like the graph paper when its lines have been erased. Every point on the paper is called an *event*. How can we even agree on the existence of an event, or define which one we are talking about, if we can’t necessarily agree on its time or position? The relativist’s attitude is that if a firecracker pops, that’s an event, everyone agrees that it’s an event, and x and t coordinates are just an optional and arbitrary name or label for the event. Labeling an event with coordinates is like

God asking Adam to name all the birds and animals: the animals weren't consulted and didn't care.

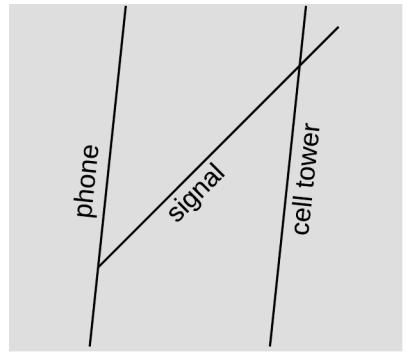
My grandparents' German shepherd lived for a certain amount of time, so he was not just a pointlike event in spacetime. We know how to represent the motion of such things as curves on an x - t graph. From the point of view of relativity, the curve *is* the thing — we make no distinction between the dog and the dog's track through spacetime. Such a track is called a *world-line*. A world-line is a set of events strung together continuously: the dog as a puppy in Walnut Creek in 1964, the dog dozing next to the TV in 1970, and so on. The strange terminology is translated from German, and is supposed to be a description of the idea that the line is the thing's track *through* the world, i.e., through spacetime.

Sometimes if we want to describe an event, we can describe it as the beginning or end of a world-line: the dog's birth, or the firecracker's self-destruction. More commonly, we pick out an event of interest as the intersection of two world-lines, as in figure e. In this figure, as is common in relativity, we omit any indications of the axes, since the idea is that events and world-lines are primary, and coordinates secondary. In this book we will use the convention that time progresses from the bottom of the diagram to the top, which means that a spacetime diagram is turned sideways compared to a graph of $x(t)$. This is the standard convention in relativity.

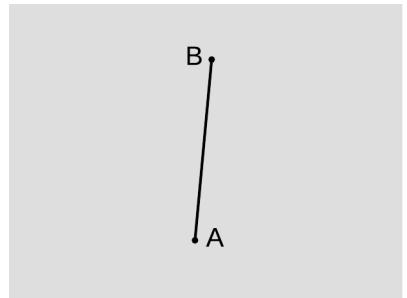
self-check A

Here is a spacetime graph for an empty object such as a house: | |. Explain why it looks like this. My grandparents had a dog door with a flap cut into their back door, so that their dog could come in and out. Draw a spacetime diagram showing the dog going out into the back yard. Can an observer using another frame of reference say that the dog didn't go outside?

▷ Answer, p. 454



e / A phone transmits a 1 or 0 to a cell tower. The phone, the tower, and the signal all have world-lines. Two events, corresponding to the transmission and reception of the signal, can be defined by the intersections of the world-lines.



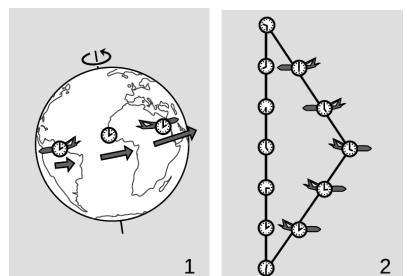
f / A clock measures the interval from A to B.

1.4 The spacetime interval

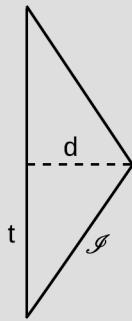
Suppose that a clock has an inertial world-line that extends from event A to event B, as in figure f. Given A and B, there is only one possible motion for the clock, because two points determine a line in spacetime, just as in Euclidean geometry. On a spacetime diagram, we have a line segment AB, and the clock gives us a measure of how “long” this line segment is. We define the spacetime interval \mathcal{I} (script “I”) as

$$\mathcal{I}^2 = (\text{time elapsed on clock})^2.$$

The squares may seem unnecessary. But without them, we would be defining a quantity that had a plus or minus sign, depending on whether B was after A or before. As we've seen in section 1.1, the laws of physics don't actually define what is the past and what is the



g / The atomic clock experiment shown in simplified form (1) and in even more simplified form (2).



h / Time dilation in the atomic clock experiment.

future, so if we are only given the world-line AB, we don't expect to be able to say whether it measures up as positive or negative. The equation only defines the absolute value of \mathcal{I} , not its sign, which kind of makes sense. We'll see later that other things also come out simpler with this style of definition.

Now let's see how this plays out in a semi-realistic context. Figure g/1 shows the general idea of how the experiment on p. 14 was done. There were three sets of atomic clocks. Some stayed in the lab, on the surface of the earth, while others went around the world, either to the east or to the west, while the planet spun on its axis. That's pretty complicated, so let's consider the simplified version in figure g/2. One clock moves inertially while the other does an out-and-back trip. The diagram is drawn in the frame of reference of the laboratory where the first clock remains. The result is a diagram that looks like a triangle

Now suppose, as in figure h, that the distance traveled by the plane, as measured in the lab frame, is d . The time measured by the lab clock over the bottom half of the diagram is t . There will be some spacetime interval \mathcal{I} measured by the flying clock during this half of its journey.

Although the slanting world-line with \mathcal{I} written next to it looks longer than the one with t next to it, we actually have $\mathcal{I} < t$. If time worked the way Galileo and Newton believed it did, then we would simply have $\mathcal{I} = t$. Keep in mind that this is basically just an $x(t)$ graph, and diagonal distances measured literally on such a graph never actually mean anything. If it helps your intuition, you can think of the line segments as roads, and the quantities t and \mathcal{I} as indications of how many gas stations you pass on those roads. In this particular diagram, the longer road just happens to have fewer gas stations.

The fact that \mathcal{I} is less than t is referred to as time dilation. We say that time dilation is a "relativistic" effect, meaning an effect that nobody expected until Einstein's theory of relativity came along.

Now we expect, for the reasons discussed above, that things will come out nicer if we talk about \mathcal{I}^2 and t^2 rather than \mathcal{I} and t . We also know that \mathcal{I}^2 will be at its greatest when $d = 0$. In general, when we look at any smooth function near a maximum, we expect it to look like a concave-down parabola, so we should have

$$\mathcal{I}^2 \approx t^2 - (\text{const.})d^2.$$

We happen to have had the historical hindsight to have known that it would be more convenient in the long run to express this in terms of the squares, but as long as d is small, that's not a substantive assumption whose failure would invalidate the approximation. We would just have an equation analogous to this one, but without the squares on \mathcal{I} and t , and with a different value for the constant.

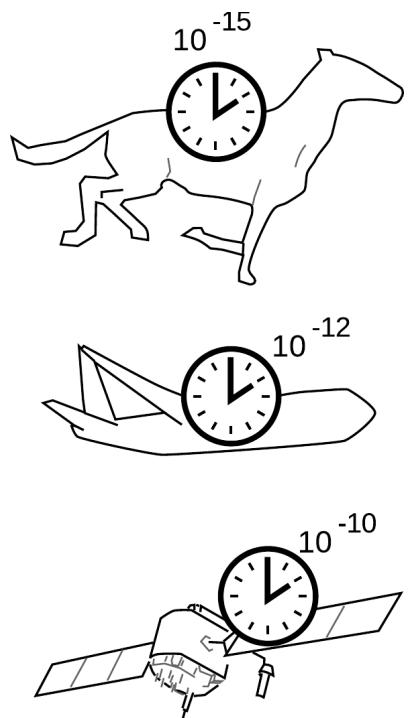
self-check B

What would figure h look like for $d = 0$, and how would \mathcal{I} turn out?
Does this make sense? What about $d < 0$? ▷ Answer, p. 454

1.5 The correspondence principle

What is this mysterious constant in our equation for \mathcal{I} ? One thing we can tell is that it must be very small when written in everyday units, because we don't notice these relativistic effects on time in everyday life. Furthermore, the units of the variables occurring in the equation show that the constant must also have units of s^2/m^2 , or one over velocity squared. In other words, we could write it as $1/c^2$, where c is some universal constant that has units of speed. The velocity of the plane, relative to the lab, is d/t , and so we can say that relativistic effects are small when the velocities are small compared to c .

It makes sense that relativistic effects like time dilation are small in everyday life. Newton's laws have already been thoroughly tested by experiments under a wide variety of conditions, so a new theory like relativity must agree with Newton's to a good approximation, within the Newtonian theory's realm of applicability. This requirement of backward-compatibility is known as the *correspondence principle*. Newton lived in an era when the fastest mode of transportation was a galloping horse, and the best pendulum clocks would accumulate errors of perhaps a minute over the course of several days. A horse is much slower than a jet plane, so the distortion of time would have had a relative size of only $\sim 10^{-15}$ — much smaller than the clocks were capable of detecting. At the speed of a passenger jet, the effect is about 10^{-12} , and state-of-the-art atomic clocks in 1971 were capable of measuring that. A GPS satellite travels much faster than a jet airplane, and the effect on the satellite turns out to be $\sim 10^{-10}$. The general idea here is that all physical laws are approximations, and approximations aren't simply right or wrong in different situations. Approximations are better or worse in different situations, and the question is whether a particular approximation is good enough in a given situation to serve a particular purpose. The faster the motion, the worse the Newtonian approximation of absolute time. Whether the approximation is good enough depends on what you're trying to accomplish. The correspondence principle says that the approximation must have been good enough to explain all the experiments done in the centuries before Einstein came up with relativity.



i / The correspondence principle requires that the relativistic distortion of time become small for small velocities.

1.6 A universal speed limit

1.6.1 Theoretical argument

We have only given an argument that our equation for the interval should be valid for small velocities, but in fact experiments show that it is valid at all speeds,³ so we have

$$\mathcal{I}^2 = t^2 - \frac{d^2}{c^2}.$$

One thing we can see from this equation is that something goes very wrong if the speeds get too high. If a clock could travel faster than c , then we would have $d = vt > ct$, and the right-hand side would be negative. The interval would then be the square root of a negative number, which wouldn't seem to make sense. This suggests that c can be interpreted as a sort of universal speed limit for the motion of any material object. For something moving exactly at c , we have $\mathcal{I} = 0$ between any two points on its world-line.

1.6.2 Experimental evidence

Experiments confirm this. As with the Hafele-Keating airplane experiment, the best educational example is neither the first historically nor the latest and most recent one. The clearest example I know of was done by Bertozzi in 1964.⁴ If you find this kind of thing mind-blowing or hard to believe, you may find it interesting to watch the half-hour educational video made by Bertozzi, in which every part of the experimental setup is demonstrated and explained.⁵

This experiment directly acts out a scenario often proposed by beginners at relativity who find this sort of thing paradoxical. Suppose we use a force to accelerate an object such as a spaceship to a speed very close to c , say 99.999% of c . Then we keep on applying the force for a little longer. Doesn't the object go faster than c ? Doing anything like this with a human-scale object like a spaceship is impractical because of the vast amounts of energy that would be required (problem 1, p. 27). Bertozzi used electrons. The electrons were accelerated by an electric field E through a distance ℓ_1 . Applying Newton's laws gives Newtonian predictions a_N for the acceleration and t_N for the time required.⁶

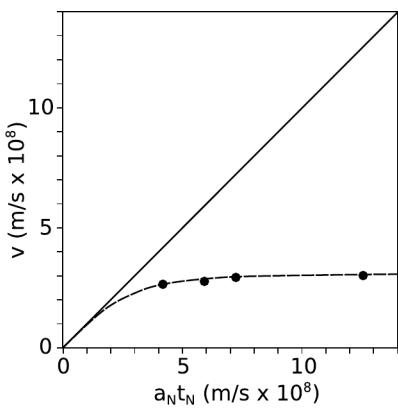
The electrons were then allowed to fly down a pipe for a further distance $\ell_2 = 8.4$ m without being acted on by any force. The time of flight t_2 for this second distance was used to find the final velocity $v = \ell_2/t_2$ to which they had actually been accelerated.

³This also ends up being necessary to make the theoretical framework fit together. See p. 90.

⁴Bertozzi, "Speed and kinetic energy of relativistic electrons," Am. J. Phys. 32 (1964) 551

⁵<https://tinyurl.com/bertozziexperiment>

⁶Newton's second law gives $a_N = F/m = eE/m$. The constant-acceleration equation $\Delta x = (1/2)at^2$ then gives $t_N = \sqrt{2m\ell_1/eE}$.



j / Electrons subjected to a constant force do not have constant acceleration and never surpass c .

Figure j shows Bertozzi’s data, supplemented by some data at lower speeds from a later, similar experiment.⁷ According to Newton, an acceleration a_N acting for a time t_N should produce a final velocity $a_N t_N$. The solid line in the graph shows the prediction of Newton’s laws, which is that a constant force exerted steadily over time will produce a velocity that rises linearly and without limit.

The experimental data, shown as black dots, clearly tell a different story. The velocity never goes above a certain maximum value, which we identify as c . The dashed line shows the predictions of special relativity (ch. 6), which are in good agreement with the experimental results.

As another example of the kind of experimental work that has tested relativity’s claims about c , there was a dramatic incident in 2011 in which particle physicists believed they had detected signals at a laboratory near Rome that had been emitted in an accelerator experiment in the Alps at a distance of 731.296 km (measured with GPS), and that these signals had arrived with a time delay of 2.43928 ms. The speed implied by these measurements exceeded c by 0.002%. Although this might seem like only a tiny discrepancy, the theory it violated had by then been so firmly established by decades of experiments that the claim set off a frenzy of theoretical and experimental work to try to disprove it, explain it, or use it as evidence for new theories of physics. A year later, the scientific collaboration that had found the observed result announced publicly, and with considerable embarrassment, that they had found the result to be due to two mistakes: a loose cable and a faulty electronic clock.

1.6.3 Invariance of c

Based on your knowledge of physics according to Galileo and Newton, you might think that it would not make sense for there to be a universal constant that was interpreted as a velocity, since velocities have different values depending on one’s frame of reference. We’ll take this up in more detail in sec. 3.6.2, p. 78, but basically the straight addition of velocities in relative motion turns out to be a low-velocity approximation. At high velocities this addition rule becomes inaccurate, and as we approach c it breaks down completely. (For more detail on this, see sec. 4.4.2, p. 94.)

We describe this by saying that c is *invariant*, i.e., it is the same in all frames of reference. Another good example of an invariant quantity is electric charge: if charge were not invariant, then, for example, the hydrogen atom would not be electrically neutral, as we observe it to be in experiments. The interval \mathcal{I} is also invariant, because it’s defined with reference to an actual measuring device:

⁷Marvel and Jolivette, open access at arxiv.org/abs/1108.5977. To make the low-energy portion of the graph legible, Bertozzi’s highest-energy data point is omitted.

a clock that travels along a certain world-line. If the clock is, for example, an hourglass, then every observer is forced to agree that at event A the sand was all still in the top, and at event B the last grain fell down to the bottom.

In the logical framework used by Einstein in his original 1905 paper on relativity, the invariance of c is one of the two basic postulates that he uses as logical foundations to derive everything else. If we instead take it as a postulate that the interval is invariant and behaves according to $\mathcal{I}^2 = t^2 - d^2/c^2$, then something moving at c has $\mathcal{I} = 0$, and since \mathcal{I} is invariant, we reach the same conclusion that motion at c is an invariant fact.

1.6.4 Speed of light

In SI units, this maximum speed has the numerical value of about 3.0×10^8 m/s. We often refer to it as the speed of light, since visible light travels at c . Light is a wave disturbance in the electric and magnetic fields, and the visible spectrum from red to violet constitutes one part of a much larger electromagnetic spectrum, which includes phenomena as apparently disparate as radio waves and x-rays. Different parts of the spectrum are distinguished either by their frequency (the number of vibrations per unit time) or, equivalently, by their wavelength (the distance between successive wave crests). Fundamentally, it's best to think of c not as the speed of light but as a speed limit, or as a sort of conversion factor between time and space units.

1.7 Natural units

Because c can be thought of as a conversion factor between the units for time and distance, we can save work by using a system of units in which the units of time and distance are the same. As an example, astronomers often use the year as the unit of time, and the light-year as their measure of distance. A light-year is the distance traveled by light in one year. The distance from the sun to the nearest star is about four light-years. In these units, c equals 1 — one light-year per year. In these units, our equation for the spacetime interval is simply

$$\mathcal{I}^2 = t^2 - d^2,$$

which is a lot simpler than the form of the equation we were writing before. We don't have to use light-years and year. For example, we could also use seconds and light-seconds, or, to better than 2% precision, feet and nanoseconds. Units in which $c = 1$ are called *natural* units. In natural units, velocities are always unitless. For example, $v = 0.5$ means a velocity equal to half the speed of light.

Using natural units for relativity is, well, natural. Because spacetime doesn't have natural "graph-paper lines" built in, space and time aren't completely separate from each other in relativity. That's

why we talk about “spacetime” instead of just “space and time.” There is a good analogy with our energy units. James Joule demonstrated with a paddlewheel experiment in 1876 that heat and mechanical work were things that could be converted into each other. We would say today that they are both measures of energy, and we would use units of joules for both. Given this insight, it would be masochism to keep on using different units for the different types of energy.

Written in the natural-units form $\mathcal{I}^2 = t^2 - d^2$, our equation for the interval looks very much like the Pythagorean theorem, and it does a similar job: measuring “how far” two points are. The only difference in the forms of the equations is the minus sign. Using different units for t and d would be like going out to survey a plot of land and using different units for east-west and north-south coordinates.

Even if we intend to use data expressed in SI units, often it is easier to do all of our algebra in natural units, which are simpler because $c = 1$, and all factors of c can therefore be omitted. Then at the very end we reinsert factors of c so that the units make sense in SI.

Motion of a ray of light

example 1

- ▷ The motion of a certain ray of light is given by the equation $x = -t$. Is this expressed in natural units, or in SI units? Convert to the other system.
- ▷ The equation is in natural units. It wouldn’t make sense in SI units, because we would have meters on the left and seconds on the right. To convert to SI units, we insert a factor of c in the only possible place that will cause the equation to make sense: $x = -ct$.

Planet of the Apes

example 2

- ▷ In the original movie version of *Planet of the Apes*, the human astronauts find that they have returned to earth after 2.006×10^3 years, but the time interval seems to them to have been only 1.5 years. How fast did their spaceship go?
- ▷ Let t be the time on earth as they do the outward leg of their trip and \mathcal{I} the time they experience on this leg. These times are half the quantities for the whole trip. In natural units, we have $\mathcal{I}^2 = t^2 - d^2$, so $d = \sqrt{t^2 - \mathcal{I}^2}$, and their speed was $v = d/t$, or

$$v = \sqrt{1 - \left(\frac{\mathcal{I}}{t}\right)^2}.$$

If we needed any reminding that this was in natural units rather than SI, we would notice that because clearly the equation doesn’t make sense in SI. The left-hand side would have units of m/s, while the right-hand side is unitless. There is only one way to fix

the units by inserting factors of c , and that's to write

$$v = c \sqrt{1 - \left(\frac{\mathcal{I}}{t}\right)^2}.$$

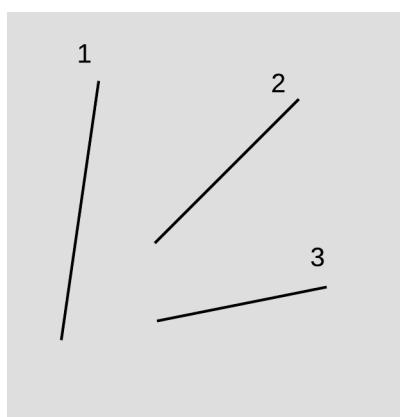
This is what we would have gotten if we had done the whole calculation in SI, but the algebra would have been messier. Plugging in numbers, we get 2.9979234×10^8 m/s, which is just a tad less than $c = 2.9979246 \times 10^8$ m/s. The numerical result is in fact easier to read and more informative in natural units: $v = 0.99999960$.

Spacetime diagrams are conventionally drawn in natural units, so that motion at c appears as a world-line with a slope of 1.

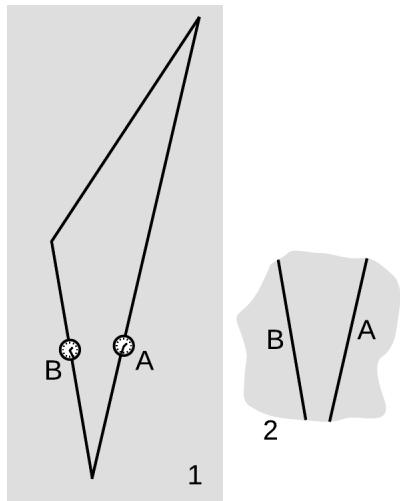
self-check C

Of the three world-lines in the figure, which could represent a material object? Which could represent a flash of light? Are these things moving inertially, or non-inertially?

▷ Answer, p. 454



Self-check C.



k / A small portion of a time-dilation experiment.

1.8 A speed limit for cause and effect

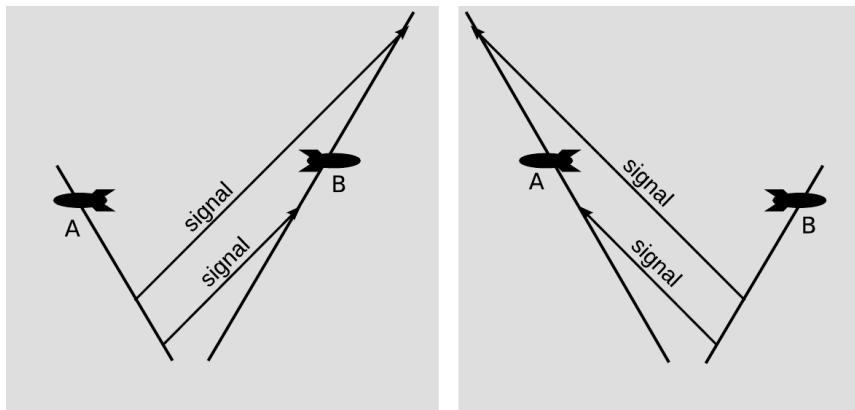
Figure k/1 shows a time-dilation experiment like the ones we've been considering. Alice moves inertially on world-line A, along with her clock, while her twin sister Betty takes the noninertial path B. The time elapsed on Alice's clock will be greater, because her world-line is the inertial one. But suppose we focus on the smaller part of spacetime shown in k/2. During this time, Betty isn't undergoing any accelerations. *Both* twins' motion is inertial.

Now suppose that during the time shown in k/2 they pull out their phones and get in contact. We might imagine that they could now establish whose time was slower than normal and whose is faster. Will Alice hear Betty's voice as slowed-down and Darth Vadery, while Betty complains that Alice sounds like a chipmunk? But no, this is impossible because of the perfect symmetry in k/2. Alice is free to choose her own ship as a frame of reference, in which case she considers herself to be at rest while Betty moves. But the situation is completely symmetrical, so Betty can say the same thing. Because motion is relative, we can't say who is "really" moving and who is "really" at rest. And yet they can't both say that the other's time is normal, because then there would be no way for any discrepancy to accumulate by the time they were reunited — as we see in the actual experiment.

One way to try to resolve this paradox would be to give up on the idea that motion is always relative, which was an assumption that we had to appeal to explicitly. As described in more detail in sec. 4.1, p. 85, this was the consensus when Einstein was a university student, around 1896. Einstein rejected this idea and retained as a central pillar of his description of spacetime the idea that motion is

relative. This is why the theory is called *relativity*. The version of the theory that doesn't include gravity is called special relativity, and the later version with gravity is general relativity.

In relativity, the key to resolving this paradox is that the description of the scenario contains a hidden assumption, that communication using cell phones is instantaneous. It isn't. A cell phone is a kind of radio, and its signals are encoded as radio waves, which propagate at c .



I / Signals don't resolve the dispute over who is really slow.

Figure 1 shows what actually happens to Alice and Betty if, for example, one twin makes two hand claps near her phone, separated by a one-second interval. If the two signals traveled at infinite speed, then their world-lines would be horizontal — they would be received at the same time they were sent. But because they actually travel at c , which is finite, they are sent at one time and arrive at some later time, and the signals' world-lines have some slope. We use natural units on our diagrams so that this is a 45-degree angle. We can see on the diagram that when Alice sends the two hand claps, Betty receives them at times that are spread to more than one second apart, but exactly the same thing happens when we flip the diagram. The situation is completely symmetric, and each twin perceives the other's transmission as having been slowed down.

The parable of Alice and Betty shows that logical consistency is preserved — but *only* if there is no mechanism for sending signals instantaneously, as in the “subspace radio” of the Star Trek universe.⁸ We therefore have an unexpected link between the nature of time and the way in which fields of force propagate across the universe. Force laws like Coulomb's law $F = kq_1q_2/r^2$ and Newton's law of gravity $F = Gm_1m_2/r^2$ look like they imply instantaneous action at a distance. There is no t in either equation. But this can't be how things actually operate, because it leads to the paradox of

⁸See also p. 92 for a different argument that leads to a stronger form of this conclusion.



m / Circular ripples propagate outward in a puddle at a fixed, finite speed.

Alice and Betty.

Switching from Star Trek to Star Wars for our metaphors, we find that if there is a “disturbance in the force,” then it must propagate at some finite speed, like ripples on a pond, figure m. This leads us to a far-reaching conclusion: that the universe must really be made out of waves (or particles that also have wave properties, as we’ll see when we take up our study of quantum physics at the end of this book).

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 On p. 20, I remarked that accelerating a macroscopic (i.e., not microscopic) object to relativistic speeds would require an unreasonable amount of energy. Suppose that the starship Enterprise from Star Trek has a mass of 8.0×10^7 kg, about the same as the Queen Elizabeth 2. Compute the kinetic energy it would have to have if it was moving at $0.1c$. Compare with the total energy content of the world's nuclear arsenals, which is about 10^{21} J. Although you don't yet know the exact relativistic equation for kinetic energy, this speed is small enough compared to c so that based on the correspondence principle, it's reasonable to imagine that using the Newtonian $K = (1/2)mv^2$ will give a good approximation. ✓

2 We can't go faster than the speed of light, but because of relativistic time dilation, it is theoretically possible to travel to arbitrarily distant parts of the galaxy within a human lifetime.

(a) Suppose that we want to visit a star at distance d in light-years, and we want our one-way trip, at constant velocity, to seem like a time interval \mathcal{I} to us. Starting from the equation $\mathcal{I}^2 = t^2 - d^2$ (the version expressed in natural units), find the time t that this takes according to observers on earth. ✓

(b) The photo shows the constellation Canis Major, including the stars Sirius and Adhara. Because we can't see depth when we look at the celestial sphere, we may be misled into imagining that these two stars are close to one another. Actually Sirius is one of our nearest neighbors, at 8.6 light-years, while Adhara is 570 light-years away. For each of these stars, take $\mathcal{I} = 10$ years and evaluate t . ✓

(c) The data in this problem are conveniently expressed in natural units, but for practice, let's see how your answer to part a would have looked in SI units. Insert factors of c as demonstrated in examples 1-2, p. 23. Don't redo the algebra for part a, which would be more complicated. The point here is to practice how to put factors of c in at the end of the calculation, as demonstrated in those examples. ✓



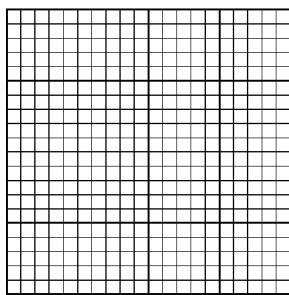
Problem 2. The brightest star is Sirius, which is the dog's head. The hind foot is Adhara.

Exercise 1A: Spacetime

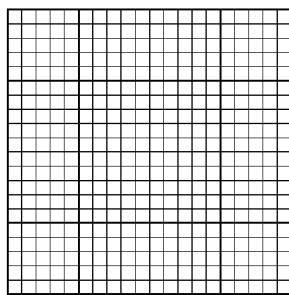
The following paper-and-pencil exercises involve spacetime diagrams. Tracks on them are called world-lines. As in the other examples in this book, a time axis is always closer to vertical, and a position axis closer to horizontal, and natural units are used, so that the universal speed c is a slope of $+1$ or -1 .

Draw spacetime diagrams of: (1) a box with a particle bouncing back and forth inside it, (2) a ray of light being absorbed by a leaf on a tree, (3) an atomic nucleus splitting up into two parts (fissioning), (4) a cloud of gas collapsing gravitationally to form our solar system.

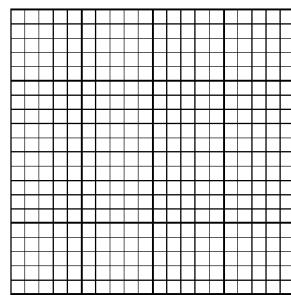
1



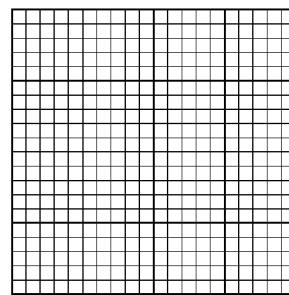
2



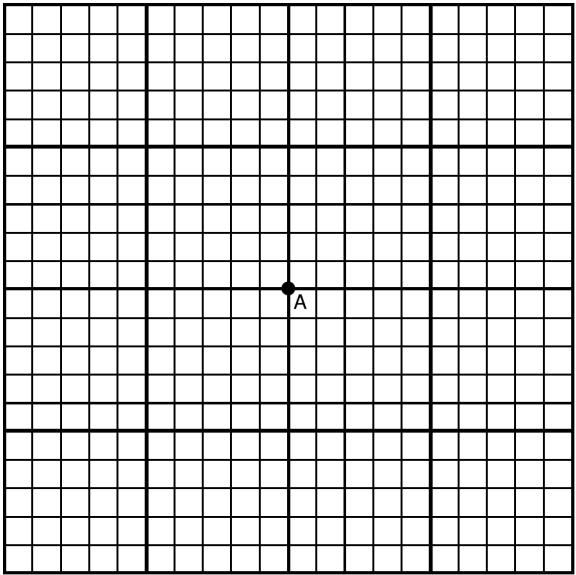
3



4



5. Event A is given. Mark examples of events B-F that satisfy these criteria. A material object travels from A to B. An object traveled from C to A. A ray of light emitted at A is received at D. A ray of light emitted at E is received at A. F can have no possible cause-and-effect relationship with A.

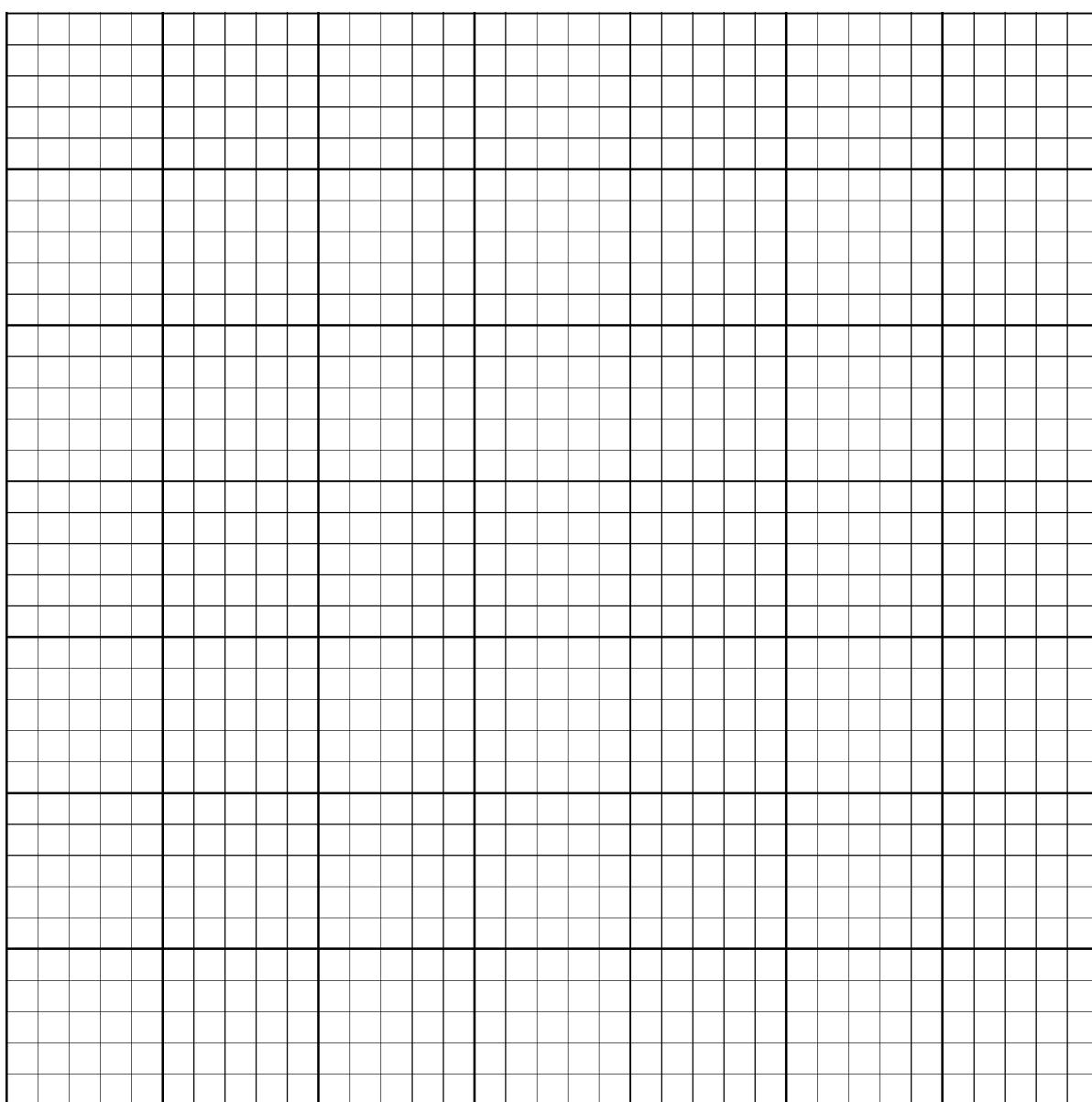


Exercise 1B: Signals and time

Relativity says we can never directly know what's happening "now" at some distant place. At best we can receive signals from a distance and get indirect information.

1. Alice stays on earth while Betty flies away at $3/5$ of the speed of light. On the graph paper, draw a spacetime diagram in the frame of the earth showing their world-lines, starting with Betty's departure.
2. After a certain amount of time \mathcal{I}_A on Alice's clock (say 10 units, if that's convenient on the graph paper you're using), she sends a radio beep toward Betty. Draw the world-line of this signal and find the event at which Betty receives it.
3. Find the time \mathcal{I}_B on Betty's clock between her departure and the time when she gets the signal. Compare with the time Alice waited before transmitting, using the ratio $D = \mathcal{I}_B/\mathcal{I}_A$.

Interpretation: This ratio is the factor by which Alice's signals seem slowed down to Betty. A similar effect would occur without relativity, e.g., for sound waves and race cars, simply because the distance is widening, but the size of the effect would be different.



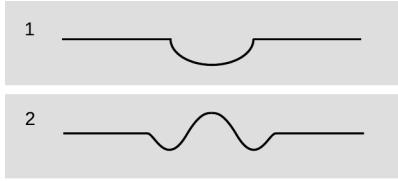


The vibrations of this electric bass string are converted to electrical vibrations, then to sound vibrations, and finally to vibrations of our eardrums.

Chapter 2

Waves

At the end of chapter 1, we saw that because of the way time works in our universe, it is not possible for light and matter to be made out of particles, as envisioned by Isaac Newton. Instead, our universe must be made out of waves (or, as we'll see when we study quantum physics in ch. 13-18, things that have both particle and wave characteristics). In this chapter we will study mechanical waves, which are made out of vibrations of physical objects. These are concrete and easy to conceptualize. Then in chapter 3 we will turn to electromagnetic waves, such as visible light, which will be our first example of the *fundamental* kinds of waves that the universe is ultimately composed of.



a / Your finger makes a depression in the surface of the water, 1. The wave pattern starts evolving, 2, after you remove your finger.

2.1 Wave motion

Let's start with an intuition-building exercise that deals with waves in matter. Put your fingertip in the middle of a cup of water and then remove it suddenly. You'll have noticed two results that are surprising to most people. First, the flat surface of the water does not simply sink uniformly to fill in the volume vacated by your finger. Instead, ripples spread out, and the process of flattening out occurs over a long period of time, during which the water at the center vibrates above and below the normal water level. This type of wave motion is the topic of the present section. Second, you've found that the ripples bounce off of the walls of the cup, in much the same way that a ball would bounce off of a wall. In the next section we discuss what happens to waves that have a boundary around them. Until then, we confine ourselves to wave phenomena that can be analyzed as if the medium (e.g., the water) was infinite and the same everywhere.

It isn't hard to understand why removing your fingertip creates ripples rather than simply allowing the water to sink back down uniformly. The initial crater, a/1, left behind by your finger has sloping sides, and the water next to the crater flows downhill to fill in the hole. The water far away, on the other hand, initially has no way of knowing what has happened, because there is no slope for it to flow down. As the hole fills up, the rising water at the center gains upward momentum, and overshoots, creating a little hill where there had been a hole originally. The area just outside of this region has been robbed of some of its water in order to build the hill, so a depressed "moat" is formed, a/2. This effect cascades outward, producing ripples.

There are three main ways in which wave motion differs from the motion of objects made of matter.

2.1.1 Superposition

If you watched the water in the cup carefully, you noticed the ghostlike behavior of the reflected ripples coming back toward the center of the cup and the outgoing ripples that hadn't yet been reflected: they passed right through each other. This is the first, and the most profound, difference between wave motion and the motion of objects: waves do not display any repulsion of each other analogous to the normal forces between objects that come in contact. Two wave patterns can therefore overlap in the same region of space, as shown in figure b. Where the two waves coincide, they add together. For instance, suppose that at a certain location in at a certain moment in time, each wave would have had a crest 3 cm above the normal water level. The waves combine at this point to make a 6-cm crest. We use negative numbers to represent depressions in the

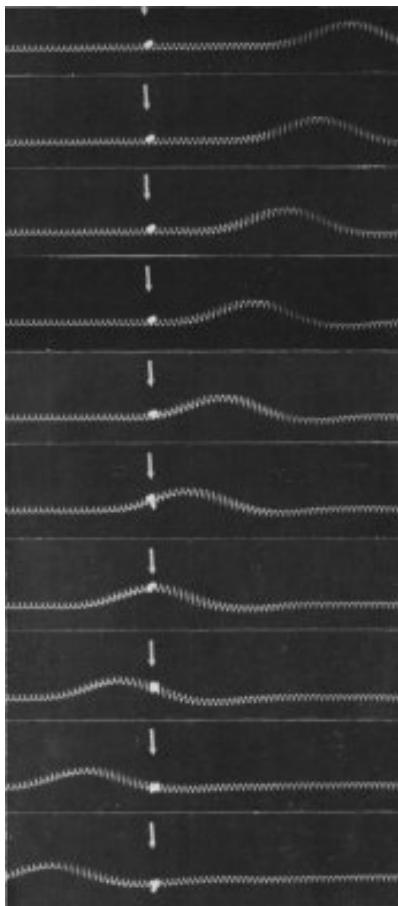


b / The two circular patterns of ripples pass through each other. Unlike material objects, wave patterns can overlap in space, and when this happens they combine by addition.

water. If both waves would have had a troughs measuring -3 cm , then they combine to make an extra-deep -6 cm trough. A $+3\text{ cm}$ crest and a -3 cm trough result in a height of zero, i.e., the waves momentarily cancel each other out at that point. This additive rule is referred to as the principle of superposition, “superposition” being merely a fancy word for “adding.”

Superposition can occur not just with sinusoidal waves like the ones in the figure above but with waves of any shape. The figures on the following page show superposition of wave pulses. A pulse is simply a wave of very short duration. These pulses consist only of a single hump or trough. If you hit a clothesline sharply, you will observe pulses heading off in both directions. This is analogous to the way ripples spread out in all directions when you make a disturbance at one point on water. The same occurs when the hammer on a piano comes up and hits a string.

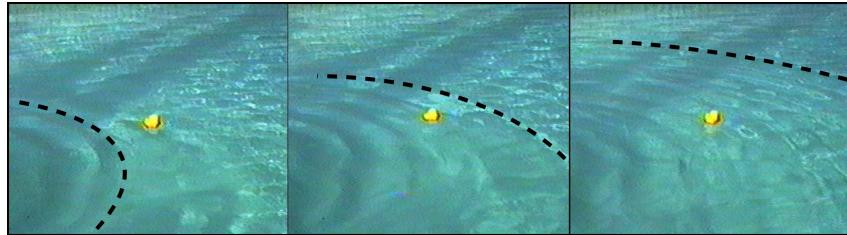
Experiments to date have not shown any deviation from the principle of superposition in the case of light waves. For other types of waves, it is typically a very good approximation for low-energy waves.



d / As the wave pulse goes by, the ribbon tied to the spring is not carried along. The motion of the wave pattern is to the right, but the medium (spring) is moving from side to side, not to the right. (*PSSC Physics*)



e / Example 1. The surfer is dragging his hand in the water.



c / As the wave pattern passes the rubber duck, the duck stays put. The water isn't moving with the wave.

2.1.2 The medium is not transported with the wave.

The sequence of three photos in figure c shows a series of water waves before it has reached a rubber duck (left), having just passed the duck (middle) and having progressed about a meter beyond the duck (right). The duck bobs around its initial position, but is not carried along with the wave. This shows that the water itself does not flow outward with the wave. If it did, we could empty one end of a swimming pool simply by kicking up waves! We must distinguish between the motion of the medium (water in this case) and the motion of the wave pattern through the medium. The medium vibrates; the wave progresses through space.

self-check A

In figure d, you can detect the side-to-side motion of the spring because the spring appears blurry. At a certain instant, represented by a single photo, how would you describe the motion of the different parts of the spring? Other than the flat parts, do any parts of the spring have zero velocity?

▷ Answer, p. 454

Surfing

example 1

The incorrect belief that the medium moves with the wave is often reinforced by garbled secondhand knowledge of surfing. Anyone who has actually surfed knows that the front of the board pushes the water to the sides, creating a wake — the surfer can even drag his hand through the water, as in figure e. If the water was moving along with the wave and the surfer, this wouldn't happen. The surfer is carried forward because forward is downhill, not because of any forward flow of the water. If the water was flowing forward, then a person floating in the water up to her neck would be carried along just as quickly as someone on a surfboard. In fact, it is even possible to surf down the back side of a wave, although the ride wouldn't last very long because the surfer and the wave would quickly part company.

2.1.3 A wave's velocity depends on the medium.

A material object can move with any velocity, and can be sped up or slowed down by a force that increases or decreases its kinetic energy. Not so with waves. The speed of a wave, depends on the properties of the medium (and perhaps also on the shape of the wave, for certain types of waves). Sound waves travel at about 340 m/s in air, 1000 m/s in helium. If you kick up water waves in a pool, you will find that kicking harder makes waves that are taller (and therefore carry more energy), not faster. The sound waves from an exploding stick of dynamite carry a lot of energy, but are no faster than any other waves. In the following section we will give an example of the physical relationship between the wave speed and the properties of the medium.

Breaking waves

example 2

The velocity of water waves increases with depth. The crest of a wave travels faster than the trough, and this can cause the wave to break.

Once a wave is created, the only reason its speed will change is if it enters a different medium or if the properties of the medium change. It is not so surprising that a change in medium can slow down a wave, but the reverse can also happen. A sound wave traveling through a helium balloon will slow down when it emerges into the air, but if it enters another balloon it will speed back up again! Similarly, water waves travel more quickly over deeper water, so a wave will slow down as it passes over an underwater ridge, but speed up again as it emerges into deeper water.

Hull speed

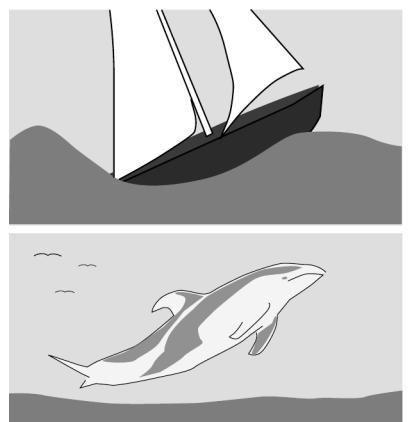
example 3

The speeds of most boats, and of some surface-swimming animals, are limited by the fact that they make a wave due to their motion through the water. The boat in figure g is going at the same speed as its own waves, and can't go any faster. No matter how hard the boat pushes against the water, it can't make the wave move ahead faster and get out of the way. The wave's speed depends only on the medium. Adding energy to the wave doesn't speed it up, it just increases its amplitude.

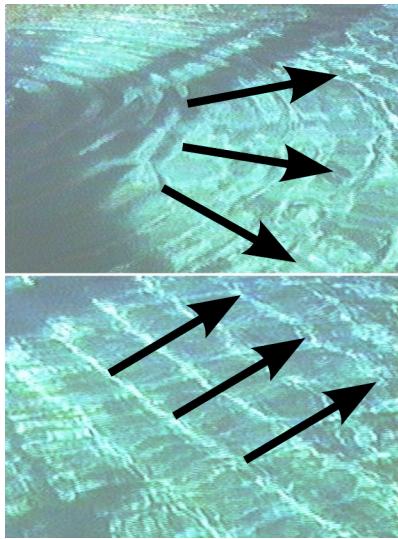
A water wave, unlike many other types of wave, has a speed that depends on its shape: a broader wave moves faster. The shape of the wave made by a boat tends to mold itself to the shape of the boat's hull, so a boat with a longer hull makes a broader wave that moves faster. The maximum speed of a boat whose speed is limited by this effect is therefore closely related to the length of its hull, and the maximum speed is called the hull speed. Sailboats designed for racing are not just long and skinny to make them more streamlined — they are also long so that their hull speeds will be high.



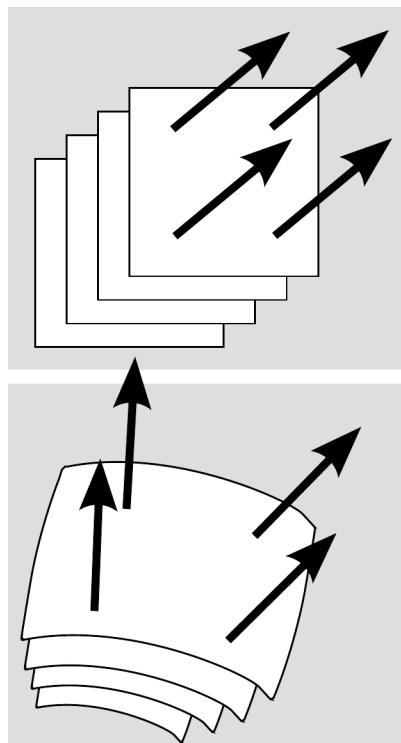
f / Example 2: a breaking wave.



g / Example 3. The boat has run up against a limit on its speed because it can't climb over its own wave. Dolphins get around the problem by leaping out of the water.



h / Circular and linear wave patterns.



i / Plane and spherical wave patterns.

2.1.4 Wave patterns and energy transport

If the magnitude of a wave's velocity vector is preordained, what about its direction? Waves spread out in all directions from every point on the disturbance that created them. If the disturbance is small, we may consider it as a single point, and in the case of water waves the resulting wave pattern is the familiar circular ripple, h/1. If, on the other hand, we lay a pole on the surface of the water and wiggle it up and down, we create a linear wave pattern, h/2. For a three-dimensional wave such as a sound wave, the analogous patterns would be spherical waves and plane waves, i.

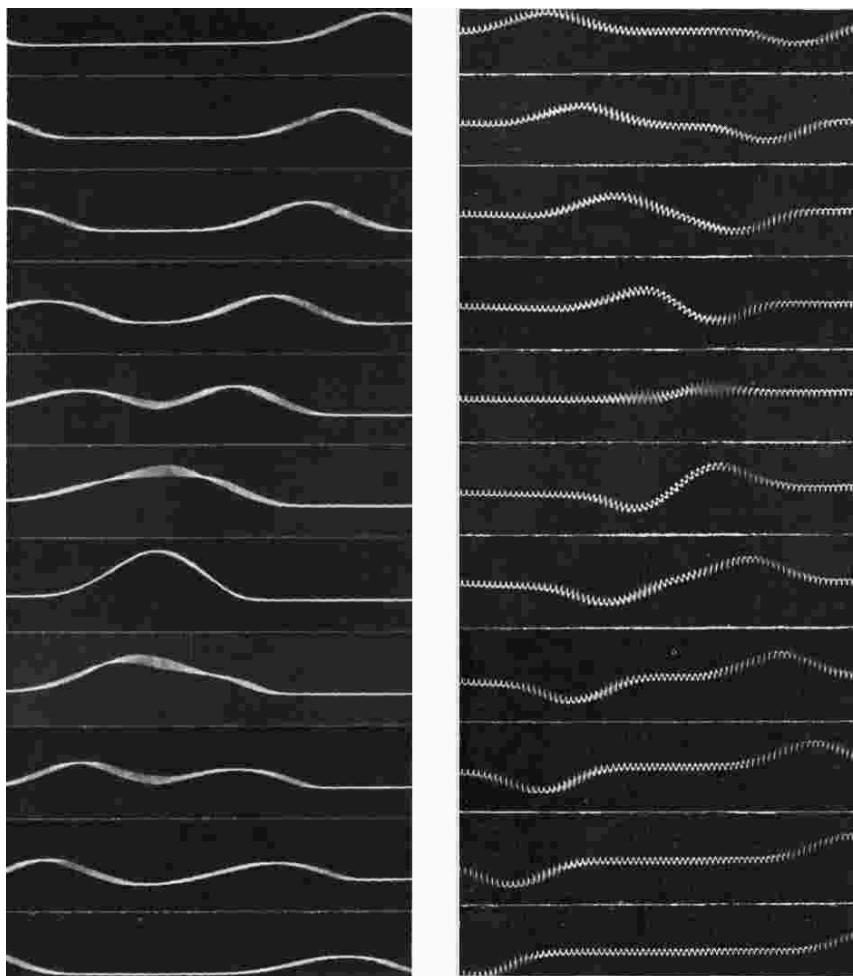
Infinitely many patterns are possible, but linear or plane waves are often the simplest to analyze, because the velocity vector is in the same direction no matter what part of the wave we look at. Since all the velocity vectors are parallel to one another, the problem is effectively one-dimensional. Throughout this chapter and the next, we will restrict ourselves mainly to wave motion in one dimension, while not hesitating to broaden our horizons when it can be done without too much complication.

In figures h and i, energy is being transported in the direction of the black arrows. How do we tell how much energy a wave has? We already know that for an oscillating mass such as a pendulum, the energy depends on the amplitude as $E \propto A^2$ when A is small. This follows simply because when a smooth function has a minimum, it generically looks like a parabola close to that minimum. It also makes sense that the energy doesn't change when we flip the sign of A , which is just a change of phase. The same logic applies to waves. As we will see in more detail ch. 5, the constant of proportionality depends on the medium. In the case of an electromagnetic wave (ch. 3), you already know that the energy densities of the electric and magnetic fields are proportional to the square of the fields.

This is a good time at which to vaccinate you against a common “gotcha” involving waves. The speed at which a wave transports energy and information is called the group velocity, v_g . The speed at which the peaks of the wave pattern move is called the phase velocity, v_p . The simplest kinds of waves, which we've been implicitly assuming so far, are those for which $v_p = v_g$. Their wave patterns glide through space without changing shape. In sec. 16.2, p. 373, we will take a closer look at *dispersive* waves, which have $v_p \neq v_g$, change their shape as they propagate, and have a velocity that depends not just on the medium but also on the wavelength.

Discussion questions

A The left panel of the figure shows a sequence of snapshots of two positive pulses on a coil spring as they move through each other. In the right panel, which shows a positive pulse and a negative one, the fifth frame has the spring just about perfectly flat. If the two pulses have essentially canceled each other out perfectly, then why does the motion pick up again? Why doesn't the spring just stay flat?



j / Discussion question A.

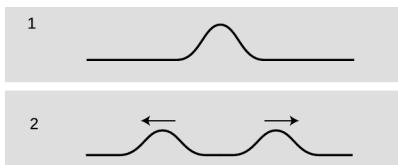
B Sketch two positive wave pulses on a string that are overlapping but not right on top of each other, and draw their superposition. Do the same for a positive pulse running into a negative pulse.

C A traveling wave pulse is moving to the right on a string. Sketch the velocity vectors of the various parts of the string. Now do the same for a pulse moving to the left.

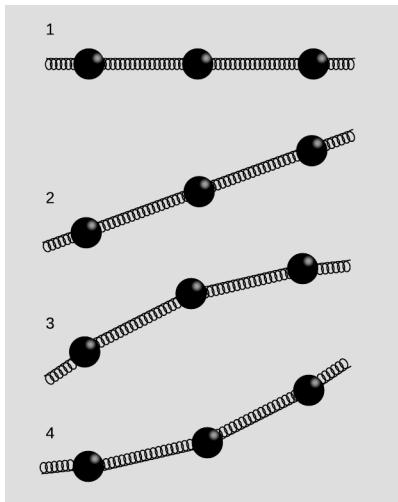
D In a spherical sound wave spreading out from a point, how would the energy of the wave fall off with distance?



k / Hitting a key on a piano causes a hammer to come up from underneath and hit a string (actually a set of three). The result is a pair of pulses moving away from the point of impact.



l / A pulse on a string splits in two and heads off in both directions.



m / Modeling a string as a series of masses connected by springs.

2.2 Waves on a string

So far you've learned some counterintuitive things about the behavior of waves, but intuition can be trained. The first half of this section aims to build your intuition by investigating a simple, one-dimensional type of wave: a wave on a string. If you have ever stretched a string between the bottoms of two open-mouthed cans to talk to a friend, you were putting this type of wave to work. Stringed instruments are another good example. Although we usually think of a piano wire simply as vibrating, the hammer actually strikes it quickly and makes a dent in it, which then ripples out in both directions. Since this chapter is about free waves, not bounded ones, we pretend that our string is infinitely long.

After the qualitative discussion, we will use simple approximations to investigate the speed of a wave pulse on a string. This quick and dirty treatment is then followed by a rigorous attack using the methods of calculus, which turns out to be both simpler and more general.

2.2.1 Intuitive ideas

Consider a string that has been struck, l/1, resulting in the creation of two wave pulses, l/2, one traveling to the left and one to the right. This is analogous to the way ripples spread out in all directions from a splash in water, but on a one-dimensional string, "all directions" becomes "both directions."

We can gain insight by modeling the string as a series of masses connected by springs, m. (In the actual string the mass and the springiness are both contributed by the molecules themselves.) If we look at various microscopic portions of the string, there will be some areas that are flat, 1, some that are sloping but not curved, 2, and some that are curved, 3 and 4. In example 1 it is clear that both the forces on the central mass cancel out, so it will not accelerate. The same is true of 2, however. Only in curved regions such as 3 and 4 is an acceleration produced. In these examples, the vector sum of the two forces acting on the central mass is not zero. The important concept is that curvature makes force: the curved areas of a wave tend to experience forces resulting in an acceleration toward the mouth of the curve. Note, however, that an uncurved portion of the string need not remain motionless. It may move at constant velocity to either side.

2.2.2 Approximate treatment

We now carry out an approximate treatment of the speed at which two pulses will spread out from an initial indentation on a string. For simplicity, we imagine a hammer blow that creates a triangular dent, n/1. We will estimate the amount of time, t , required until each of the pulses has traveled a distance equal to the width of the pulse itself. The velocity of the pulses is then $\pm w/t$.

As always, the velocity of a wave depends on the properties of the medium, in this case the string. The properties of the string can be summarized by two variables: the tension, T , and the mass per unit length, μ (Greek letter mu).

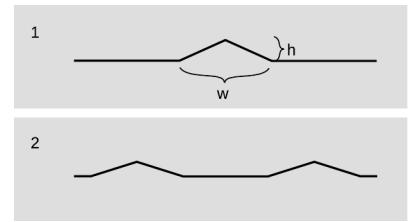
If we consider the part of the string encompassed by the initial dent as a single object, then this object has a mass of approximately μw (mass/length \times length=mass). (Here, and throughout the derivation, we assume that h is much less than w , so that we can ignore the fact that this segment of the string has a length slightly greater than w .) Although the downward acceleration of this segment of the string will be neither constant over time nor uniform across the pulse, we will pretend that it is constant for the sake of our simple estimate. Roughly speaking, the time interval between $n/1$ and $n/2$ is the amount of time required for the initial dent to accelerate from rest and reach its normal, flattened position. Of course the tip of the triangle has a longer distance to travel than the edges, but again we ignore the complications and simply assume that the segment as a whole must travel a distance h . Indeed, it might seem surprising that the triangle would so neatly spring back to a perfectly flat shape. It is an experimental fact that it does, but our analysis is too crude to address such details.

The string is kinked, i.e., tightly curved, at the edges of the triangle, so it is here that there will be large forces that do not cancel out to zero. There are two forces acting on the triangular hump, one of magnitude T acting down and to the right, and one of the same magnitude acting down and to the left. If the angle of the sloping sides is θ , then the total force on the segment equals $2T \sin \theta$. Dividing the triangle into two right triangles, we see that $\sin \theta$ equals h divided by the length of one of the sloping sides. Since h is much less than w , the length of the sloping side is essentially the same as $w/2$, so we have $\sin \theta = 2h/w$, and $F = 4Th/w$. The acceleration of the segment (actually the acceleration of its center of mass) is

$$\begin{aligned} a &= \frac{F}{m} \\ &= \frac{4Th}{\mu w^2}. \end{aligned}$$

The time required to move a distance h under constant acceleration a is found by solving $h = (1/2)at^2$ to yield

$$\begin{aligned} t &= \sqrt{2h/a} \\ &= w \sqrt{\frac{\mu}{2T}}. \end{aligned}$$



n / A triangular pulse spreads out.

Our final result for the speed of the pulses is

$$v = w/t$$

$$= \sqrt{\frac{2T}{\mu}}.$$

The remarkable feature of this result is that the velocity of the pulses does not depend at all on w or h , i.e., any triangular pulse has the same speed. It is an experimental fact (and we will also prove rigorously below) that any pulse of any kind, triangular or otherwise, travels along the string at the same speed. Of course, after so many approximations we cannot expect to have gotten all the numerical factors right. The correct result for the speed of the pulses is

$$v = \sqrt{\frac{T}{\mu}}.$$

The importance of the above derivation lies in the insight it brings —that all pulses move with the same speed — rather than in the details of the numerical result. The reason for our too-high value for the velocity is not hard to guess. It comes from the assumption that the acceleration was constant, when actually the total force on the segment would diminish as it flattened out.

2.2.3 Treatment using calculus

After expending considerable effort for an approximate solution, we now display the power of calculus with a rigorous and completely general treatment that is nevertheless much shorter and easier. Let the flat position of the string define the x axis, so that y measures how far a point on the string is from equilibrium. The motion of the string is characterized by $y(x, t)$, a function of two variables. Knowing that the force on any small segment of string depends on the curvature of the string in that area, and that the second derivative is a measure of curvature, it is not surprising to find that the infinitesimal force dF acting on an infinitesimal segment dx is given by

$$dF = T \frac{\partial^2 y}{\partial x^2} dx.$$

(This can be proved by vector addition of the two infinitesimal forces acting on either side.) The symbol ∂ stands for a partial derivative, e.g., $\partial/\partial x$ means a derivative with respect to x that is evaluated while treating t as a constant. The acceleration is then $a = dF/dm$, or, substituting $dm = \mu dx$,

$$\frac{\partial^2 y}{\partial t^2} = \frac{T}{\mu} \frac{\partial^2 y}{\partial x^2}.$$

The second derivative with respect to time is related to the second derivative with respect to position. This is no more than a fancy

mathematical statement of the intuitive fact developed above, that the string accelerates so as to flatten out its curves.

Before even bothering to look for solutions to this equation, we note that it already proves the principle of superposition, because the derivative of a sum is the sum of the derivatives. Therefore the sum of any two solutions will also be a solution.

Based on experiment, we expect that this equation will be satisfied by any function $y(x, t)$ that describes a pulse or wave pattern moving to the left or right at the correct speed v . In general, such a function will be of the form $y = f(x - vt)$ or $y = f(x + vt)$, where f is any function of one variable. Because of the chain rule, each derivative with respect to time brings out a factor of v . Evaluating the second derivatives on both sides of the equation gives

$$(\pm v)^2 f'' = \frac{T}{\mu} f''.$$

Squaring gets rid of the sign, and we find that we have a valid solution for any function f , provided that v is given by

$$v = \sqrt{\frac{T}{\mu}}.$$

2.2.4 Significance of the result

This specific result for the speed of waves on a string, $v = \sqrt{T/\mu}$, is utterly unimportant. Don't memorize it. Don't take notes on it. Try to erase it from your memory.

What *is* important about this result is that it is an example of two things that are usually true, at least approximately, for mechanical waves in general:

1. The speed at which a wave moves does not depend on the size or shape of the wave.
2. The speed of a mechanical wave depends on a combination of two properties of the medium: some measure of its *inertia* and some measure of its *tightness*, i.e., the strength of the force trying to bring the medium back toward equilibrium.

self-check B

- (a) What is it about the equation $v = \sqrt{T/\mu}$ that relates to fact 1 above?
- (b) In the equation $v = \sqrt{T/\mu}$, which variable is a measure of inertia, and which is a measure of tightness? (c) Now suppose that we produce compressional wave pulses in a metal rod by tapping the end of the rod with a hammer. What physical properties of the rod would play the roles of inertia and tightness? How would you expect the speed of compressional waves in lead to compare with their speed in aluminum?

▷ Answer, p. 454

2.3 Sound waves

The phenomenon of sound is easily found to have all the characteristics we expect from a wave phenomenon:

- Sound waves obey superposition. Sounds do not knock other sounds out of the way when they collide, and we can hear more than one sound at once if they both reach our ear simultaneously.

- The medium does not move with the sound. Even standing in front of a titanic speaker playing earsplitting music, we do not feel the slightest breeze.

- The velocity of sound depends on the medium. Sound travels faster in helium than in air, and faster in water than in helium. Putting more energy into the wave makes it more intense, not faster. For example, you can easily detect an echo when you clap your hands a short distance from a large, flat wall, and the delay of the echo is no shorter for a louder clap.

Although not all waves have a speed that is independent of the shape of the wave, and this property therefore is irrelevant to our collection of evidence that sound is a wave phenomenon, sound does nevertheless have this property. For instance, the music in a large concert hall or stadium may take on the order of a second to reach someone seated in the nosebleed section, but we do not notice or care, because the delay is the same for every sound. Bass, drums, and vocals all head outward from the stage at 340 m/s, regardless of their differing wave shapes. (The speed of sound in a gas is related to the gas's physical properties in example 9 on p. 194.)

If sound has all the properties we expect from a wave, then what type of wave is it? It is a series of compressions and expansions of the air. Even for a very loud sound, the increase or decrease compared to normal atmospheric pressure is no more than a part per million, so our ears are apparently very sensitive instruments. In a vacuum, there is no medium for the sound waves, and so they cannot exist. The roars and whooshes of space ships in Hollywood movies are fun, but scientifically wrong.

2.4 Periodic waves

2.4.1 Period and frequency of a periodic wave

You choose a radio station by selecting a certain frequency. We have already defined period and frequency for vibrations,

$$T = \text{period} = \text{seconds per cycle}$$

$$f = \text{frequency} = 1/T = \text{cycles per second}$$

$$\omega = \text{angular frequency} = 2\pi f = \text{radians per second}$$

but what do they signify in the case of a wave? We can recycle our previous definition simply by stating it in terms of the vibrations

that the wave causes as it passes a receiving instrument at a certain point in space. For a sound wave, this receiver could be an eardrum or a microphone. If the vibrations of the eardrum repeat themselves over and over, i.e., are periodic, then we describe the sound wave that caused them as periodic. Likewise we can define the period and frequency of a wave in terms of the period and frequency of the vibrations it causes. As another example, a periodic water wave would be one that caused a rubber duck to bob in a periodic manner as they passed by it.

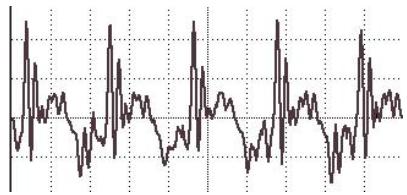
The period of a sound wave correlates with our sensory impression of musical pitch. A high frequency (short period) is a high note. The sounds that really define the musical notes of a song are only the ones that are periodic. It is not possible to sing a nonperiodic sound like “sh” with a definite pitch.

The frequency of a light wave corresponds to color. Violet is the high-frequency end of the rainbow, red the low-frequency end. A color like brown that does not occur in a rainbow is not a periodic light wave. Many phenomena that we do not normally think of as light are actually just forms of light that are invisible because they fall outside the range of frequencies our eyes can detect. Beyond the red end of the visible rainbow, there are infrared and radio waves. Past the violet end, we have ultraviolet, x-rays, and gamma rays.

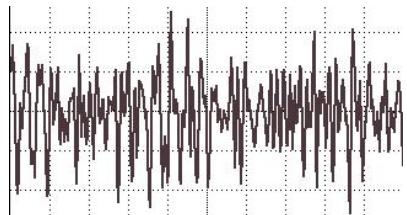
2.4.2 Graphs of waves as a function of position

Some waves, like sound waves, are easy to study by placing a detector at a certain location in space and studying the motion as a function of time. The result is a graph whose horizontal axis is time. With a water wave, on the other hand, it is simpler just to look at the wave directly. This visual snapshot amounts to a graph of the height of the water wave as a function of position. Any wave can be represented in either way.

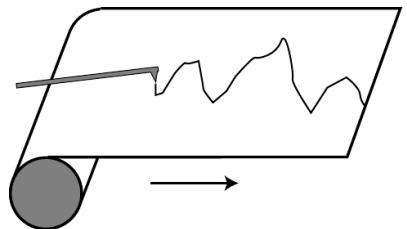
An easy way to visualize this is in terms of a strip chart recorder, an obsolescing device consisting of a pen that wiggles back and forth as a roll of paper is fed under it. It can be used to record a person’s electrocardiogram, or seismic waves too small to be felt as a noticeable earthquake but detectable by a seismometer. Taking the seismometer as an example, the chart is essentially a record of the ground’s wave motion as a function of time, but if the paper was set to feed at the same velocity as the motion of an earthquake wave, it would also be a full-scale representation of the profile of the actual wave pattern itself. Assuming, as is usually the case, that the wave velocity is a constant number regardless of the wave’s shape, knowing the wave motion as a function of time is equivalent to knowing it as a function of position.



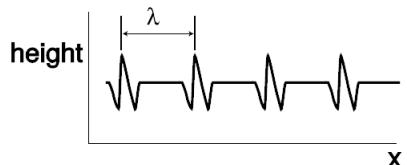
o / A graph of pressure versus time for a periodic sound wave, the vowel “ah.”



p / A similar graph for a non-periodic wave, “sh.”



q / A strip chart recorder.



r / A water wave profile created by a series of repeating pulses.

2.4.3 Wavelength

Any wave that is periodic will also display a repeating pattern when graphed as a function of position. The distance spanned by one repetition is referred to as one wavelength. The usual notation for wavelength is λ , the Greek letter lambda. Wavelength is to space as period is to time.

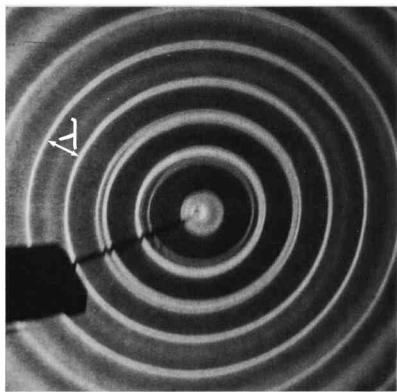
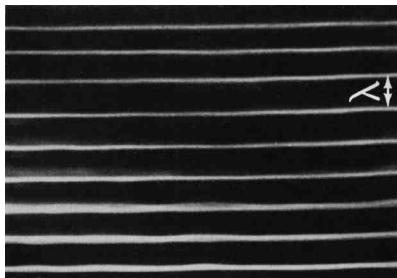
2.4.4 Wave velocity related to frequency and wavelength

Suppose that we create a repetitive disturbance by kicking the surface of a swimming pool. We are essentially making a series of wave pulses. The wavelength is simply the distance a pulse is able to travel before we make the next pulse. The distance between pulses is λ , and the time between pulses is the period, T , so the speed of the wave is the distance divided by the time,

$$v = \lambda/T.$$

This important and useful relationship is more commonly written in terms of the frequency,

$$v = f\lambda.$$



s / Wavelengths of linear and circular waves.

Wavelength of radio waves

example 4

▷ The speed of light is 3.0×10^8 m/s. What is the wavelength of the radio waves emitted by KMHD, a station whose frequency is 89.1 MHz?

▷ Solving for wavelength, we have

$$\begin{aligned}\lambda &= v/f \\ &= (3.0 \times 10^8 \text{ m/s}) / (89.1 \times 10^6 \text{ s}^{-1}) \\ &= 3.4 \text{ m}\end{aligned}$$

The size of a radio antenna is closely related to the wavelength of the waves it is intended to receive. The match need not be exact (since after all one antenna can receive more than one wavelength!), but the ordinary “whip” antenna such as a car’s is 1/4 of a wavelength. An antenna optimized to receive KMHD’s signal would have a length of $(3.4 \text{ m})/4 = 0.85 \text{ m}$.

The equation $v = f\lambda$ defines a fixed relationship between any two of the variables if the other is held fixed. The speed of radio waves in air is almost exactly the same for all wavelengths and frequencies (it is exactly the same if they are in a vacuum), so there is a fixed relationship between their frequency and wavelength. Thus we can say either “Are we on the same wavelength?” or “Are we on the same frequency?”

A different example is the behavior of a wave that travels from a region where the medium has one set of properties to an area where the medium behaves differently. The frequency is now fixed, because otherwise the two portions of the wave would otherwise get out of step, causing a kink or discontinuity at the boundary, which would be unphysical. (A more careful argument is that a kink or discontinuity would have infinite curvature, and waves tend to flatten out their curvature. An infinite curvature would flatten out infinitely fast, i.e., it could never occur in the first place.) Since the frequency must stay the same, any change in the velocity that results from the new medium must cause a change in wavelength.

The velocity of water waves depends on the depth of the water, so based on $\lambda = v/f$, we see that water waves that move into a region of different depth must change their wavelength, as shown in figure u. This effect can be observed when ocean waves come up to the shore. If the deceleration of the wave pattern is sudden enough, the tip of the wave can curl over, resulting in a breaking wave.

2.4.5 Sinusoidal waves

Sinusoidal waves are the most important special case of periodic waves. In fact, many scientists and engineers would be uncomfortable with defining a waveform like the “ah” vowel sound as having a definite frequency and wavelength, because they consider only sine waves to be pure examples of a certain frequency and wavelengths. Their bias is not unreasonable, since the French mathematician Fourier showed that any periodic wave with frequency f can be constructed as a superposition of sine waves with frequencies $f, 2f, 3f, \dots$. In this sense, sine waves are the basic, pure building blocks of all waves. (Fourier’s result so surprised the mathematical community of France that he was ridiculed the first time he publicly presented his theorem.)

However, what definition to use is really a matter of convenience. Our sense of hearing perceives any two sounds having the same period as possessing the same pitch, regardless of whether they are sine waves or not. This is undoubtedly because our ear-brain system evolved to be able to interpret human speech and animal noises, which are periodic but not sinusoidal. Our eyes, on the other hand, judge a color as pure (belonging to the rainbow set of colors) only if it is a sine wave.

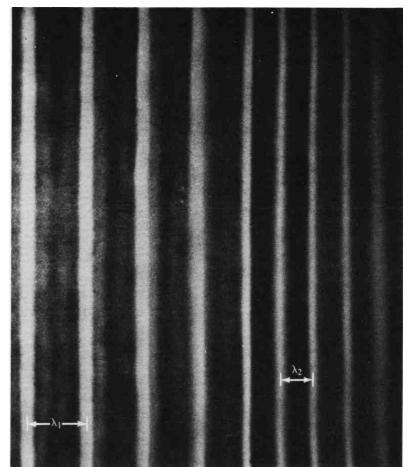
2.4.6 The wavenumber

When discussing vibrations, it’s conventional to define both a frequency f (cycles per second) and a frequency $\omega = 2\pi f$ (radians per second). It’s convenient to work with ω because it’s easier to write expressions like $\sin \omega t$ than $\sin 2\pi ft$, and $\omega = \sqrt{k/m}$ rather than $f = (1/2\pi)\sqrt{k/m}$ for the frequency of a harmonic oscillator.

For similar reasons, when discussing sinusoidal waves, we define



t / Ultrasound, i.e., sound with frequencies higher than the range of human hearing, was used to make this image of a fetus. The resolution of the image is related to the wavelength, since details smaller than about one wavelength cannot be resolved. High resolution therefore requires a short wavelength, corresponding to a high frequency.



u / A water wave traveling into a region with different depth will change its wavelength.

a variable k , called the wavenumber, which is the number of radians per meter, $k = 2\pi/\lambda$. The wavenumber has the same relationship to the wavelength as the angular frequency ω has to the period. In terms of the wavenumber, we have a compact way of writing the amplitude as a function of position and time for a traveling sinusoidal wave,

$$u = A \sin(kx - \omega t + \delta),$$

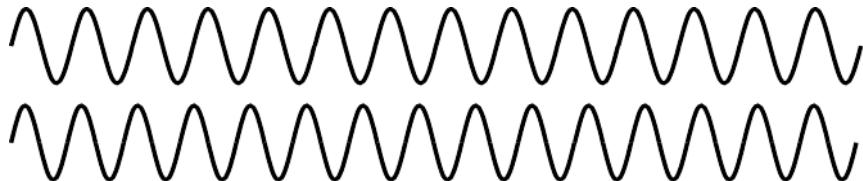
where δ is a constant phase. The wavenumber has units of radians per meter, and since radians aren't really units, this is the same as m^{-1} . When the direction of propagation is not limited to the x axis, we can generalize k to make a vector, the wave-vector, p. 51. The wavenumber is the x component of that vector, and a positive wavenumber indicates propagation in the positive x direction.

If we surf the wave while it moves forward one wavelength, then the kx term in the equation above increases by 2π . Since the input to the sine function (the phase of the wave) stays constant for the surfer, ωt must also change by 2π during this time. With some simple algebra (problem 5, p. 58), we find that the speed of the wave (phase velocity, not group velocity) is

$$v = \frac{\omega}{k}.$$

Discussion question

A Suppose we superimpose two sine waves with equal amplitudes but slightly different frequencies, as shown in the figure. What will the superposition look like? What would this sound like if they were sound waves?



Discussion question A.

2.5 The Doppler effect

Figure v shows the wave pattern made by the tip of a vibrating rod which is moving across the water. If the rod had been vibrating in one place, we would have seen the familiar pattern of concentric circles, all centered on the same point. But since the source of the waves is moving, the wavelength is shortened on one side and lengthened on the other. This is known as the Doppler effect.

Note that the velocity of the waves is a fixed property of the medium, so for example the forward-going waves do not get an extra boost in speed as would a material object like a bullet being shot forward from an airplane.

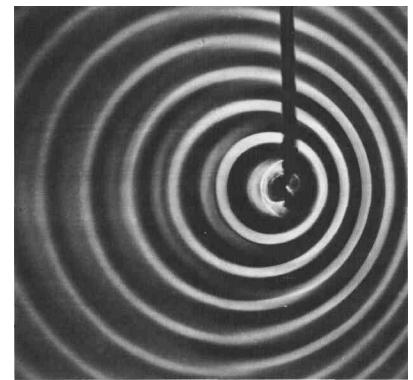
We can also infer a change in frequency. Since the velocity is constant, the equation $v = f\lambda$ tells us that the change in wavelength must be matched by an opposite change in frequency: higher frequency for the waves emitted forward, and lower for the ones emitted backward. The frequency Doppler effect is the reason for the familiar dropping-pitch sound of a race car going by. As the car approaches us, we hear a higher pitch, but after it passes us we hear a frequency that is lower than normal.

The Doppler effect will also occur if the observer is moving but the source is stationary. For instance, an observer moving toward a stationary source will perceive one crest of the wave, and will then be surrounded by the next crest sooner than she otherwise would have, because she has moved toward it and hastened her encounter with it. Roughly speaking, the Doppler effect depends only the relative motion of the source and the observer, not on their absolute state of motion (which is not a well-defined notion in physics) or on their velocity relative to the medium.

Restricting ourselves to the case of a moving source, and to waves emitted either directly along or directly against the direction of motion, we can easily calculate the wavelength, or equivalently the frequency, of the Doppler-shifted waves. Let u be the velocity of the source. The wavelength of the forward-emitted waves is shortened by an amount uT equal to the distance traveled by the source over the course of one period. Using the definition $f = 1/T$ and the equation $v = f\lambda$, we find for the wavelength λ' of the Doppler-shifted wave the equation

$$\lambda' = \left(1 - \frac{u}{v}\right) \lambda.$$

A similar equation can be used for the backward-emitted waves, but with a plus sign rather than a minus.



v / The pattern of waves made by a point source moving to the right across the water. Note the shorter wavelength of the forward-emitted waves and the longer wavelength of the backward-going ones.

Doppler-shifted sound from a race car

example 5

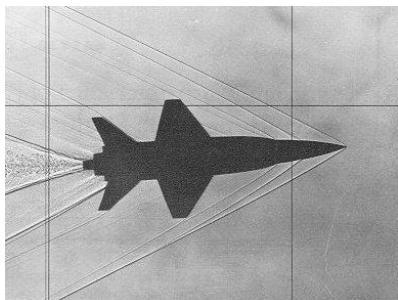
▷ If a race car moves at a velocity of 50 m/s, and the velocity of sound is 340 m/s, by what percentage are the wavelength and frequency of its sound waves shifted for an observer lying along its line of motion?

▷ For an observer whom the car is approaching, we find

$$1 - \frac{u}{v} = 0.85,$$

so the shift in wavelength is 15%. Since the frequency is inversely proportional to the wavelength for a fixed value of the speed of sound, the frequency is shifted upward by

$$1/0.85 = 1.18,$$



w / Sound waves are created by the X-15 rocket plane, flying at 3.5 times the speed of sound.

i.e., a change of 18%. (For velocities that are small compared to the wave velocities, the Doppler shifts of the wavelength and frequency are about the same.)

Discussion questions

A If an airplane travels at exactly the speed of sound, what would be the wavelength of the forward-emitted part of the sound waves it emitted? How should this be interpreted, and what would actually happen? What happens if it's going faster than the speed of sound? Sketch spherical wavefronts in the style of figure v.

B If bullets go slower than the speed of sound, why can a supersonic fighter plane catch up to its own sound, but not to its own bullets?

C If someone inside a plane is talking to you, will their speech be Doppler shifted?

2.6 Reflection and standing waves

2.6.1 Reflection of waves

Reflection of light from a mirror and echoes of sound are familiar examples of the phenomenon of wave reflection. In ch. 5 we will see in more mathematical detail that whenever a wave encounters a change of medium, it is partially reflected back into the original medium and partially transmitted into the new one.

Figure x shows a water wave being reflected from a wall. You've probably noticed this kind of thing when taking a bath. This example is particularly simple because the medium simply ends at the wall. Therefore there is no transmitted wave, and 100% of the incident energy is reflected.

Reflections can sometimes be *inverted*, meaning that the amplitude reverses its sign. We will discuss this in more detail in sec. 5.3.3.

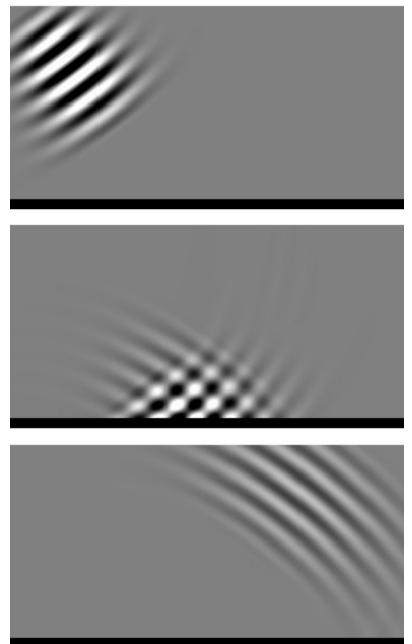
2.6.2 Standing waves

Figure z shows sinusoidal wave patterns made by shaking a rope. I used to enjoy doing this at the bank with the pens on chains, back in the days when people actually went to the bank. You might think that I and the person in the photos had to practice for a long time in order to get such nice sine waves. In fact, a sine wave is the only shape that can create this kind of wave pattern, called a standing wave, which simply vibrates back and forth in one place without moving. The sine wave just creates itself automatically when you find the right frequency, because no other shape is possible.

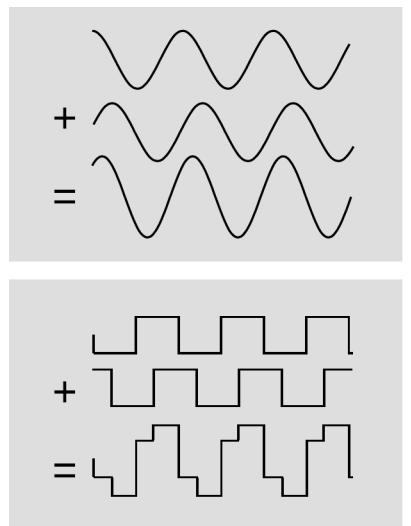
If you think about it, it's not even obvious that sine waves should be able to do this trick. After all, waves are supposed to travel at a set speed, aren't they? The speed isn't supposed to be zero! Well, we can actually think of a standing wave as a superposition of a moving sine wave with its own reflection, which is moving the opposite way. Sine waves have the unique mathematical property, y, that the sum of sine waves of equal wavelength is simply a new sine wave with the same wavelength. As the two sine waves go back and forth, they always cancel perfectly at the ends, and their sum appears to stand still. As each wave gets to the end of the rope, it is reflected, so our supply of waves is always being replenished.

Standing wave patterns are rather important, since atoms are really standing-wave patterns of electron waves. You are a standing wave!

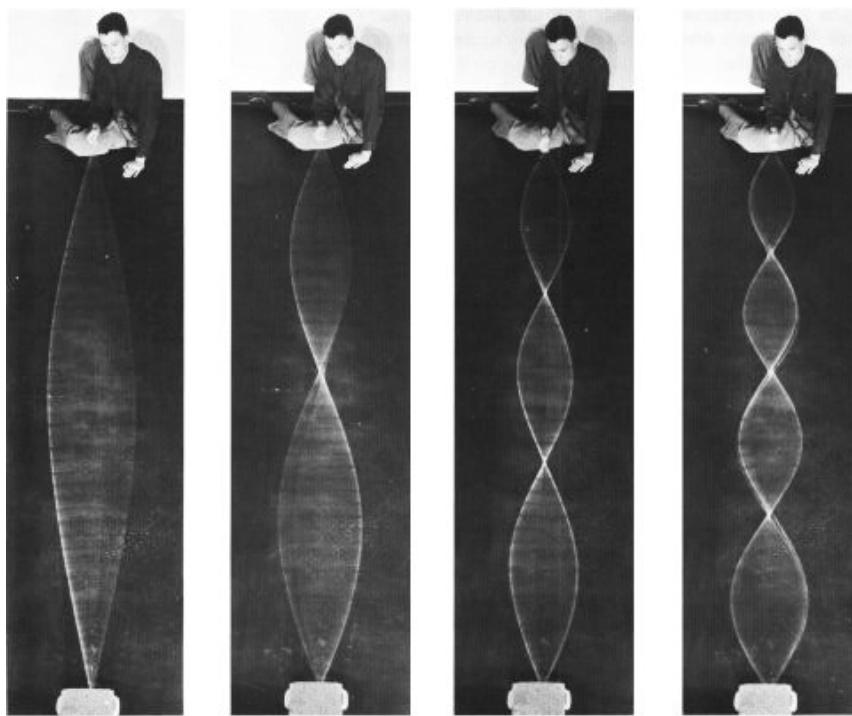
In this chapter we will consider only the standing-wave patterns obtained when the wave's amplitude is zero at the ends, which is the case for an example like figure z because the spring is fixed at the ends. We will consider more general standing-wave patterns in



x / A simulation of a water wave being reflected from a wall.



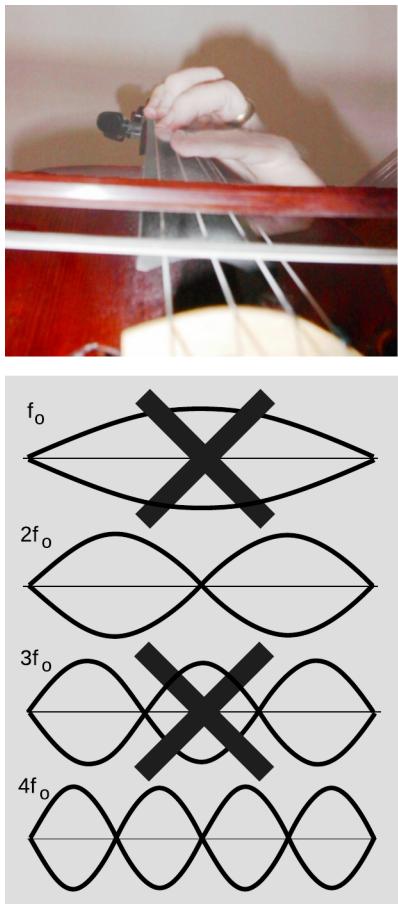
y / Sine waves add to make sine waves. Other functions don't have this property.



z / Standing waves on a spring.

one dimension in sec. 5.4.

The wavelength in a standing wave can only have a certain set of values. If the velocity is fixed, then this gives a list of possible frequencies. For the standing wave patterns described in this chapter, the pattern of these frequencies is $f_0, 2f_0, 3f_0, \dots$ (We'll see in sec. 5.4 that for other types of standing wave patterns, the pattern of frequencies can be different from this.) The frequencies in this set are called the *harmonics* of f_0 . If you walk up to the spring in figure *z* and distort it into some random shape, then by Fourier's theorem (p. 45), this can be analyzed into a superposition of standing wave patterns. When you release the spring, it will begin vibrating in a *superposition* of all of the harmonic frequencies. This happens with all musical instruments, and also with the human voice, but the ear-brain system is adapted to recognizing the pattern and perceives it as a single fused sensation, not as separate musical notes. The relative strengths of the harmonics is what makes vowels sound different from each other, and is also one of the factors that makes one instrument sound different from another. (Other factors are actually usually more important.)



aa / Example 6.

Harmonics on string instruments

example 6

Figure *aa* shows a violist playing what string players refer to as a natural harmonic. The term "harmonic" is used here in a somewhat different sense than in physics. The musician's pinkie is pressing very lightly against the string — not hard enough to

make it touch the fingerboard — at a point precisely at the center of the string's length. As shown in the diagram, this allows the string to vibrate at frequencies $2f_0, 4f_0, 6f_0, \dots$, which have stationary points at the center of the string, but not at the odd multiples $f_0, 3f_0, \dots$. Since all the overtones are multiples of $2f_0$, the ear perceives $2f_0$ as the basic frequency of the note. In musical terms, doubling the frequency corresponds to raising the pitch by an octave.

2.7 Waves in two or three dimensions

2.7.1 The wave-vector

Up until now we've concentrated on waves in one dimension, a category that is elastic enough to include examples like a plane wave or a transverse wave on a string, since they propagate in a single direction. As a simple example of a wave that really does have to be treated as fully two-dimensional, let's consider a square, elastic membrane, figure ab, with sides of length b . Vibrations like this can be made with soap bubble films.

The example I've drawn has an upward bulge at the back left side, and a trough at the front right. It is visually plausible to say that this wave has two different wavelengths, b along the x axis and $2b$ along y . If we ignore the time dependence, then the shape of this wave can be written as

$$u = A \sin\left(\frac{2\pi}{b}x\right) \sin\left(\frac{2\pi}{2b}y\right),$$

where u is the height of the membrane at a certain point.

To make the writing easier, we generalize the wavenumber (p. 45) to define a vector quantity \mathbf{k} , called the wave-vector. In terms of \mathbf{k} 's components, our example becomes

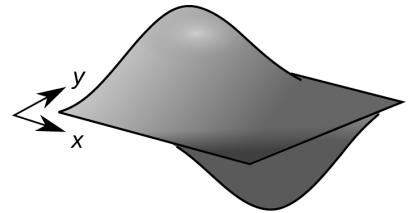
$$u = A \sin k_x x \sin k_y y,$$

where $k_x = 2\pi/b$ and $k_y = 2\pi/2b$.

A traveling sine wave has a beautifully simple form when written in terms of the wave-vector,

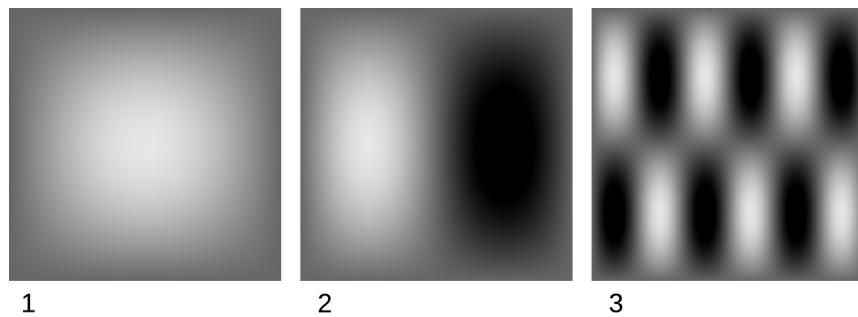
$$u = A \sin(\mathbf{k} \cdot \mathbf{r} - \omega t + \delta),$$

where $\mathbf{r} = (x, y)$ is the position vector, and δ is a constant phase. The vector \mathbf{k} points in the direction of propagation, and its magnitude is the wavenumber k . We recall that $v = \omega/k$ (which is the phase velocity, not the group velocity).



ab / A standing wave pattern on an elastic membrane, attached at its edges on a square frame.

ac / Three standing-wave patterns of a square membrane, example 7 and problem 17, p. 60.



Standing waves on a square membrane example 7

▷ Figure ac/2 is a more economical way to represent the pattern of vibrations originally drawn in figure ab, and the other two panels of the figure are to be interpreted similarly. Let ω_1 , ω_2 , and ω_3 be the frequencies of vibration. Find ω_2/ω_1 .

▷ The wave-vector in ac/2 was found earlier. Call this \mathbf{k}_2 . To cut down on writing, let $k_0 = 2\pi/2b$, so that in x - y component form we have $\mathbf{k}_2 = k_0(2, 1)$, with magnitude

$$k_2 = k_0 \sqrt{5}.$$

A similar calculation for ac/1 gives

$$k_1 = k_0 \sqrt{2}.$$

The result is

$$\frac{\omega_2}{\omega_1} = \frac{k_2 v}{k_1 v} = \sqrt{\frac{5}{2}}.$$

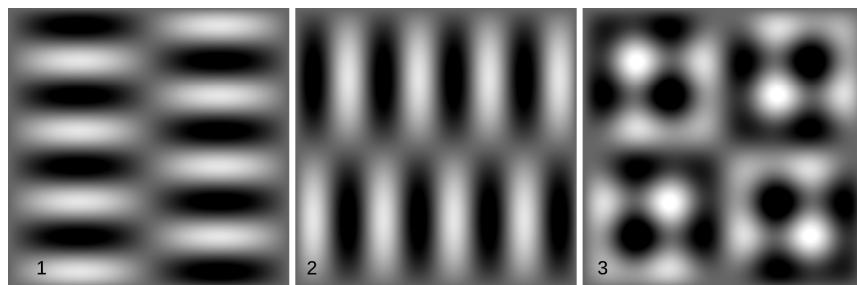
The calculation of ω_3/ω_1 is problem 17.

2.7.2 Degeneracy

On the square membrane we had the standing-wave pattern , but we could equally well have drawn , which is the same pattern turned 90 degrees. Because of the square's symmetrical shape, these will have the same frequency. When the frequencies of two wave patterns coincide like this, we call them *degenerate*. Not coincidentally, we say in geometry that all squares are rectangles, but a square is a *degenerate* rectangle. A “generic” or “general” rectangle has unequal sides. You have probably had some exposure in a chemistry class to the idea that an electron is a wave. Because of the perfect symmetry of an atom, we frequently get perfectly degenerate electron-wave patterns. In the language used there, would be analogous to a p orbital oriented along the x axis, while would be oriented along y .

2.7.3 Separability

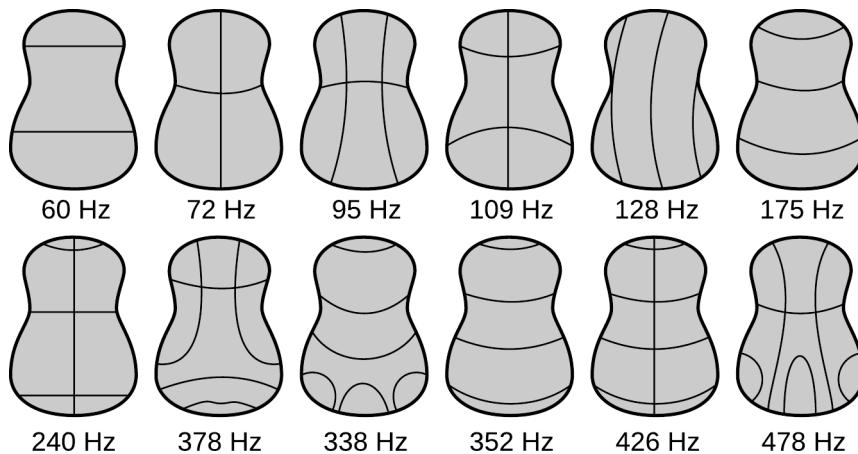
The wave patterns that we have been discussing on the square membrane are special because they can be written as the product of two functions, a function of x and a function of y . Such a wave is called *separable*. For a pattern like , we can draw any horizontal line across the square, and the behavior of the function along that line is always the same regardless of which section we choose. It will be a two-hump wave like , multiplied by some constant. Similarly, the vertical sections all look like a one-hump wave.



ad / Patterns 1 and 2 are separable. Pattern 3, a superposition of 1 and 2, is not.

Figure ad shows that the superposition of two separable waves does not have to be separable.

2.7.4 The Laplacian



ae / Standing-wave patterns on a guitar. Each pattern is labeled with its frequency $f = \omega/2\pi$.



af / A car crashes into the corner of a building. Waves from the impact propagate toward the back of the car through the sheet metal.

Sometimes we get waves that simply aren't sine waves. An example is figure ae, where the shape of the guitar's body prevents simple patterns like the ones in figure ac from existing. Similar examples from civil and mechanical engineering are the wave disturbances in structures such as bridges and cars, which wiggle and vibrate in complicated ways when subjected to forces (figure af). A third type of example is the electron waves that occur inside atoms. In all these cases, we can't just write down a single wave-vector for

the whole wave pattern, and we have to solve some kind of wave equation. A typical wave equation in *one* dimension is the equation we found for waves on a string, p. 54, which can be written as

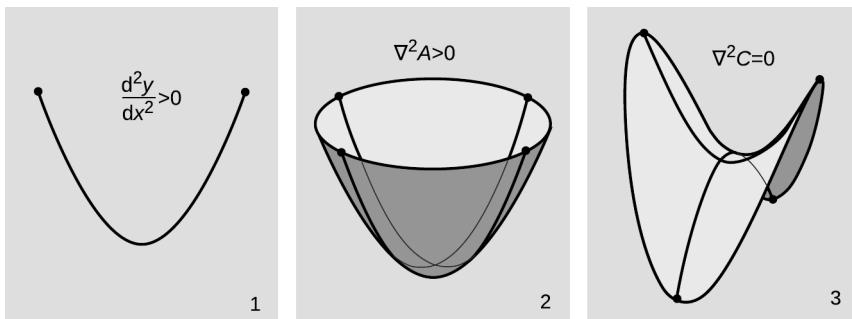
$$\text{acceleration} = v^2 \frac{\partial^2 u}{\partial x^2}.$$

This is basically an expression of Newton's second law. The left-hand side is the acceleration, and the right-hand side is the unbalanced force that acts because of the curvature of the string (figure m, p. 38). We would get waves that were not sine waves when v^2 was not constant. In more than one dimension, we need to replace $\partial^2 u / \partial x^2$ with

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

(adding a third term for z if the wave was in three dimensions). This operation on the function u is notated $\nabla^2 u$, and the derivative-like operator $\nabla^2 = \partial^2 / \partial x^2 + \partial^2 / \partial y^2$ is called the Laplacian. It occurs in many places in physics. For example, in electrostatics, the potential ϕ in a region of vacuum must be a solution of the equation $\nabla^2 \phi = 0$. Like the second derivative, the Laplacian is essentially a measure of curvature. Or, as shown in figure ag, we can think of it as a measure of how much the value of a function at a certain point differs from the average of its value on nearby points (which amounts to the same thing).

In a simulation like the one in figure af, a computer breaks down a continuous body such as a piece of sheet metal into a grid of squares or triangles. It can find a numerical approximation to the Laplacian by finding the difference between the position of one node on the grid and the average position of its neighboring nodes. This produces a prediction of the acceleration, which is then used to evolve the motion forward in time as the simulation goes on.



ag / 1. The one-dimensional version of the Laplacian is the second derivative. It is positive here because the average of the two nearby points is greater than the value at the center. 2. The Laplacian of the function A in example 8 is positive because the average of the four nearby points along the perpendicular axes is greater than the function's value at the center. 3. $\nabla^2 C = 0$. The average is the same as the value at the center.

Examples of the Laplacian in two dimensions *example 8*

- ▷ Compute the Laplacians of the following functions in two dimensions, and interpret them: $A = x^2 + y^2$, $B = -x^2 - y^2$, $C = x^2 - y^2$.
- ▷ The first derivative of function A with respect to x is $\partial A / \partial x = 2x$. Since y is treated as a constant in the computation of the partial derivative $\partial / \partial x$, the second term goes away. The second derivative of A with respect to x is $\partial^2 A / \partial x^2 = 2$. Similarly we have $\partial^2 A / \partial y^2 = 2$, so $\nabla^2 A = 4$.

All derivative operators, including ∇^2 , have the linear property that multiplying the input function by a constant just multiplies the output function by the same constant. Since $B = -A$, and we have $\nabla^2 B = -4$.

For function C , the x term contributes a second derivative of 2, but the y term contributes -2 , so $\nabla^2 C = 0$.

The interpretation of the positive sign in $\nabla^2 A = 4$ is that A 's graph is shaped like a trophy cup, and the cup is concave up. $\nabla^2 B < 0$ is because B is concave down. Function C is shaped like a saddle. Since its curvature along one axis is concave up, but the curvature along the other is down and equal in magnitude, the function is considered to have zero concavity over all.

A three-dimensional example: the probability cloud in hydrogen *example 9*

You have probably been exposed in a chemistry class to the idea that electrons are waves, and that the waves form probability clouds in atoms. These are ideas that we will take up in detail at the end of this course. In this connection, it is natural to con-

sider the Laplacian of the function

$$\Psi = ue^{-r/a},$$

where $r = \sqrt{x^2 + y^2 + z^2}$ is the distance from the origin. For a hydrogen atom in its ground state, this turns out to be the form of the electron's wave (example 5, p. 401). In the following, the result $\partial r / \partial x = x/r$ comes in handy. Computing the partial derivatives that occur in the Laplacian, we obtain for the x term

$$\begin{aligned}\frac{\partial \Psi}{\partial x} &= \frac{\partial \Psi}{\partial r} \frac{\partial r}{\partial x} \\ &= -\frac{x}{ar} \Psi \\ \frac{\partial^2 \Psi}{\partial x^2} &= -\frac{1}{ar} \Psi - \frac{x}{a} \left(\frac{\partial}{\partial x} \frac{1}{r} \right) \Psi + \left(\frac{x}{ar} \right)^2 \Psi \\ &= -\frac{1}{ar} \Psi + \frac{x^2}{ar^3} \Psi + \left(\frac{x}{ar} \right)^2 \Psi,\end{aligned}$$

so

$$\nabla^2 \Psi = \left(-\frac{2}{ar} + \frac{1}{a^2} \right) \Psi.$$

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 The musical note middle C has a frequency of 262 Hz. What are its period and wavelength? ✓

2 Singing that is off-pitch by more than about 1% sounds bad. How fast would a singer have to be moving relative to the rest of a band to make this much of a change in pitch due to the Doppler effect? ✓

3 The following is a graph of the height of a water wave as a function of *position*, at a certain moment in time.



Trace this graph onto another piece of paper, and then sketch below it the corresponding graphs that would be obtained if

- the amplitude and frequency were doubled while the velocity remained the same;
- the frequency and velocity were both doubled while the amplitude remained unchanged;
- the wavelength and amplitude were reduced by a factor of three while the velocity was doubled.

Explain all your answers. [Problem by Arnold Arons.]

4 (a) The graph shows the height of a water wave pulse as a function of position. Draw a graph of height as a function of time for a specific point on the water. Assume the pulse is traveling to the right.

- Repeat part a, but assume the pulse is traveling to the left.
- Now assume the original graph was of height as a function of time, and draw a graph of height as a function of position, assuming the pulse is traveling to the right.
- Repeat part c, but assume the pulse is traveling to the left.

Explain all your answers. [Problem by Arnold Arons.]



Problem 4

5 Complete the algebra on p. 46 leading to the relation $v = \omega/k$. Check that the units of this relationship make sense.

6 At a particular moment in time, a wave on a string has a shape described by $y = 3.5 \cos(0.73\pi x + 0.45\pi t + 0.37\pi)$. The stuff inside the cosine is in radians. Assume that the units of the numerical constants are such that x , y , and t are in SI units. \triangleright Hint, p. 443
(a) Is the wave moving in the positive x or the negative x direction?
(b) Find the wave's period, frequency, wavelength.
(c) Find the wave's velocity.
(d) Find the maximum velocity of any point on the string, and compare with the magnitude and direction of the wave's velocity.

✓

7 Near the cloud-tops of Jupiter, sound waves have a velocity of about 950 m/s. A sound wave with a frequency $f = 100$ Hz is traveling in this environment in the positive x direction. Calculate the wavenumber k and frequency ω , and write an equation for the wave's amplitude u as a function of x and t . \triangleright Solution, p. 443



Problem 8.

8 The figure shows one wavelength of a steady sinusoidal wave traveling to the right along a string. Define a coordinate system in which the positive x axis points to the right and the positive y axis up, such that the flattened string would have $y = 0$. Copy the figure, and label with $y = 0$ all the appropriate parts of the string. Similarly, label with $v = 0$ all parts of the string whose velocities are zero, and with $a = 0$ all parts whose accelerations are zero. There is more than one point whose velocity is of the greatest magnitude. Pick one of these, and indicate the direction of its velocity vector. Do the same for a point having the maximum magnitude of acceleration. Explain all your answers.

[Problem by Arnold Arons.]

9 (a) Find an equation for the relationship between the Doppler-shifted frequency of a wave and the frequency of the original wave, for the case of a stationary observer and a source moving directly toward or away from the observer.

✓

- (b) Check that the units of your answer make sense.
(c) Check that the dependence on v_s makes sense.

10 Suggest a quantitative experiment to look for any deviation from the principle of superposition for surface waves in water. Try to make your experiment simple and practical.

- 11** (a) On a typical harp, the string for the note A (the one just above middle C) has a length of 41 cm. Find the wavelength of the lowest-frequency standing wave. \checkmark
- (b) The frequency of this standing wave pattern is 440 Hz. Find the speed of waves on this string. \checkmark
- (c) The nylon string used for this purpose has a mass per unit length $\mu = 1.1 \times 10^{-3}$ kg/m. Find the tension that must be applied to the string. \checkmark

12 Consider the standing-wave patterns of a string of length L that is fixed at both ends. The speed of waves on the string is v . Let N be the number of humps in the standing wave, i.e., the number of extrema of displacement.

- (a) Sketch the first few patterns and label them with their values of N .
- (b) As a warm-up with concrete numbers, consider the case where $L = 1$ m. Find the wavelengths of the patterns you drew in part a.
- (c) Find the wavelength λ in terms of L and N . Check that your equation reproduces the numbers from part b, and also that it reproduces the answer of problem 11a.
- (d) Find the frequency in terms of N , v , and L . \checkmark

13 In the following, x and y are variables, while u and v are constants. Compute (a) $\partial(ux \ln(vy))/\partial x$, (b) $\partial(ux \ln(vy))/\partial y$. \checkmark

14 Let $\Psi = e^{2x+y}$. Compute $\nabla^2\Psi$. [If you get $9e^{2x+y}$, then you've made the mistake described in problem 15.] \checkmark

15 (a) Consider the function defined by $f(x, y) = (x - y)^2$. Visualize the graph of this function as a surface. (This is a simple enough example that you should not have to resort to computer software.) Use this visualization to determine the behavior of the sign of the Laplacian, as in example 8 on p. 55.

(b) Consider the following incorrect calculation of this Laplacian. We take the first derivatives and find

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} = 0.$$

Next we take the second derivatives, but those are zero as well, so the Laplacian is zero. Critique this calculation in two ways: (1) by comparing with part a; (2) by comparing with a correct calculation.

Remark: In general, if we have a function f of two variables, the quantity $Q = \partial f / \partial x + \partial f / \partial y$ can never be of physical interest, because it doesn't behave in a sensible way when we rotate our coordinate axes. You may want to prove this by showing that by rotating your coordinate system, you can get a completely different answer than the one calculated in part b. \triangleright Solution, p. 443

16 Consider the traveling wave

$$u = A \sin(k_x x + k_y y - \omega t + \delta).$$

Calculate $\nabla^2 u$, and show that it is the same as $-k^2 u$, where $k = |\mathbf{k}|$ is the magnitude of the wave-vector.

17 Complete example 7 on p. 52 by finding ω_3 / ω_1 . \checkmark

18 As discussed in section 2.2, the speed at which a disturbance travels along a string under tension is given by $v = \sqrt{T/\mu}$, where μ is the mass per unit length, and T is the tension.

(a) Suppose a string has a density ρ , and a cross-sectional area A . Find an expression for the maximum tension that could possibly exist in the string without producing $v > c$, which is impossible according to relativity. Express your answer in terms of ρ , A , and c . The interpretation is that relativity puts a limit on how strong any material can be. \checkmark

(b) Every substance has a tensile strength, defined as the force per unit area required to break it by pulling it apart. The tensile strength is measured in units of N/m^2 , which is the same as the pascal (Pa), the mks unit of pressure. Make a numerical estimate of the maximum tensile strength allowed by relativity in the case where the rope is made out of ordinary matter, with a density on the same order of magnitude as that of water. (For comparison, kevlar has a tensile strength of about 4×10^9 Pa, and there is speculation that fibers made from carbon nanotubes could have values as high as 6×10^{10} Pa.) \checkmark

(c) A black hole is a star that has collapsed and become very dense, so that its gravity is too strong for anything ever to escape from it. For instance, the escape velocity from a black hole is greater than

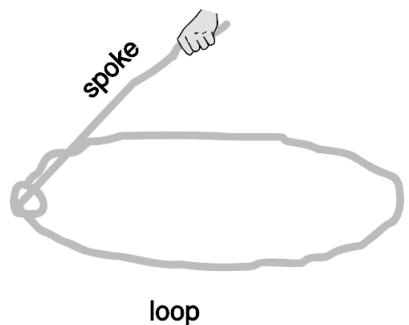
c , so a projectile can't be shot out of it. Many people, when they hear this description of a black hole in terms of an escape velocity, wonder why it still wouldn't be possible to extract an object from a black hole by other means. For example, suppose we lower an astronaut into a black hole on a rope, and then pull him back out again. Why might this not work?

19 The simplest trick with a lasso is to spin a flat loop in a horizontal plane. The whirling loop of a lasso is kept under tension mainly due to its own rotation. Although the spoke's force on the loop has an inward component, we'll ignore it. The purpose of this problem, which is based on one by A.P. French, is to prove a cute fact about wave disturbances moving around the loop. As far as I know, this fact has no practical implications for trick roping! Let the loop have radius r and mass per unit length μ , and let its angular velocity be ω .

(a) Find the tension, T , in the loop in terms of r , μ , and ω . Assume the loop is a perfect circle, with no wave disturbances on it yet.

▷ Hint, p. 443 ▷ Answer, p. 459 ✓

(b) Find the velocity of a wave pulse traveling around the loop. Discuss what happens when the pulse moves in the same direction as the rotation, and when it travels contrary to the rotation. ✓ *



Problem 19.



Chapter 3

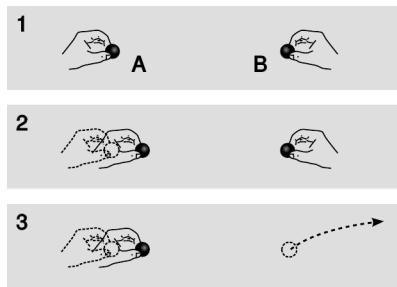
Electromagnetic waves

In ch. 13 we saw that, because of the relative nature of time, the universe cannot work according to Isaac Newton's vision of instantaneous action at a distance, as claimed by force laws like Coulomb's law $F = kq_1q_2/r^2$ and Newton's law of gravity $F = Gm_1m_2/r^2$. Instead, a "disturbance in the force" must propagate outward from the source as a wave. In ch. 31 we learned about general properties of waves, using mechanical waves as our examples because they were concrete and easy to conceptualize. We now turn to electromagnetic waves, which are one of the *fundamental* kinds of waves that the universe is ultimately composed of.

3.1 Energy and momentum of electric and magnetic fields

3.1.1 A thought experiment

The following argument shows that electric and magnetic fields must contain both energy (as you probably already knew) and momentum. In figure a/1, Andy and Bob hold positive charges A and B at some distance from one another. If Andy chooses to move his charge closer to Bob's, a/2, Andy will have to do some mechanical work against the electrical repulsion, burning off some of the calories from that chocolate cheesecake he had at lunch. This reduction in his body's chemical energy is offset by a corresponding increase in the electrical potential energy $q\Delta\phi$. Not only that, but Andy feels the resistance stiffen as the charges get closer together and the repulsion strengthens. He has to do a little extra work, but this is all properly accounted for in the electrical potential energy.



a / Fields carry energy.

Furthermore, we know that Newton's third law holds for the ordinary contact forces, such as normal and frictional forces, with which we have experience in mechanics. Therefore when Andy's hand makes a force on his charge, the charge makes a force back on his hand that is equal in strength and opposite in direction. In the mechanical context that we have experience with, this balance of forces is what ensures conservation of momentum. Andy recoils slightly, absorbing some leftward momentum from the charge. (This momentum may be transmitted through his feet to the earth, which then recoils at some very small velocity.)

But now suppose, a/3, that Bob decides to play a trick on Andy by tossing charge B far away just as Andy is getting ready to move charge A. We have already established that Andy can't feel charge B's motion instantaneously, so the force on B must actually be propagated through some kind of "force ripples." Of course this experiment is utterly impractical, but suppose for the sake of argument that the time it takes the effect to propagate across the diagram is long enough so that Andy can complete his motion before he feels the effect of B's disappearance. He is still getting stale information about B's position. As he moves A to the right, he feels a repulsion, because the field in his region of space is still the field caused by B in its *old* position. He has burned some chocolate cheesecake calories, and it appears that conservation of energy has been violated, because these calories can't be properly accounted for by any interaction with B, which is long gone. Conservation of momentum is also violated.

If we hope to preserve the laws of conservation of energy and momentum, then the only possible conclusion is that the field ripples themselves carry away the cheesecake energy, as well as the momentum. In fact, this example represents an impractical method

of transmitting radio waves. Andy does work on charge A, and that energy goes into the radio waves. Even if B had never existed, the radio waves would still have carried energy and momentum.

3.1.2 Expressions for the energy and momentum density

From your previous study of electromagnetism, you know that a time-varying electric field will induce a curly magnetic field, so in an electromagnetic wave we expect both an \mathbf{E} and a \mathbf{B} to exist. In fact, if we hadn't already known this fact about induction, we could have inferred it at this point. In Andy and Bob's experiment, we should be able to look at the empty space in the middle of the diagram, measure the fields, and infer not just their energy content, which is a scalar, but also their momentum and direction of propagation, which have directions in space. If only an electric field existed, then the only direction we could infer would be the direction of \mathbf{E} , but that can't be the right way to tell which direction the wave is traveling. For example, we could do the experiment twice, once with charges A and B both positive, then again with both charges negative. In both cases the wave should propagate to the right, but all the electric fields would have flipped.

This problem is resolved by the presence of the magnetic field. Given two vectors \mathbf{E} and \mathbf{B} , we can form the vector cross product $\mathbf{E} \times \mathbf{B}$. This is the *only* such vector we can form, so it must be the one that tells us the direction of the wave's momentum and the direction it's going. We thus infer that there is no such thing as an electric wave or a magnetic wave, only electromagnetic waves containing both fields.

self-check A

Why can't we use $\mathbf{E} \times \mathbf{E}$ or $\mathbf{B} \times \mathbf{B}$ for this purpose? \triangleright Answer, p. 455

Since energy is a scalar, similar arguments lead to the conclusion that the energy content of the fields must depend on $\mathbf{E} \cdot \mathbf{E}$ and $\mathbf{B} \cdot \mathbf{B}$, i.e., on the squared magnitudes of the fields. (We could in principle have an energy that went like $\mathbf{E} \cdot \mathbf{B}$, but this would lead to effects that we don't observe, such as the deflection of a magnetic compass when we place it next to a battery.)

The energy and momentum densities go like this:

$$\begin{aligned} dU_E &= \frac{1}{8\pi k} E^2 dv \\ dU_B &= \frac{c^2}{8\pi k} B^2 dv \\ d\mathbf{p} &= \frac{1}{4\pi k} \mathbf{E} \times \mathbf{B} dv \end{aligned}$$

Here U stands for energy (to avoid a notational clash with \mathbf{E} for the electric field), k is the Coulomb constant (sometimes also written as $1/4\pi\epsilon_0$), c is the speed of light, and v indicates volume. The “d” notation looks like the Leibniz notation for a derivative, but these are

not derivatives. In these expressions d just means “a little bit of.” The reason for the d ’s is that in general the fields are nonuniform, so that we can’t speak of “the” value of E^2 or B^2 over some large volume. Only by taking an infinitesimal volume near one point can we speak of the field as having a definite value. The quantity dU is then understood as the infinitesimal energy contained within this volume.

We have already justified the structure of these expressions, but not the constant factors in front. The factors of $1/k$ and c^2 have to be there because of units. In natural units, the c^2 wouldn’t exist. Its presence here tells us that there is a relativistic link between **E** and **B**. We’ll see this in more detail later, but it’s not particularly surprising. For example, an electric charge has only an **E** field according to an observer in the frame of reference where the charge is at rest, but observers in other frames will say that the charge is moving, and therefore it will have a **B** field as well.

3.1.3 Examples of the momentum

You have probably already learned about U_E and U_B , and also experienced them directly. For example, in an oscillating LC circuit, these are the energies that are being shuffled back and forth between the inductor and the capacitor. But you have probably never directly experienced the momentum of the fields.

It is in fact plausible that the proportionality constant occurring in the equation for the momentum density is such that the momentum of light is too small to notice in everyday life. For material objects moving at speeds small compared to c , the kinetic energy and momentum are given by $K = (1/2)mv^2$ and $p = mv$, so that the ratio of momentum to energy is $p/K = 2/v$. Therefore objects moving very fast have very little momentum in proportion to their energy. We see this, for example, in an old-fashioned CRT television tube, in which the electron beam moves at extremely high speeds (perhaps 10^6 m/s); the energy is enough to make a bright image on the screen, but the device doesn’t recoil from the beam’s momentum when we turn it on, nor does it shake and rattle as the beam is steered back and forth across the screen to paint the picture. Although the equations above do not actually hold in detail for light (the final result ends up being off by a factor of 2, as shown in sec. 3.5.1, p. 73), it still makes sense that the momentum-to-energy ratio is extremely small, because the speed, c , is so big.

A comet’s tail

example 1

Halley’s comet, shown in figure b, has a very elongated elliptical orbit, like those of many other comets. About once per century, its orbit brings it close to the sun. The comet’s head, or nucleus, is composed of dirty ice, so the energy deposited by the intense sunlight gradually removes ice from the surface and turns it into water vapor.

The sunlight does not just carry energy, however. If it only carried energy, then the water vapor would just form a spherical halo that would surround the nucleus and travel along with it. The light also carries momentum. Once the steam comes off, the momentum of the sunlight impacting on it pushes it away from the sun, forming a tail as shown in the top image. (Some comets also have a second tail, which is propelled by electrical forces rather than by the momentum of sunlight.)

The Nichols radiometer

example 2

Figure c shows a simplified drawing of the 1903 experiment by Nichols and Hull that verified the predicted momentum of light waves. Two circular mirrors were hung from a fine quartz fiber, inside an evacuated bell jar. A 150 mW beam of light was shone on one of the mirrors for 6 s, producing a tiny rotation, which was measurable by an optical lever (not shown). The force was within 0.6% of the theoretically predicted value of 0.001 μN . For comparison, a short clipping of a human hair weighs $\sim 1 \mu\text{N}$.

The hydrogen bomb

example 3

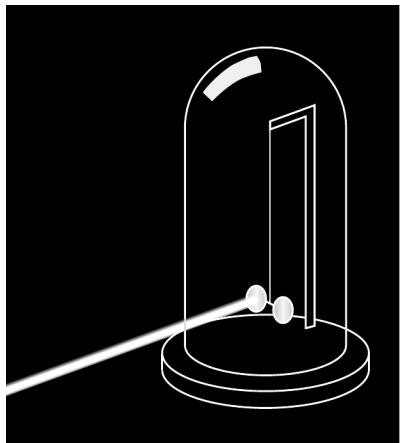
The technological feasibility of the hydrogen bomb was considered uncertain for some time after the end of World War II. The general idea was to use a fission bomb to implode hydrogen fuel and create conditions of high temperature and density in order to initiate nuclear fusion reactions. If a few properties of certain nuclei had been slightly different, the human race might not have been afflicted with this weapon. The first successful design concept was created in 1951 by Stanislaw Ulam and Edward Teller, both of them Jewish refugees whose moral and political calculus analogized Stalin to Hitler. A crucial trick was the use of radiation pressure from x-rays to implode the hydrogen fuel. Although this pressure was smaller than the pressure of the imploding material particles, the radiation traveled faster and got to the fuel first.

Because the momentum of light waves is so small in cases like examples 1 and 2, one might wonder why we should even bother discussing it. Is it purely an impractical and theoretical consideration? The answer is that it is very practical in the sense that it helps us to understand important practical facts about these waves. One such fact is that, as we have already seen, wave disturbances in the field must be both electric and magnetic.

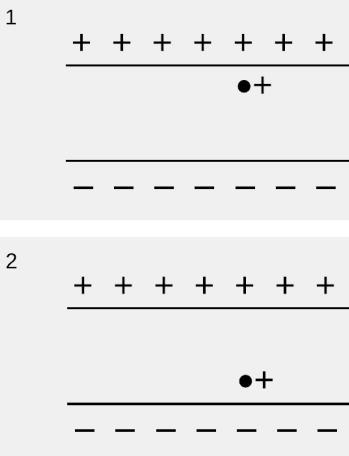
We can also see that such waves must have fields with nonvanishing components perpendicular to the direction in which the wave is traveling, because the cross product $\mathbf{E} \times \mathbf{B}$ is perpendicular to both \mathbf{E} and \mathbf{B} . In fact, for the simplest wave patterns (such as a laser beam or a small enough piece of sunlight), we will see that the fields are purely perpendicular to the direction of propagation — they have no component at all parallel to the momentum. Such a wave is referred to as a transverse wave, as opposed to a longitudinal



b / Halley's comet, example 1.



c / Example 2.



d / Discussion question A.

wave. In the examples in ch. 2, the waves on a string are transverse, while sound waves are longitudinal.

Discussion question

A The figure shows a positive charge in the gap between two capacitor plates. First make a large drawing of the field pattern that would be formed by the capacitor itself, without the extra charge in the middle. Next, show how the field pattern changes when you add the particle at these two positions. Compare the energy of the electric fields in the two cases. Does this agree with what you would have expected based on your knowledge of electrical forces?

3.2 Geometry of a plane wave

3.2.1 E and B perpendicular to the direction of propagation

The momentum density $(1/4\pi k)\mathbf{E} \times \mathbf{B}$ is proportional to the momentum density of our plane wave, and therefore points in the direction of propagation. Since a vector cross product is perpendicular to both of the vectors, it follows that both fields lie in the plane perpendicular to the direction of propagation.

3.2.2 E and B equal in energy

Figure e shows two examples of electric and magnetic fields that we imagine as possible fields in an electromagnetic plane wave. We draw the arrows for the \mathbf{E} and \mathbf{B} vectors with equal lengths on the page, which suggests that they are “equal” in the sense of carrying equal energy, which we are now going to prove. The angle ϕ is drawn as an arbitrary angle, although we will prove later that it must be a right angle. The first example is drawn with \mathbf{E} clockwise from \mathbf{B} , so that by the right-hand rule, the direction of propagation is out of the page. The second example has been flipped around so that the momentum vector is into the page, and has also been rotated in the plane of the page.

Our argument for equal energy sharing between the electric and magnetic fields works by colliding these two waves head-on. Before the waves collide, they each carry energy $U_E + U_B$, for a total energy of $2U_E + 2U_B$. Now suppose that the rotation is chosen as in the figure, so that when the waves superpose, the electric fields cancel. At this moment, the total energy is $U_{B'}$, where \mathbf{B}' is the result of vector addition of the two magnetic field vectors at an angle of $\pi - 2\phi$ relative to each other.

That was one possible choice of the rotation. But we can also choose the rotation such that the *magnetic* fields cancel, so that the total energy is $U_{E'}$, where \mathbf{E}' is the result of a similar vector addition problem involving the same angle.

Requiring conservation of energy in both examples, we have $U_E + U_B = U_{B'} = U_{E'}$, but since the two vector addition problems involve

the same angle, we must have $U_{\mathbf{E}} = U_{\mathbf{B}}$, as claimed.

Since the energy densities are $(1/8\pi k)E^2$ and $(c^2/8\pi k)B^2$, it follows that $E = cB$ (problem 1, p. 82). With a couple of centuries of hindsight, it would have been better if we had constructed a system of units in which E and B had the same units, and in fact they do have the same units in the cgs system. In the SI their units are different, but ignoring the factor of c , we can say that this means the magnitudes of the electric and magnetic fields in a plane wave are “equal.”

3.2.3 \mathbf{E} and \mathbf{B} perpendicular to each other

Continuing the analysis of the colliding waves, we find that the angle ϕ between \mathbf{E} and \mathbf{B} must be a right angle. We have four units of energy before the waves collide: one unit in each wave’s electric field, and one unit in each magnetic field. When the waves collide in an orientation such that the electric fields cancel, then $U_{\mathbf{B}'}$ has a value, in these units, of $4 \sin^2 \phi$, which gives conservation of energy only if ϕ is a right angle. We arrive at the geometry shown in figure f.

Because the angle ϕ is fixed at 90° , it’s not a property like color or brightness that can distinguish one light wave from another. But we are always free to take a diagram like figure f and simply spin the whole book around by an angle θ . This is referred to as the polarization of the wave.

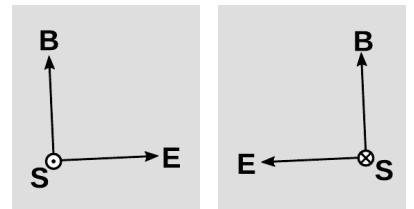
Crossed polarizing films

example 4

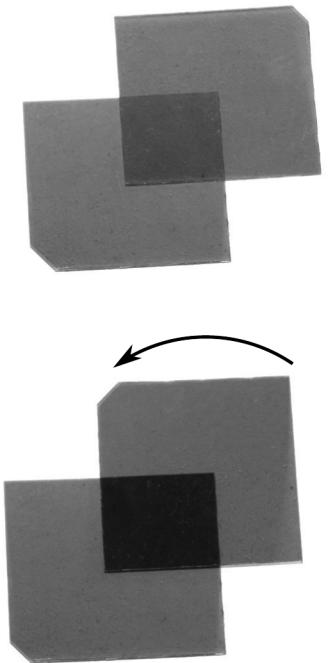
A polarizing filter is one that passes light that has its polarization oriented in a certain direction, while blocking it if the polarization is in the perpendicular direction. The photos show two polarizing filters that overlap, with the light coming from the back being a random mixture of small wave-trains with random polarizations.

At the first filter, light with the “right” orientation gets through, while light with the “wrong” orientation is blocked. Of course, a randomly chosen angle will not be at exactly 0° or 90° . A wave with an intermediate angle of polarization can be broken down into *components* (say the components of \mathbf{E} , although it doesn’t matter in principle whether we talk about \mathbf{E} or \mathbf{B} , since their orientations are fixed relative to each other). On the average, these components are equal in energy, so half the light is transmitted.

The filters overlap, so the light now has to pass through the second filter. In the top photo, the two filters have been oriented the same way, so that in principle any light that passes through the first filter should also get through the second without any reduction in intensity. Because the filters are nonideal, we do observe some further reduction in brightness where the filters overlap, but not very much.



f / The geometry of a plane wave, with $\phi = 90^\circ$.



g / Example 4.

In the bottom photo, one filter has been rotated by 90° . Any component that passes the first filter is in exactly the wrong direction to get through the second, so we see black where the filters overlap.

Discussion question

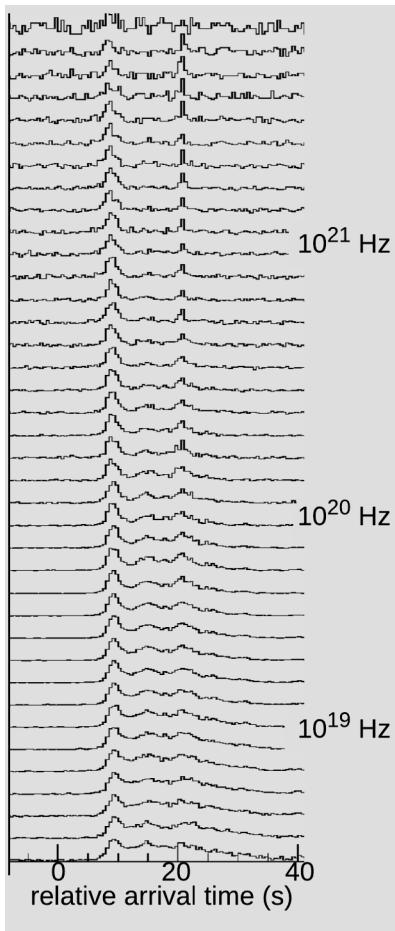
A Suppose someone tells you that a beam of light consists of a stream of electrons moving through space. Use the experiment in example 4 to convince them that they're wrong.

3.3 Propagation at a fixed velocity

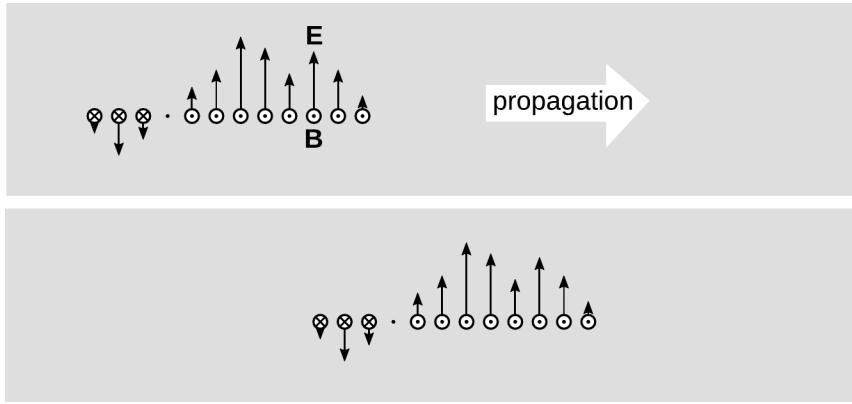
If we look at the speed of waves in general, we see two main types of behavior. In one type, exemplified by sound waves or waves on a string, all waves travel at the same speed, regardless of their amplitude or frequency. Water waves are a good example of another type, called a dispersive wave. A dispersive wave travels at different velocities depending on its frequency. Only a perfect sine wave has a definite frequency, so typically if we generate a wave with some randomly chosen shape, it will act like a mixture of different frequencies. (There is a mathematical theorem called Fourier's theorem that says we can always analyze any wave as a superposition of sine waves.) These different frequencies will travel at different speeds, so the wave acts like a bunch of runners over the course of a long-distance race: at the start they're all crowded together, but as time goes on, the faster ones pull ahead, the slower ones fall behind, and the pack spreads out in space, often to the point where people are running all alone and can't see their competitors. In a wave, this causes the wave pattern to spread out, or disperse. When we see that a pulse on a string propagates without changing its shape, as in figure d, p. 34, we can tell that there is no dispersion. Dispersion is an important phenomenon in general, and we will encounter it again in our studies of optics (sec. 11.3.3, p. 244) and quantum physics.

We know on both theoretical and empirical grounds that electromagnetic waves are nondispersive when they travel in a vacuum. Figure h shows some astronomical evidence that is extremely impressive for its accuracy. Electromagnetic waves (gamma rays) with different frequencies, spanning several orders of magnitude, were generated, probably by matter falling into a black hole. These waves then traveled for 9 billion years before reaching earth, where the different frequencies arrived within seconds of one another.

Theoretically, we have the following argument. In general, the ratio of an object's energy to its momentum depends on its speed (cf. p. 66). But the geometrical facts we've found about electromagnetic waves guarantee that the energy and momentum scale up and down with amplitude in exactly the same way, and are independent



h / Arrival times of waves with different frequencies from gamma-ray burst 160625B, from Wei *et. al.*, 2018. In this histogram, the vertical axis is a count of the number of wave pulses arriving per unit time.



i / A plane wave propagating to the right, shown at one time (top) and a later one (bottom).

of the shape of the wave. For example, if the angle ϕ between the \mathbf{E} and \mathbf{B} vectors could vary, or if $|\mathbf{E}|$ and $|\mathbf{B}|$ could vary independently, then we could get different momenta for the same energy — but these things are *not* independently variable. Since electromagnetic waves have a fixed ratio of energy to momentum, they must travel at a fixed speed. Thus they are nondispersive, so a plane wave glides along rigidly as in figure i, without changing shape.

If all electromagnetic waves travel at a fixed speed, then what speed is that? If this speed is invariant (the same in all frames of reference), then it must be c , which is the only invariant speed (sec. 1.6.3, p. 21). Experiments verify that this is the case, although historically there was considerable confusion on this point (see p. 79 and ch. 4).

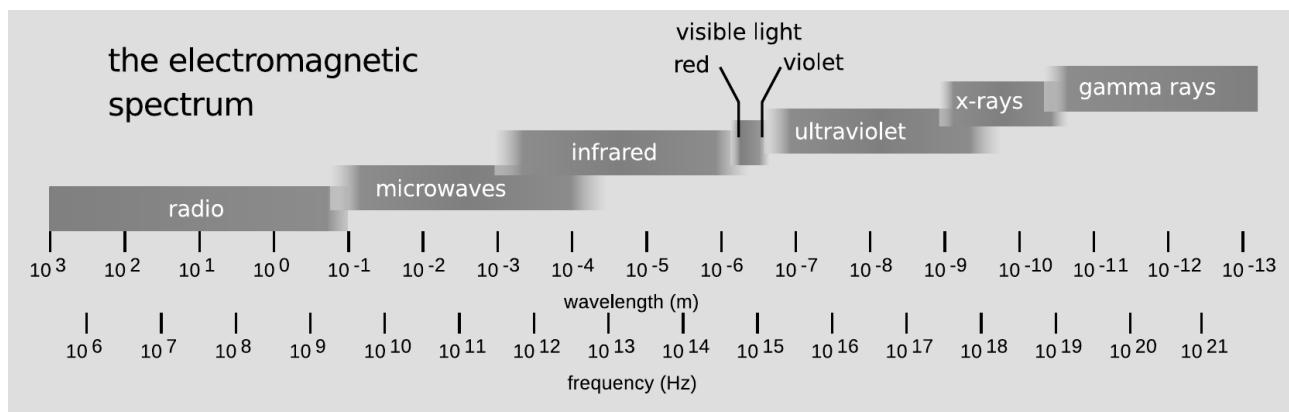
By the way, all of this applies only to a vacuum. For example, there is dispersion when light travels through glass: we observe that blue travels more slowly than red by about 1%. The theoretical argument about energy and momentum doesn't apply here because there are transfers of energy and momentum between the light and the glass while the light is passing through.

3.4 The electromagnetic spectrum

Heinrich Hertz (for whom the unit of frequency is named) verified Maxwell's ideas experimentally. Hertz was the first to succeed in producing, detecting, and studying electromagnetic waves in detail using antennas and electric circuits. To produce the waves, he had to make electric currents oscillate very rapidly in a circuit. In fact, there was really no hope of making the current reverse directions at the frequencies of 10^{15} Hz possessed by visible light. The fastest electrical oscillations he could produce were 10^9 Hz. He succeeded in showing that, just like visible light, the waves he produced were polarizable, and could be reflected and refracted (i.e., bent, as by

a lens), and he built devices such as parabolic mirrors that worked according to the same optical principles as those employing light. Hertz's results were convincing evidence that light and electromagnetic waves were one and the same.

Together, the experimentalist Hertz and the theorist Maxwell showed that electromagnetic waves were in fact the structure underlying a variety of apparently disparate phenomena, including visible light, radio waves, and other phenomena such as x-rays. All of these types of radiation differ only in their frequency, and they lie along a unified electromagnetic spectrum (figure below) in which the visible rainbow spectrum is only a narrow slice.



An electromagnetic wave can be characterized either by its frequency or by its wavelength. The terminology for the various parts of the spectrum is worth memorizing, and is most easily learned by recognizing the logical relationships between the wavelengths and the properties of the waves with which you are already familiar. Radio waves have wavelengths that are comparable to the size of a radio antenna, i.e., meters to tens of meters. Microwaves were named that because they have much shorter wavelengths than radio waves; when food heats unevenly in a microwave oven, the small distances between neighboring hot and cold spots is half of one wavelength of the standing wave the oven creates. The infrared, visible, and ultraviolet obviously have much shorter wavelengths, because otherwise the wave nature of light would have been as obvious to humans as the wave nature of ocean waves. To remember that ultraviolet, x-rays, and gamma rays all lie on the short-wavelength side of visible, recall that all three of these can cause cancer. (As you'll see when you learn about quantum physics, there is a basic physical reason why the cancer-causing disruption of DNA can only be caused by very short-wavelength electromagnetic waves. Contrary to popular belief, microwaves cannot cause cancer, which is why we have microwave ovens and not x-ray ovens!)

3.5 Momentum and rate of energy flow

3.5.1 Momentum of a plane wave

Recalling the relations $d\mathbf{p}/dv = (1/4\pi k)\mathbf{E} \times \mathbf{B}$, $dU_E/dv = (1/8\pi k)E^2$, and $dU_B/dv = (c^2/8\pi k)B^2$, it is straightforward to show that for a plane wave, the energy and momentum are related by

$$dU = c dp.$$

This turns out to be a more general relation that, according to relativity, applies to anything without mass.

3.5.2 Rate of energy flow

Intuitively we feel that sunlight *flows* through a window. We have been focusing on the momentum density as a measure of this rate of flow, but it would be equally valid to quantify it in units of power per unit area (watts/meter²). These two figures must somehow be equivalent, since we can't change one without changing the other by the same factor. The relationship between them is

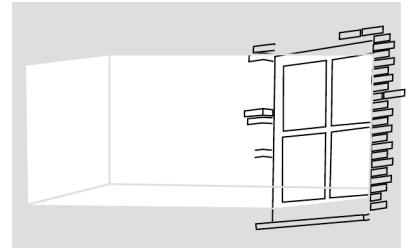
$$\frac{\text{power}}{\text{area}} = \frac{\text{momentum}}{\text{volume}} \times c^2.$$

To see this, consider an imaginary rectangular box of length ℓ , with the window of area A forming one end. Its volume is $v = \ell A$. Let's say the light is flowing directly along the length of this box, striking the window flat-on. At a given instant, the box contains momentum p and energy cp . The time it will take for this entire box worth of light to flow through the window is $t = \ell/c$, so that the power per unit area is $(cp/t)/A = c^2 p / (\ell A) = (p/v)c^2$.

Summarizing, we find that the vector cross product $\mathbf{E} \times \mathbf{B}$ can be interpreted either as a measure of momentum density or as a measure of the rate of flow of energy.

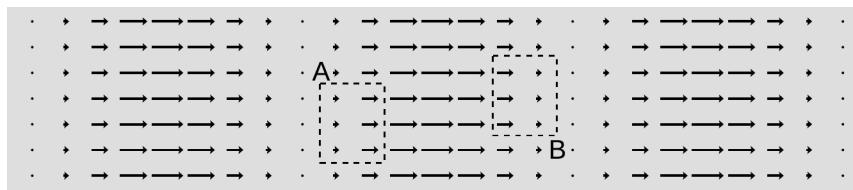
The quantity $(c^2/4\pi k)\mathbf{E} \times \mathbf{B}$, which is just the momentum density multiplied by c^2 , is often notated \mathbf{S} , and is referred to as the Poynting vector, after John Henry Poynting — a wonderful coincidence, because the vector *points* in the direction of the momentum and energy flow. Poynting coined the term “greenhouse effect” in 1909. The magnitude of the Poynting vector is power per unit area.

It makes sense that the momentum density and the rate of energy flow differ by the factor c^2 , which is huge in SI units. SI units were chosen so that their sizes would be of a convenient order of magnitude in everyday life. We know from ordinary experience that the energy flux of a blast of desert sun can be physically staggering, whereas the momentum of the same sunlight is totally undetectable in everyday life.



j / Light fills an imaginary rectangular box, flowing through a window of area A .

k / The Poynting vector of a sinusoidal plane wave, example 5. There is a net flow of energy out of region A, and a net flow into B.



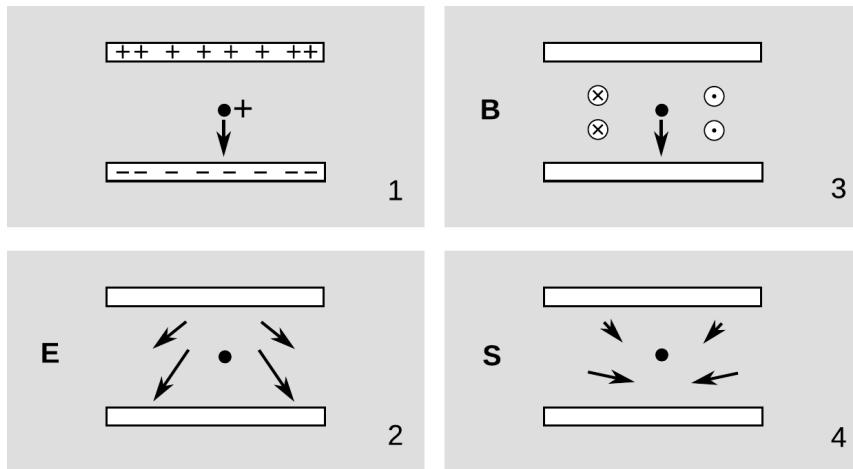
Poynting vector of a plane wave

example 5

Figure k shows the Poynting vector, in the sea-of-arrows representation, for the example of a sinusoidal plane wave. In a coordinate system where $+z$ is to the right, frozen at one point in time, this wave could be described by $\mathbf{E} = A\hat{x}\sin kz$ and $c\mathbf{B} = A\hat{y}\sin kz$, so that $\mathbf{S} = (A^2c/4\pi k)\hat{z}\sin^2 kz$. The figure shows examples of regions that have a net flow of energy in or out.

In example 5, we have regions of space that are gaining energy, and others that are losing it. It's because there are "winners and losers" that these energy flows are physically observable. As a technical aside, it is also possible to have examples in which the Poynting vector is nonzero, but there is no physically observable flow of energy, because every region of space is having energy flow in as fast as it flows out ([280](#)).

I / The energy flow for a point charge released from rest in a capacitor. The \mathbf{E} , \mathbf{B} , and \mathbf{S} vectors are shown at four sample points.



A charge accelerated inside a capacitor

example 6

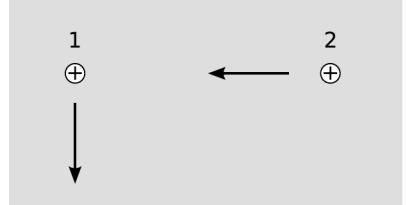
In discussion question A, p. 68, we convinced ourselves that if a charge was released inside a capacitor, the kinetic energy it gained could be properly accounted for by the energy lost from the electric field. This can also be calculated quantitatively [280](#). Figure I shows how this plays out in terms of the Poynting vector. The setup is recapitulated in I/1. The electric field, I/2, is the superposition of the capacitor's nearly uniform downward field and the outward field pattern of the particle. As the particle moves downward, it creates a magnetic field, I/3, similar to that of a current-carrying wire. Taking the vector cross product $\mathbf{E} \times \mathbf{B}$ gives

us the Poynting vector \mathbf{S} (ignoring constants of proportionality). We see that the energy flow is out of the electric field in the top of the capacitor, and into the center, where the particle is.

It is also interesting to consider the case where the capacitor in example 6 is infinite in size, i.e., we simply fill the whole universe with a uniform electric field. In this case, the Poynting vector tells us that the energy flow comes from infinity ([281](#)).

Discussion question

A Positive charges 1 and 2 are moving as shown. What electric and magnetic forces do they exert on each other? (To find the directions of the relevant magnetic fields, you can pretend that the charges are wires, and you will need to use the right-hand rule.) What does this imply for conservation of momentum?



Discussion question A.

3.6 Relativistic consequences

3.6.1 $E=mc^2$

Fields carry inertia

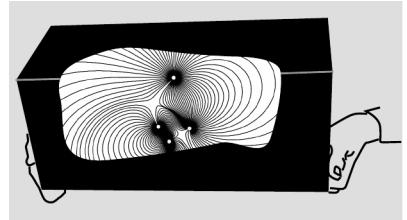
Suppose you're given a black box, figure m. You're not allowed to open it, but you're able shake it around and measure its momentum. By trial and error, you find that there is some frame in which its momentum is zero. (If there are things moving around inside, this may not be the frame in which the externally visible cardboard sides of the box are at rest.) This is what we might as well call the box's rest frame, the frame in which it is at rest, in some over-all sense.

Next you can measure its nonzero momentum p when you shake it around at various velocities. Knowing p at a particular v allows you to infer the mass, $m = p/v$.

Now suppose you do this, not knowing that inside the box is an electromagnetic field, which has *zero* mass. The energies and momenta you measure are those of the *fields* alone. You will find a frame in which the momentum is zero. This could be a frame in which the field is purely electric. If you now set the box in motion, the original electric field pattern turns into a new electric field plus a magnetic field pattern.¹ These electric and magnetic fields have some momentum density, proportional to $\mathbf{E} \times \mathbf{B}$. You measure the total momentum. You infer a certain mass.

Hm. This seems like mass without mass. There are no material particles inside the box, and yet the box acts like it has mass.

Suppose that the field is purely electric in the box's rest frame,



m / The black box has electromagnetic fields inside. If we shake it, it has inertia.

¹We assume that the fields are transported with the cardboard box, so that the result of moving the box at velocity v is the same as if we left the box unaccelerated and simply took our measurements while *we* were moving at v . In reality the results of accelerating the box would depend on the details of how the fields were created and sustained.

and we have a way to make this electric field stronger or weaker. When we do this and then set the box in motion, the energy of the fields and the mass we infer change by equal factors. For example, if we increase the electric field by a factor of 3, then the energy goes up by a factor of 9. But when the box is moving, this also has the effect of multiplying \mathbf{B} by a factor of 3 (because the transformation of the fields is linear), so $\mathbf{E} \times \mathbf{B}$ goes up by a factor of 9. This means that the momentum goes up by 9 times, and so does the mass that we infer at a given velocity.

Equivalence of mass and energy

In this example, energy and mass are *equivalent*. Based on units, the relation must be of the form $E = (\text{constant})mc^2$, where the constant is unitless. Einstein showed that the unitless constant was equal to 1, and was the same for any system, regardless of what type or types of energy are involved.² This is the famous $E = mc^2$, which states that mass and energy are equivalent.

The equation $E = mc^2$ tells us how much energy is equivalent to how much mass: the conversion factor is the square of the speed of light, c . Since c a big number, you get a really really big number when you multiply it by itself to get c^2 . This means that even a small amount of mass is equivalent to a very large amount of energy. Conversely, an ordinary amount of energy corresponds to an extremely small mass (example 7), and this is why nobody discovered mass-energy equivalence experimentally hundreds of years before Einstein.

It's fairly easy to see that if mass is equivalent to one form of energy, then it must be equivalent to all other forms of energy, with the same conversion factor. Let's take heat as an example. Suppose a rocket ship contains some electrical energy stored in a battery. What if we believed that $E = mc^2$ applied to electromagnetic energy but not to heat. Then the pilot of the rocket could use a battery to run a heater, decreasing the mass of the ship. Since momentum $p = mv$ is conserved, this would require that the ship speed up!

This would not only be strange, but it would violate the principle that motion is relative, because the result of the experiment would be different depending on whether the ship was at rest or not. The only logical conclusion is that all forms of energy are equivalent to mass. Running the heater then has no effect on the motion of the ship, because the total energy in the ship was unchanged; one form of energy (electrical) was simply converted to another (heat).

A somewhat different, and equally valid, way of looking at $E = mc^2$ is that energy and mass are not separately conserved. Therefore

²Here the “system” has to be an isolated one. If the system is not isolated, then it can be exchanging energy and momentum with the outside world. The analysis then gets more complicated, and $E = mc^2$ can be false.

we can have processes that convert one to the other.

A rusting nail

example 7

▷ An iron nail is left in a cup of water until it turns entirely to rust. The energy released is about 0.5 MJ. In theory, would a sufficiently precise scale register a change in mass? If so, how much?

▷ The energy will appear as heat, which will be lost to the environment. The total mass-energy of the cup, water, and iron will indeed be lessened by 0.5 MJ. (If it had been perfectly insulated, there would have been no change, since the heat energy would have been trapped in the cup.) The speed of light is $c = 3 \times 10^8$ meters per second, so converting to mass units, we have

$$\begin{aligned} m &= \frac{E}{c^2} \\ &= \frac{0.5 \times 10^6 \text{ J}}{(3 \times 10^8 \text{ m/s})^2} \\ &= 6 \times 10^{-12} \text{ kilograms.} \end{aligned}$$

The change in mass is too small to measure with any practical technique. This is because the square of the speed of light is such a large number.

Electron-positron annihilation

example 8

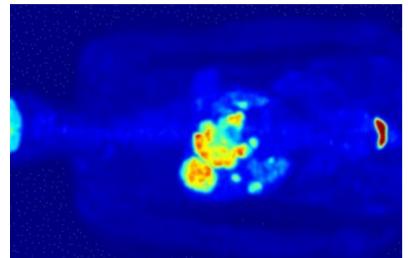
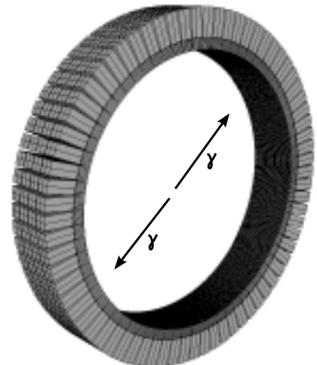
Natural radioactivity in the earth produces positrons, which are like electrons but have the opposite charge. A form of antimatter, positrons annihilate with electrons to produce gamma rays, a form of high-frequency light. Such a process would have been considered impossible before Einstein, because conservation of mass and energy were believed to be separate principles, and this process eliminates 100% of the original mass. The amount of energy produced by annihilating 1 kg of matter with 1 kg of antimatter is

$$\begin{aligned} E &= mc^2 \\ &= (2 \text{ kg}) (3.0 \times 10^8 \text{ m/s})^2 \\ &= 2 \times 10^{17} \text{ J,} \end{aligned}$$

which is on the same order of magnitude as a day's energy consumption for the entire world's population!

Positron annihilation forms the basis for the medical imaging technique called a PET (positron emission tomography) scan, in which a positron-emitting chemical is injected into the patient and mapped by the emission of gamma rays from the parts of the body where it accumulates.

Electron-positron annihilation is analyzed in more detail in example 6, p. 153.



n / Top: A PET scanner. Middle: Each positron annihilates with an electron, producing two gamma-rays that fly off back-to-back. When two gamma rays are observed simultaneously in the ring of detectors, they are assumed to come from the same annihilation event, and the point at which they were emitted must lie on the line connecting the two detectors. Bottom: A scan of a person's torso. The body has concentrated the radioactive tracer around the stomach, indicating an abnormal medical condition.

3.6.2 Einstein's motorcycle

In a vibration, such as the motion of a pendulum or a mass on a spring, we would define the amplitude either as the position of the object relative to equilibrium, or as some other, closely related quantity such as the object's velocity. In these examples, position and velocity are not independent measures of amplitude. They are closely related, and we can't change one without changing the other proportionately. A wave is a kind of vibration that exists across a whole region of space, so the same ideas recur. For example, the amplitude of a sound wave could be defined in multiple ways: in terms of the displacement of the air, its velocity, or the pressure or density. Again, all of these things are related and cannot be controlled independently.

In the case of an electromagnetic wave, we could define the amplitude in terms of either the electric field or the magnetic field. We have already seen that in an electromagnetic wave we can't have one of these be zero while the other is nonzero. Shortly we will see that in a plane wave, they are in fact directly proportional to each other.

An electromagnetic wave, unlike the mechanical waves in ch. 2, isn't a vibration of any material medium such as air or a piece of string. What vibrates is the fields — invisible, intangible, and massless. Electromagnetic waves are transverse. But because they are not vibrations of a material medium, this doesn't mean that anything is actually *traveling* from side to side. In an example like the photos of the coil spring in figure d, p. 34, we could imagine a bug sitting on the spring. The bug moves to the side and then back as the pulse passes through, but nothing analogous happens in an electromagnetic wave. The bug would simply notice a change in the fields over time, but would not move.

Early physicists working on the description of electromagnetic radiation had never had any previous experience with waves that were not mechanical vibrations of a physical medium. James Clerk Maxwell gave a complete and correct mathematical description of electromagnetism in 1865, and his equations worked just fine as a description of radiation purely in terms of non-material fields. But old habits died hard, and as late as the 1930's, it was common to hear physicists referring to electromagnetic waves as vibrations in a mysterious medium called the "aether."

We can laugh at the silly people who believed in the aether, but getting rid of it might seem to cause more problems than it solves. At the Swiss Federal Polytechnic school, a physics student, about 20 years old, was sitting in the back of a classroom, absorbing a lecture on electromagnetism, when he came back to a troubling, half-formed fantasy that he had originally imagined at the age of 16. Suppose, Albert Einstein daydreamed, that I ride on a motorcycle at nearly the speed of light, chasing a light wave as it passes over me. What

would I observe? And what would happen if I rode *at* the speed of light? Physicists at that point in history didn't have a valid answer to these questions unless there was something like the aether. We take up this train of thought again in ch. 4.

Notes for chapter 3

274 Unobservable Poynting vectors

Cases exist where the Poynting vector is nonzero, but there is no flow of energy that is actually observable.

One such example would be the static field of a bar magnet immersed in a uniform ambient magnetic field along the magnet's axis. If you work out the right-hand rule for yourself at various points in space, you should be able to convince yourself that the energy flow goes in circles, like a game of musical chairs. Thus although it seems weird that a static field can "have" momentum and a flow of energy, there are no observable consequences because no region of space can gather up the energy. It's a bit like the situation of a rich teenager who "has" a few million dollars in a trust fund, but can't touch it until she's 21.

274 Electric force on a test charge

The force on a test charge is $\mathbf{F} = q\mathbf{E}$.

When we insert a test charge in an ambient field, the total field at any given point in space becomes the vector sum $\mathbf{E} + \mathbf{E}_q$, where \mathbf{E}_q is the field contributed by the test charge itself. The energy density at this point is proportional to the squared magnitude of this field, $(\mathbf{E} + \mathbf{E}_q) \cdot (\mathbf{E} + \mathbf{E}_q)$. Multiplying this expression out, we get terms $\mathbf{E} \cdot \mathbf{E}$ and $\mathbf{E}_q \cdot \mathbf{E}_q$, which are constants and therefore don't have any effect on our analysis, but in addition we get a term $2\mathbf{E} \cdot \mathbf{E}_q$. It is only this latter term than can change if we move q around, so the force \mathbf{F} on q is proportional to it. Since \mathbf{E}_q is proportional to q (as we can easily prove by Gauss's law), it follows that \mathbf{F} is proportional both to \mathbf{E} and to q . We conclude that $\mathbf{F} \propto q\mathbf{E}$. The remainder of the calculation is only required in order to show that the proportionality constant is 1.

Although the force will only depend on the field at the point where the test charge is, the energy depends on the fields at all points in space. Therefore we are free to take the ambient field to be any field pattern we like. We could use a uniform field filling all of space,

but then the total energy turns out to diverge. (This is discussed further in note 281.) Instead, we take the test charge to be at the origin, and use an ambient field

$$\mathbf{E} = \begin{cases} E\hat{\mathbf{z}} & \text{if } a < z < b \\ 0 & \text{elsewhere,} \end{cases}$$

where $a < 0$, $b > 0$, and E is a constant. This is the field we would get from a parallel-plate capacitor, as in discussion question A on p. 68.

Because of the symmetry of the problem under rotation about the z axis, we use cylindrical coordinates in which R is the distance from the z axis. In these coordinates, the volume of a ring of radius R , radial thickness dR , and height dz is $dv = (\text{circumference}) dR dz = 2\pi R dR dz$. The part of the energy describing the interaction is

$$\begin{aligned} U &= \int_{z=a}^b \int_{R=0}^{\infty} \frac{1}{8\pi k} 2\mathbf{E} \cdot \mathbf{E}_q dv \\ &= \frac{E}{4\pi k} \int_{z=a}^b \int_{R=0}^{\infty} E_{q,z} \cdot 2\pi R dR dz \\ &= \frac{E}{2k} \int_{z=a}^b \int_{R=0}^{\infty} \frac{kq}{r^2} \cos\theta \cdot R dR dz, \end{aligned}$$

where $r = \sqrt{R^2 + z^2}$ is the distance from the origin, and $\cos\theta = z/R$. This becomes

$$\begin{aligned} U &= \frac{Eq}{2} \int_{z=a}^b \int_{R=0}^{\infty} \frac{zR}{r^3} dR dz \\ &= \frac{Eq}{2} \int_{z=a}^b \int_{R=0}^{\infty} \frac{zR}{(R^2 + z^2)^{3/2}} dR dz. \end{aligned}$$

This is the kind of situation where the best strategy is usually to clean up the integrand by expressing it in terms of a unitless variable. For the inside integral, with respect to R , let $u = R/z$, giving

$$U = \frac{Eq}{2} \int_{z=a}^b \int_{u=0}^{\pm\infty} \frac{u}{(u^2 + 1)^{3/2}} du dz,$$

where the sign in the upper limit of the u integral is $+$ for $z > 0$ and $-$ for $z < 0$. The indefinite integral is $-(u^2 + 1)^{-1/2}$, and plugging this in at the limits of integration gives

$$\begin{aligned} U &= \frac{Eq}{2} \int_{z=a}^b \pm 1 dz \\ &= \frac{Eq}{2} (|b| - |a|). \end{aligned}$$

If we let h be the height of the charge above the lower boundary and fix $H = |a| + |b|$, then $|a| = h$ and $|b| = H - h$, so $|b| - |a| = H - 2h$ and $U = -Eqh + \text{const}$. Varying h is equivalent to moving the charge up or down, so the force is $F = -dU/dh = Eq$, which is what we wanted to prove.

275 A point charge in an infinite, uniform electric field

In this example, energy flows from infinity.

If a charged particle is released inside a capacitor, energy flows out of the electric field and into the particle as kinetic energy. We have discussed this energy transformation in discussion question A, p. 68, in example 6, p. 74, and in note 280. In the quantitative analysis of note 280, I avoided the simplest electric field pattern, which would have been a uniform field stretching out to infinity, with the excuse that the total energy would then have been infinite. By doing that, I also conveniently sidestepped the following apparent paradox.

Suppose that the electric field *is* uniform out to an infinite distance. After some time, the particle will have gained some kinetic energy. We can then allow it to hit something and stop, at which point its energy will be converted into heat. (Something similar happens when the beam of an old-fashioned CRT monitor hits the phosphor-coated glass in the front, with part of the energy also being converted into visible light.) But where has this energy come from? If the electric field is truly filling the entire universe uniformly, then it seems that the total energy in the universe's electric field can't possibly have changed. For a field of this kind, superimposing the field of a point charge at one point or another produces exactly the same total electric field pattern, just shifted rigidly through space.

We can make the excuse that the original energy was ∞ , and so is the final energy, and $\infty - \infty$ doesn't have to be zero — it's an indeterminate form. But this isn't as satisfying as an analysis of the actual energy flows.

Such an analysis can be provided simply by letting the capacitor in figure 1, p. 74, approach infinite size. The flows of energy are still qualitatively like the ones shown in figure 1/4, but the sources of this flow are now "off stage" at infinity. This seems like a perfectly natural resolution of the paradox, which we created in the first place by moving the plates of the capacitor off stage to infinity.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 The electric and magnetic fields have different units, so we can't really say whether an electric field is equal in strength to a magnetic field. However, we could consider them to be "morally equal" if their energy densities are equal. Find the value of E/B in this case.

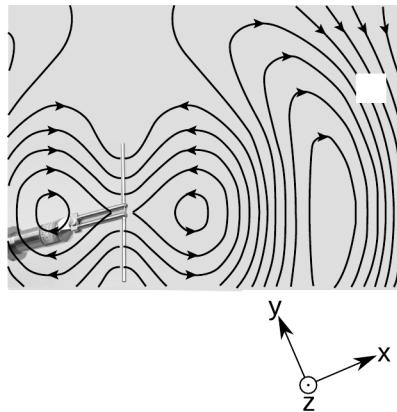
Remark: With hindsight, it was a mistake to design the SI so that E and B had different units. There are in fact other systems of units, such as the cgs system, in which their units are the same, and the expressions for their energy densities have the same constant factor in front. ✓

2 The nuclei ${}^3\text{H}$ (hydrogen-3, or tritium) and ${}^3\text{He}$ (helium-3) have almost exactly the same size and shape, but helium-3's electric field is twice as strong.

Compare the energies stored in the two fields. ✓

3 An electromagnetic plane wave has its electric field in the $+y$ direction and its magnetic field in the $-z$ direction. Find the direction in which it is propagating.

4 The figure depicts the electric fields in the radiation pattern of a certain type of radio antenna, shown in the photo superimposed in the background. Consider the small region of space indicated by the white square. As the waves pass through this area, spreading out like ripples from the antenna, they are moving up and to the right. We therefore expect that their momentum density should be up and to the right. For the reasons discussed on p. 65, the radio wave cannot be purely electric; it must contain a magnetic field as well. For convenience in discussion, a coordinate system is given below the diagram, with the x axis pointing in the direction of propagation. Consider the following six possibilities for the direction of the magnetic field in the area of the white square: $+x$, $-x$, $+y$, $-y$, $+z$, and $-z$. Of these, which are not possible because they don't produce a momentum density in the $+x$ direction?



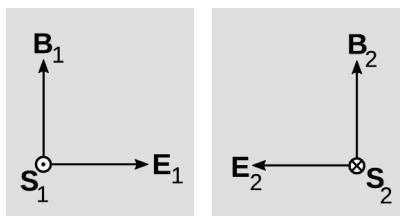
Problem 4.

5 At a particular point in time and space, an electromagnetic plane wave has an energy flux $\mathbf{S} = a\hat{\mathbf{z}}$ ($a > 0$) and an electric field in the $+x$ direction. Find the magnitude and direction of its magnetic field. ✓

6 The figure shows two electromagnetic plane waves, with their associated Poynting vectors. The two waves are equal in intensity. Suppose that these two waves are now superimposed at the same point in space. (a) Find the total Poynting vector by adding the two Poynting vectors.

- (b) Find the total Poynting vector by adding the fields, then computing the Poynting vector from the total.
 (c) Show that the results from the two methods are consistent with each other, and give a physical interpretation.

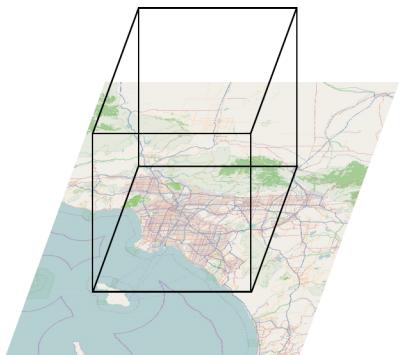
▷ Solution, p. 444



Problem 6.

7 The solar constant is defined as the average flux of electromagnetic energy coming from the sun to the earth, per square meter. It equals 1.36 kW/m^2 . Now imagine, as suggested by the figure, a giant cubical volume of space, 100 km on a side. Find the total momentum of the electromagnetic energy inside this cube. You should find that it is on the same order of magnitude as the momentum of a baseball thrown casually.

✓



Problem 7.

8 (a) A free neutron (as opposed to a neutron bound into an atomic nucleus) is unstable, and undergoes beta decay (which you may want to review). The masses of the particles involved are as follows:

neutron	$1.67495 \times 10^{-27} \text{ kg}$
proton	$1.67265 \times 10^{-27} \text{ kg}$
electron	$0.00091 \times 10^{-27} \text{ kg}$
antineutrino	$< 10^{-35} \text{ kg}$

Find the energy released in the decay of a free neutron. ✓

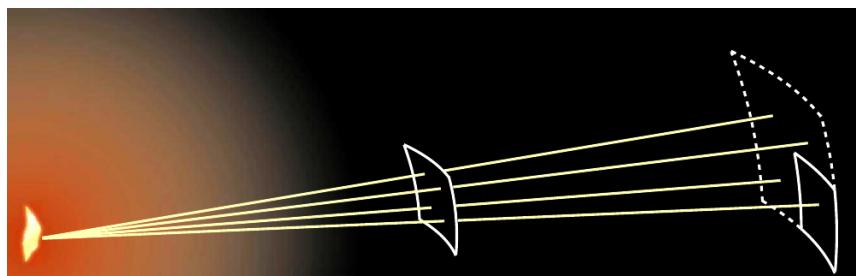
(b) Neutrons and protons make up essentially all of the mass of the ordinary matter around us. We observe that the universe around us has no free neutrons, but lots of free protons (the nuclei of hydrogen, which is the element that 90% of the universe is made of). We find neutrons only inside nuclei along with other neutrons and protons, not on their own.

If there are processes that can convert neutrons into protons, we might imagine that there could also be proton-to-neutron conversions, and indeed such a process does occur sometimes in nuclei that contain both neutrons and protons: a proton can decay into a neutron, a positron, and a neutrino. A positron is a particle with the same properties as an electron, except that its electrical charge is positive. A neutrino, like an antineutrino, has negligible mass.

Although such a process can occur within a nucleus, explain why it cannot happen to a free proton. (If it could, hydrogen would be radioactive, and you wouldn't exist!)

9 The figure shows a candle flame. The light from the flame spreads out in all directions. We pick four representative rays from among those that happen to pass through the nearer square. Of these four, only one passes through the square of equal area at twice the distance. If the two equal-area squares were people's eyes, then only one fourth of the light would go into the more distant person's eye. In other words, the energy flux from a point source goes like $1/r^2$. If the energy flux is an electromagnetic wave, determine the dependence of the electric and magnetic fields on r .

▷ Solution, p. 444



Problem 9.

10 Our intuition is that if we combine two beams of light into a single beam, traveling in the same direction, the intensities should add. But it is not so obvious how this can be, since the fields should add linearly, and the Poynting vector is proportional to the *square* of the fields. Suppose we superimpose two plane waves that are traveling in the same direction, but with random polarizations. Show that, on the average, the intensities do add. ★

Chapter 4

The Lorentz transformation

4.1 Relativity of simultaneity

When we left our hero Albert Einstein on p. 79, he was struggling with a stubborn paradox. What would happen, he wondered, if I rode on a motorcycle at nearly the speed of light, chasing a light wave as it passed over me. What would I observe? And what would happen if I rode *at* the speed of light?

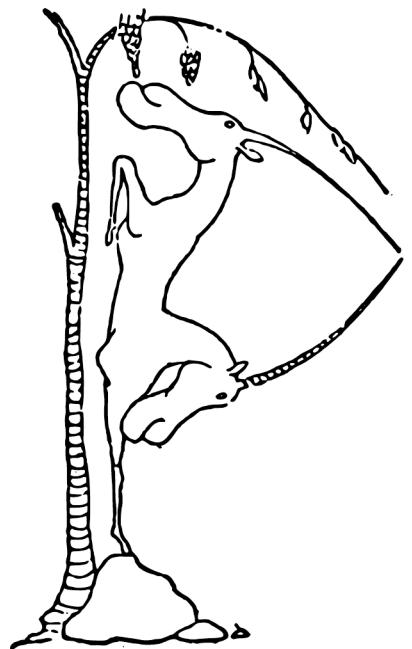
Maxwell's equations are the fundamental laws of physics governing electricity and magnetism, and they do predict the existence of electromagnetic waves that act as we described in ch. 3. Maxwell's equations contain induction terms, which tell us that a changing magnetic field induces an electric field, and conversely a changing electric field creates a magnetic one. Each field helps the other along, sort of like Doctor Dolittle's pushmi-pullyu.

But all of this breaks down in the frame of reference in which Einstein surfs the wave. In that frame, the wave is not moving, so the fields have no time variation. Therefore no induction can take place. In this frame, the wave is *not* a solution of Maxwell's equations.

His university teachers were probably still clinging to aether theories, so if he had asked them about this they might have given him the following answer. Young man, stop being so silly, because motorcycles don't go that fast. And in any case, Maxwell's equations don't work in just any old frame you choose. They're only correct in the frame of reference of the aether.

This might have been an acceptable answer in 1899, but we now know that it won't work. As we've seen, c is not just a speed at which a certain type of wave happens to travel. It's woven directly into the fabric of space and time, cause and effect. If c is some number, then it should be the same number in all frames of reference.

And yet this seems impossible, since we expect velocities to add and subtract in relative motion. If a dog is running away from me at 5 m/s relative to the sidewalk, and I run after it at 3 m/s, the dog's velocity in my frame of reference is 2 m/s. According to everything



The pushmi-pullyu.

we have learned about motion, the dog must have different speeds in the two frames: 5 m/s in the sidewalk's frame and 2 m/s in mine.

Einstein later recalled struggling with this on and off for years. The breakthrough came one day when he was visiting with his friend Michele Besso.

One beautiful day, I visited him and said, I have a problem that I have been totally unable to solve. We then had extensive discussions, and suddenly I realized the solution. The very next day, I visited him again and said, "Thanks to you, I have completely solved my problem."

My solution concerned the concept of time. Namely, time cannot be absolutely defined by itself, and there is an unbreakable connection between time and the velocity of a signal.

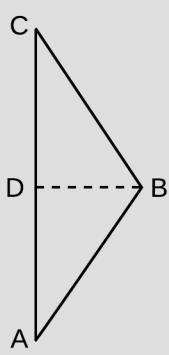
Using this idea, I could now resolve the great difficulty that I previously felt. After I had this inspiration, it took only five weeks to complete what is now known as the special theory of relativity.¹

Let's work this out with the benefit of hindsight, taking advantage of what we know from ch. 1 about the behavior of time. If we want to know how fast something is going, we need to know how far it went and how much time it took. In Einstein's era, clocks in different towns were just starting to be systematically synchronized, so if you took a train from Zurich to Bern, you might have trouble figuring out the train's speed, because the clock in the railroad station in Zurich might be off by fifteen minutes or an hour compared to the one in Bern. So in order to get started, we need some method of figuring out whether two events are *simultaneous*.

Figure a shows one method for doing this. AC is the world-line of an observer, Amy, and we draw it straight up and down because we're imagining this in her frame of reference. She sends a cannonball at some velocity from event A to event B, which is later in time and some distance away from her. It doesn't really matter what this velocity is. It can be anything at all. She hires a helper to be present at B and swat the cannonball back with the same velocity, and she gets it back at event C. If there is any doubt about whether the return velocity was the same, there are methods by which she can easily check. For example, she can attach a clock to the cannonball and ask her helper to write down the clock reading at B. If the velocities are equal, then the interval from A to B should be the same as the one from B to C. Meanwhile, she has been patiently

¹Paraphrased from a translation by John Stachel of a lecture by Einstein in Kyoto, 1922. The lecture was probably given in German and transcribed into Japanese.

a / Determining that B and D are simultaneous.



watching her own clock, which stays with her. She measures the time from A to C on her own inertially moving clock and divides by two, which tells her the time at which event D occurred. Amy has now established that B and D are simultaneous.

Of course this method for synchronization is not very practical. It's just meant to be conceptually simple. A practical example from everyday life is the synchronization of the clock on a phone, figure b, with the atomic clock aboard a GPS satellite. But the point is that we do have ways of doing synchronization. There are in fact many ways of doing clock synchronization. Another way that works is simply to synchronize our clocks side by side, and then make sure that when we move one, we only move it very slowly, rather than at the speed of a passenger jet or a rocket ship. Since relativistic effects on time are proportional to $(v/c)^2$, we can in principle always make these effects as small as desired by making v sufficiently small.

When we carry out this kind of synchronization experiment, we find out two things:

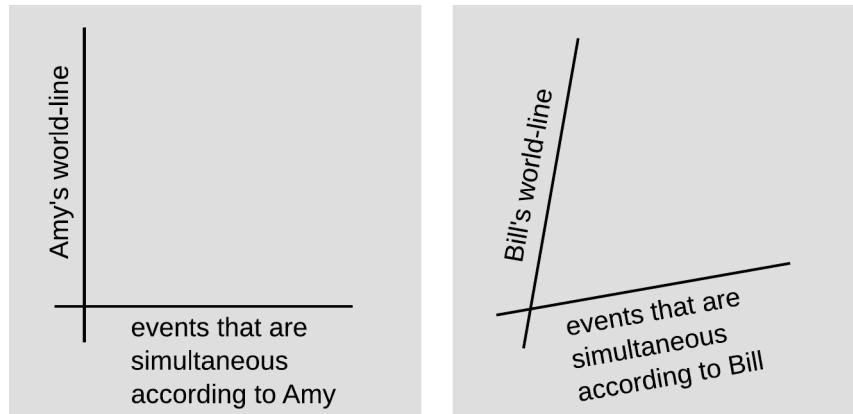
1. All of the methods give consistent results.
2. The results depend on the state of motion of the observer.

Bad news first. Observation 2 tells us that simultaneity is observer-dependent. That is, we can only say that events D and B are simultaneous *in Amy's frame of reference*. That's not so bad. We probably would have suspected this based on the parable of Alice and Betty (p. 24), which showed that we couldn't send signals instantaneously.

The good news is that according to observation 1, we are learning something about spacetime itself, and there is at least *some* notion of simultaneity that makes sense in relativity.



b / This Global Positioning System (GPS) system, running on a smartphone attached to a bike's handlebar, depends on Einstein's theory of relativity. Time flows at a different rates aboard a GPS satellite than it does on the bike, and the GPS software has to take this into account.



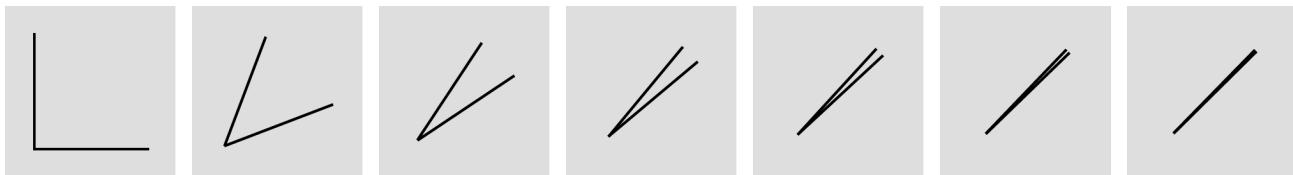
c / Observers Amy and Bill, in different states of motion, disagree on what is simultaneous.

What we end up with is the situation shown in figure c. Amy considers herself to be at rest, so she draws her own world-line

straight up and down. Events like B and D, which are simultaneous, lie along a horizontal line. If Amy was graphing spacetime on graph paper, these would be the axes of the graph, the way she would draw them.

Bill is in motion relative to Amy, so his world-line is slanted a little compared to hers. We've already established that his line of simultaneity can't be the same as hers, so we slant it as well ([Z104](#)).

We can now see the resolution of the puzzle that gave Einstein such a hard time. Suppose that Amy sits at home while Bill blasts off in a rocket ship. He accelerates, then cruises for a while and moves inertially. While he takes a bathroom break, the ship's automatic systems take measurements and establish a new inertial frame of reference. Now he starts up the engines again, accelerates to an even higher speed, and then eats some oreos while the hard-working ship again consults its atomic clocks and surveys spacetime.



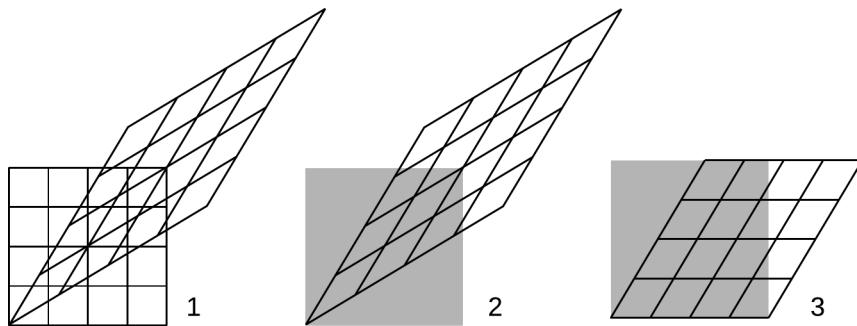
d / Bill keeps up a process of acceleration.

As far as Bill and his ship's computer are concerned, every set of axes they construct is at a nice 90-degree angle. But figure d shows how these appear to Amy. The x and t axes close in on each other like a pair of scissors. The line they're converging to is a line with a slope of c , which equals 1 because I've chosen to draw these diagrams in natural units. We find that: —

3. Velocities don't just add in relativity. Addition is a nonrelativistic approximation.
4. No continuous process of acceleration can bring a material object from $< c$ to $\geq c$.

4.2 The Lorentz transformation

We've seen how the x and t axes of spacetime change when we change frames of reference. Often it's convenient to know how the whole graph-paper grid changes. We can tell already that if we start with squares they'll become parallelograms, and we know that the diagonals with slopes of ± 1 , representing c , have to stay the same. It's not hard to prove that the areas have to stay the same ([Z104](#)). This is enough information to completely determine the result, which is shown in figure e/1. This is called the Lorentz transformation.



e / 1. The Lorentz transformation. 2. The same transformation with a simplifying visual convention applied. 3. The Galilean transformation.

This is a little messy visually because of all the intersecting lines, which create a headachy Moiré-pattern effect. For that reason, this book consistently employs the convention shown in e/2, where the original square grid is replaced by a gray rectangle.

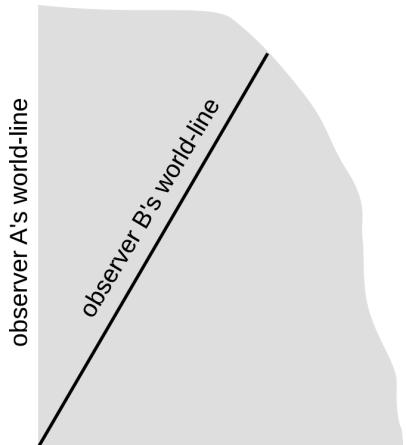
For comparison, e/3 shows the transformation as everyone before Einstein imagined it worked. This is called the Galilean transformation. It doesn't change the horizontal lines, so that all observers agree on simultaneity.

In both the Lorentz transformation and the Galilean transformation, there is the same way of interpreting a slope on the spacetime diagram as a velocity. Figure f shows an example. If you turn the book sideways, then what you're looking at can be interpreted in freshman physics terms simply as the corner of a piece of gray graph paper, with a graph of B's motion, the function $x(t)$. The slope of this graph is B's velocity relative to A, which happens to be about -0.58 in this example. B says that A's velocity is $+0.58$.

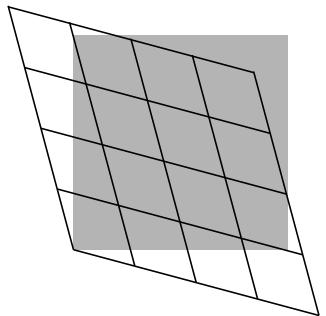
Motion in the opposite direction

example 1

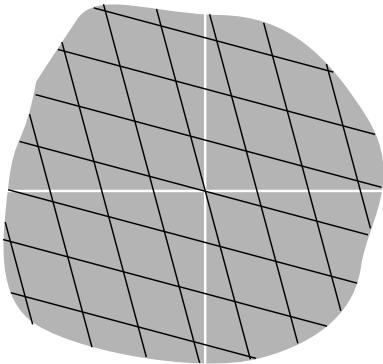
Figure g shows the case where the observer whose frame is represented by the grid is moving to the left relative to the one whose frame is represented by the gray square.



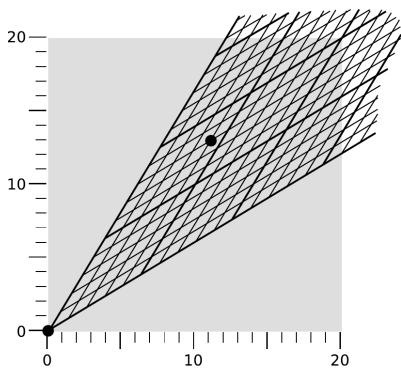
f / Turn the book sideways, and the velocity of one observer relative to the other is the slope shown in the figure.



g / Example 1.



h / Example 2.



i / Example 3.

Other quadrants

example 2

So far I've been arbitrarily choosing to draw only the first quadrant of each coordinate system. Figure h shows a region that includes all four quadrants.

For logical consistency, we need the spacetime interval \mathcal{I} to come out the same regardless of whether we use one set of coordinates or the other.² There is a nice analogy here with Euclidean geometry. In Euclidean geometry, we have a measure of distance between points that is given by $d^2 = x^2 + y^2$, and this distance stays the same even if we rotate our x and y axes. Similarly in the x - t plane of spacetime, we have $\mathcal{I}^2 = t^2 - d^2$, and this stays the same even if we change our frame of reference. Example 3 shows that this works out correctly with some sample numbers.

Invariance of the spacetime interval

example 3

Figure i shows two frames of reference in motion relative to one another at $v = 3/5$. Events are marked at coordinates that in the frame represented by the square are

$$(t, x) = (0, 0) \quad \text{and} \\ (t, x) = (13, 11).$$

The squared interval between these events is $\mathcal{I}^2 = 13^2 - 11^2 = 48$. In the frame represented by the parallelogram, the same two events lie at coordinates

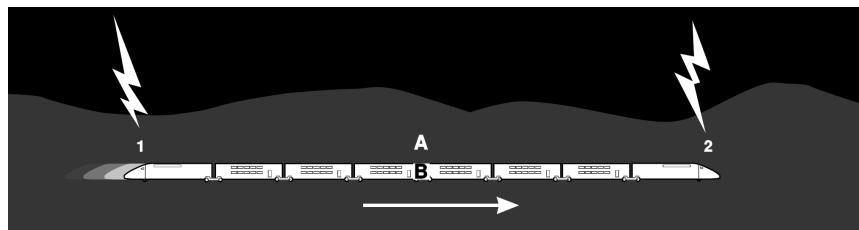
$$(t', x') = (0, 0) \quad \text{and} \\ (t', x') = (8, 4).$$

Calculating the interval using these values, the result is $\mathcal{I}^2 = 8^2 - 4^2 = 48$, which comes out the same as in the other frame.

²Some books use this as the defining property of a Lorentz transformation.

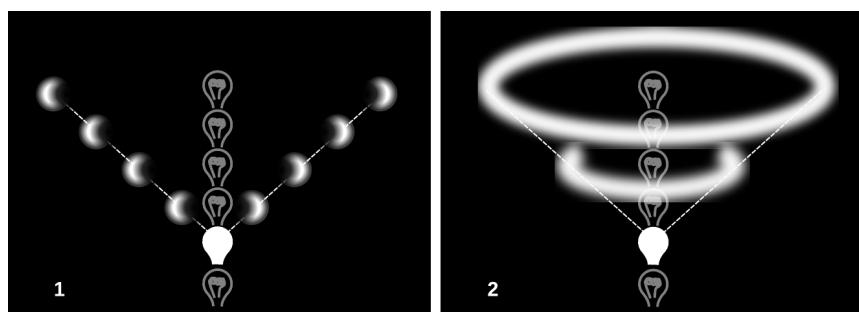
Discussion question

A The figure shows a famous thought experiment devised by Einstein. A train is moving at constant velocity to the right when bolts of lightning strike the ground near its front and back. Alice, standing on the dirt at the midpoint of the flashes, observes that the light from the two flashes arrives simultaneously, so she says the two strikes must have occurred simultaneously. Bob, meanwhile, is sitting aboard the train, at its middle. He passes by Alice at the moment when Alice later figures out that the flashes happened. Later, he receives flash 2, and then flash 1. He infers that since both flashes traveled half the length of the train, flash 2 must have occurred first. How can this be reconciled with Alice's belief that the flashes were simultaneous? Explain using a spacetime diagram.



4.3 The light cone

In figure k/1, a light bulb blinks on momentarily and then back off. Flashes of light spread out from it at c to the right and left, tracing lines which, because of our choice of units, make 45-degree angles. These lines form a geometrical figure, which is called the *light cone*, for reasons that may be more clear from figure k/2, which attempts to represent two dimensions of space as well as the time dimension.

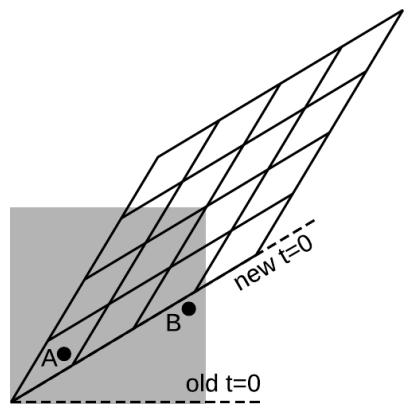


k / 1. A spacetime diagram in 1+1 dimensions, showing the light cone formed by a flash of light emitted by the bulb. 2. The light cone in 2+1 dimensions.

We can draw a light cone centered on any event, such as the one labeled \star in figure l. If \star is the Russian Revolution, then events inside \star 's future light cone can be described as all the ones that we could reach, starting from the Russian Revolution, by traveling at less than the speed of light. A signal could be sent from \star to any of these events, and \star could be the *cause* of any of these events.

Events inside \star 's light cone are said to have a *timelike* relation to \star , because they are separated from it by more time than space. For example, one of these events could be separated in time from \star by 100 years, but in space by only 50 light-years. Then a spaceship could get there by going at half the speed of light. Events are timelike in relation to each other if they have $\mathcal{I}^2 > 0$ for the interval between them. Timelike relationships can be future-timelike or past-timelike. Events outside \star 's light cone, with $\mathcal{I}^2 < 0$, are said to be *spacelike* in relation to \star , and events on the surface of the light cone, $\mathcal{I}^2 = 0$, are *lightlike*.

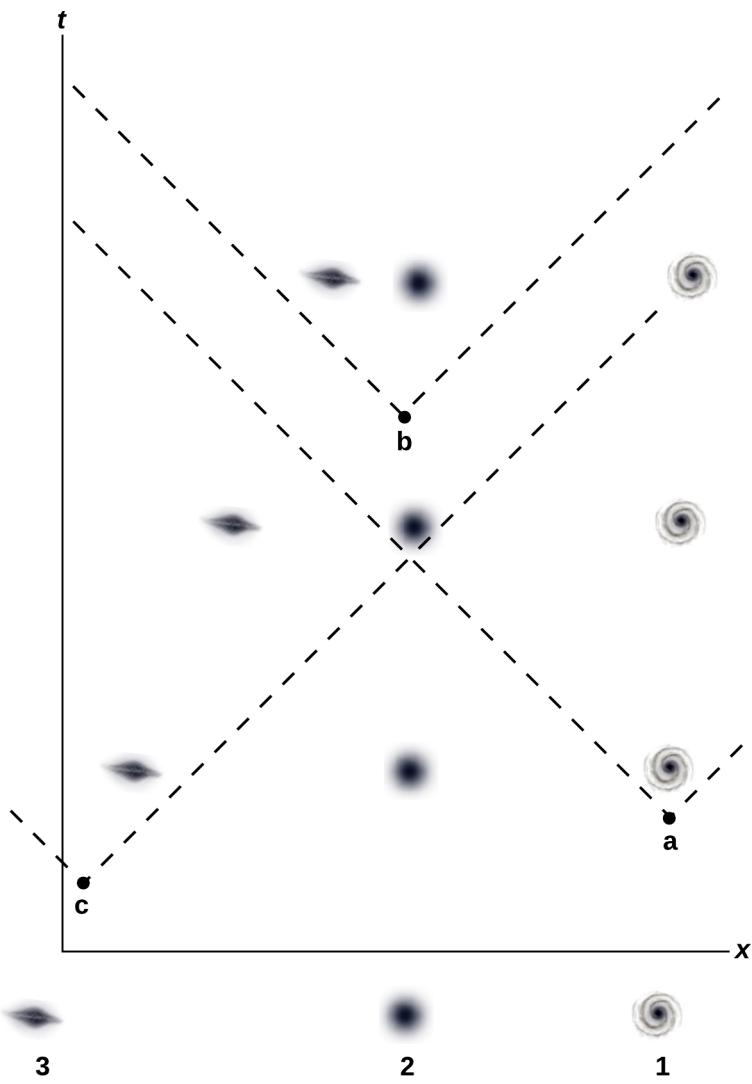
l / The light cone divides up spacetime into three categories in terms of possible cause-and-effect relationships with the event \star .



m / In the original frame, B's time is greater than A's. In the new frame, it's the other way around.

In Einstein's theory of space and time, the light cone plays the same fundamental geometrical role as the circle in Euclidean geometry. A circle looks the same regardless of how we rotate our frame of reference. The light cone looks the same regardless of our state of motion. We can see that all observers will agree on the light cone, because it can be defined in terms of the interval, which is invariant.

The light cone is fundamentally important because it tells us about the cause-and-effect structure of spacetime. Bad things happen if we have cause-and-effect relationships between events that are *spacelike* in relation to each other, like events A and B in figure m. We can then do a Lorentz boost that causes the time-order of the events to be reversed. Say A is Captain Kirk sending a subspace radio distress call, and B is Starfleet receiving it and initiating a rescue mission. In the second frame of reference, represented by the parallelogram grid, B comes before A, and we have the paradoxical situation in which the rescue mission causes the distress call. (This is a stronger conclusion than our earlier one about c as a speed limit on cause and effect, [≥104](#).)



n / Discussion question A.

A Figure n is spacetime diagram showing three galaxies. As astronomers often do, I've depicted them as photographic negatives, so that the black background of outer space looks white. The axes are drawn according to an observer at rest relative to the galaxy 2, so that that galaxy is always at the same x coordinate. Intelligent species in the three different galaxies develop radio technology independently, and at some point each begins to actively send out signals in an attempt to communicate with other civilizations. Events a, b, and c mark the points at which these signals begin spreading out across the universe at the speed of light. Find the events at which the inhabitants of galaxy 2 detect the signals from galaxies 1 and 3. According to 2, who developed radio first, 1 or 3? On top of the diagram, draw a new pair of position and time axes, for the frame in which galaxy 3 is at rest. According to 3, in what order did events a, b, and c happen?

4.4 The diagonal stretch factor and two of its applications

4.4.1 Definition of the stretch factor

When we take a square and subject it to a Lorentz transformation, turning it into a parallelogram, we stretch one axis and shrink the other. For example, if one axis is doubled in length, then the other one must be cut in half, since the area has to stay the same. We call this diagonal stretch factor D . It's a very handy tool, and we'll also see in a moment that it has a direct physical interpretation as a Doppler shift factor. Algebra shows that it is related to the relative velocity v of the two frames of reference by

$$D = \sqrt{\frac{1+v}{1-v}}.$$

For $v = 3/5$, we happen to get $D = 2$, and this is why I'll often make up examples with this velocity, so that numbers come out simple.

self-check A

Is the relation $D = \sqrt{(1+v)/(1-v)}$ expressed in natural units, or in SI units?

▷ Answer, p. 455

4.4.2 Combination of velocities

An application of D is to figure out how to combine velocities. In our example on p. 88, Bill did a series of accelerations, and we saw visually how the effect of these successive Lorentz transformations was to bring his x and t axes closer and closer together, like the blades of a pair of scissors that never quite meet. When we boost an object's speed like this, we sometimes refer to the transformation as a Lorentz boost, or simply a boost. Those drawings were constructed using successive boosts equal to 40% of the speed of light. If velocities added as in Galilean relativity, then Bill's velocity would have gone 0.4, 0.8, 1.2, . . . Let's see what actually happens. The stretch factor for $v = 0.4$ is $D = 1.53$. If we do two of these boosts in a row, then the long diagonal of one of our squares will get stretched by a factor of $1.53^2 = 2.33$. If we now solve for v in terms of D (problem 2, p. 107), and plug in $D = 2.33$, we get $v = 0.69$. A third boost gives $v = 0.85$. Rather than surpassing the speed of light, we're just approaching it.

Some notes on streamlining this type of calculation are relegated to a note ([≥104](#)).

We can also check that all of this is consistent with the fact that c is invariant (sec. 1.6.3, p. 21), i.e., the same in all frames of reference. We have $\lim_{v \rightarrow +1} D = \infty$, which we can describe in words by saying that the stretch factor for $v = +1$ is $D = \infty$. If we switch frames of reference, then this infinite D will get multiplied by some other, finite D , but it will still be infinite. Thus $v = +1$ combined with any sub-light velocity will still be $v = +1$, i.e., c .

self-check B

Carry through the same kind of reasoning for $v = -1$. ▷ Answer, p.

455

An experimental test

example 4

Relativity predicts that when a source emits light, the speed of the light should be independent of the velocity of the source relative to the observer who observes the light's speed. A sensitive test was performed by Alvager *et al.* A particle accelerator produced a beam of π^0 particles moving at $v = 0.99975$ ($D = 90$). These particles are unstable and rapidly undergo radioactive decay into a pair of gamma rays. Gamma-rays are a form of high-frequency electromagnetic radiation, so relativity predicts that they move at c , whereas according to Galilean relativity one might expect forward-emitted gammas to move at $v + c$, i.e., almost twice the speed of light. The experimentalists measured the time of flight of the gammas along a vacuum pipe having a length of 31.450 m. The resulting speed was 0.99993 ± 0.00013 , which is nowhere near the Galilean prediction of 2, and is consistent with special relativity to within the experimental error.

4.4.3 The relativistic Doppler shift

Figure o shows an elaboration on the story of Alice and Betty, p. 24, in which the twins send radio signals back and forth in an effort to decide who is really slow and who is really fast. The situation is totally symmetric, and the outcome is the same regardless of whether Alice receives Betty's signals, o/1, or the other way around, o/2. The only real difference from the earlier figure is that now I have the twins sending a steady series of beeps. It may not look like it visually, but the time between transmission of the beeps is the same before and after Alice and Betty pass by each other. (It's a bit of an optical illusion. Try measuring with a ruler.) We can see that the beeps are *received* close together during the approach, but far apart as the twins recede from each other. This is in fact a Doppler shift. If we like, we can say that instead of a series of beeps, these are just the crests of a radio wave.

In fact, it's not hard to show (Z105) that the Doppler shift in this situation is exactly the factor $D = \sqrt{(1+v)/(1-v)}$,

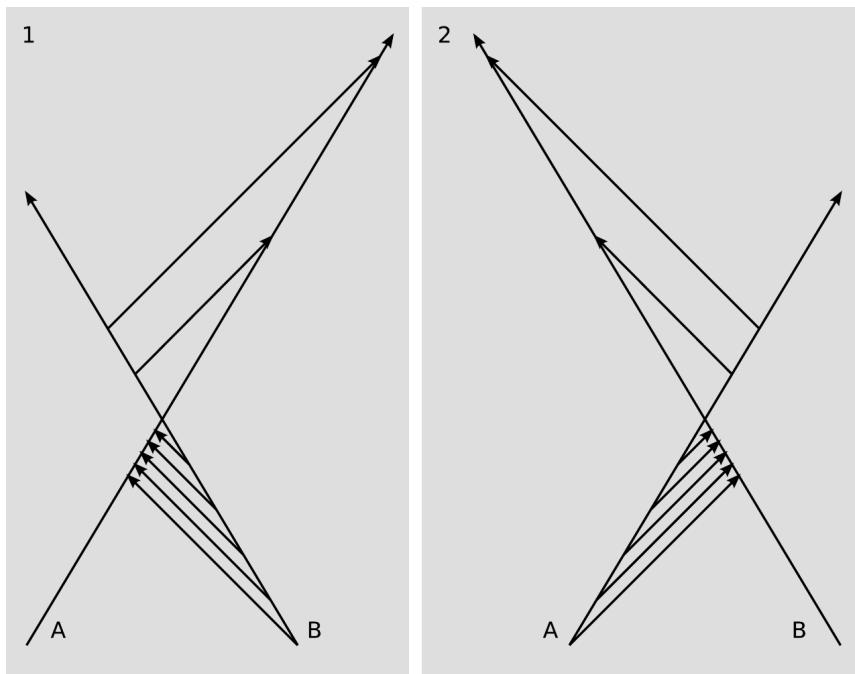
$$f' = Df.$$

This can be contrasted with the nonrelativistic Doppler frequency shift for a moving source,

$$f' = \frac{f}{1-v},$$

which we obtain from the expression on p. 47 by setting the wave speed equal to 1. The two expressions turn out to be approximately

o / Doppler shifts in signals transmitted from Alice to Betty or Betty to Alice.



equal for $v \ll c$, as demanded by the correspondence principle. The relativistic Doppler shift is actually conceptually simpler than the nonrelativistic one. Since there is no medium for the waves, the result can only depend on the velocity of the received relative to the source, whereas the Doppler shift for something like a sound wave can in general depend on the relative velocities of all three things: the source, the received, and the medium.

By the way, throughout this treatment of relativity, I've been avoiding complications by discussing only a single spatial dimension x , so that spacetime is the space of points labeled by coordinates (t, x) . Relativists call this 1+1 dimensions: 1 dimension of space and 1 of time. We actually live in 3+1 dimensions. For the most part nothing especially interesting or complicated happens when we add the y and z dimensions, but in the case of the Doppler shift it should be noted that this expression applies only when the motion is parallel to the line between the source and receiver. In relativity, unlike nonrelativistic physics, we can also have a *transverse* Doppler shift, which occurs when the motion is perpendicular to this line. This is interpreted as a pure time dilation effect, and is therefore of order $(v/c)^2$ and negligible when $v \ll c$.

Radial velocity of the Andromeda Galaxy

example 5

Atoms in stars and nebulae emit and absorb light at certain discrete frequencies, which are properties of the atoms they contain. By comparing with the frequencies observed in the laboratory, we can infer Doppler shifts. Light from the Andromeda Galaxy shows a Doppler shift such that the fractional shift $\Delta f/f$ is $D - 1 = +0.001001$, i.e., a shift that happens to be almost exactly a tenth



The visible-light spectrum of the star Sirius. The dark lines are due to absorption of light at certain frequencies in the star's outer atmosphere.

of a percent towards higher frequencies (a “blue-shift”). Solving for the velocity (problem 2, p. 107), we find $v = 1.000 \times 10^{-3}$. The positive sign means that this galaxy’s radial motion is toward us. In other words, the velocity is small compared to the speed of light, so that the fractional Doppler shift is essentially the same as v/c , as we would expect from the equation for the nonrelativistic Doppler shift.

Ives-Stilwell experiments

example 6

Measurements of Doppler shifts have provided some of the most precise tests of special relativity, with the aid of the following trick. The relativistic expression $D(v) = \sqrt{(1+v)/(1-v)}$ for the Doppler shift has the property that $D(v)D(-v) = 1$, which differs from the nonrelativistic result of $1/[(1+v)(1-v)] = 1/(1-v^2)$. Suppose we accelerate an ion up to a relativistic speed. (Ions are easier to manipulate than neutral atoms, because we can operate on them with electrical and magnetic forces.) If the ion emits an electromagnetic wave with a known frequency, we can then measure both the forward Doppler shifted frequency f_f and the backward one f_b , and compute $\sqrt{f_f f_b}$. According to relativity, this should exactly equal the frequency f_0 measured in the ion’s rest frame.

The earliest test using this idea was done by Ives and Stilwell in 1938. In a particularly exquisite modern version, Saathoff *et al.* circulated Li^+ ions at $v = .064$ in a storage ring. Since the identity $D(v)D(-v) = 1$ is independent of v , it was not necessary to measure v to the same incredible precision as the frequencies; it was only necessary that it be stable and well-defined. The resulting frequencies, in units of MHz, were:

$$\begin{aligned}f_f &= 582490203.44 \pm .09 \\f_b &= 512671442.9 \pm 0.5 \\\sqrt{f_f f_b} &= 546466918.6 \pm 0.3 \\f_0 &= 546466918.8 \pm 0.4 \text{ (from previous experimental work)}\end{aligned}$$

The spectacular agreement with theory has made this experiment a lightning rod for anti-relativity kooks.

If one is searching for small deviations from the predictions of special relativity, a natural place to look is at high velocities. Ives-Stilwell experiments have been performed at velocities as high as 0.84, and they also confirm special relativity.

Discussion question

- A A person in a spaceship moving at 99.9999999% of the speed of light relative to Earth shines a flashlight forward through dusty air, so the beam is visible. What does she see? What would it look like to an observer on Earth?

4.5 Length contraction

One of the earliest tests of relativity was an experiment by Rossi and Hall in 1941, in which they observed a certain type of subatomic particle, called a muon, produced by cosmic rays as they enter the upper atmosphere of the earth. A muon is essentially a heavier version of the electron. Muons undergo radioactive decay, lasting an average of only 2.197 microseconds before they evaporate into an electron and two neutrinos. The earth's atmosphere is several miles thick, so even traveling at near the speed of light a muon takes about ten or twenty microseconds to reach the ground. This is much longer than their average lifetime, so almost none should have gotten to Rossi and Hall's detectors. In fact they found a much higher flux of muons than should have been expected nonrelativistically. This was interpreted as a result of relativistic time dilation. Although ten or twenty microseconds passes for us, a much shorter time passes for the muons, which are going close to c .

Now what if we consider this from the point of view of the poor little muon? Newly born, it sees a planet's surface approaching. Maybe some inhabitant of this planet will love and cherish it. But wait. "Planetary atmosphere are several miles thick," it thinks to itself. "Even at nearly the speed of light, I won't make it through that in less than ten or twenty microseconds, which is much longer than I expect to live."

This seems to be a paradox, since the collision of the muon with the detector is an event, and all observers agree on whether events exist. But according to the muon, it's at rest, so it doesn't experience time dilation. So how can its survival be explained?

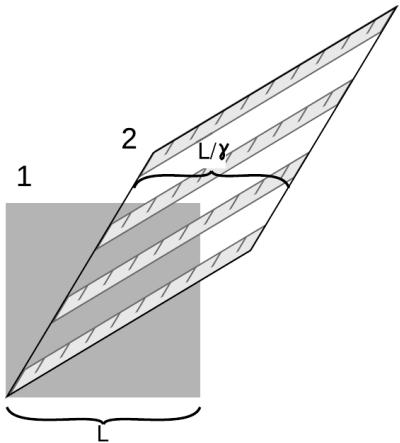
Our conclusion is that the distance in the muon's frame must be shorter than the distance in the earth's frame. This can be confirmed from an analysis of the Lorentz transformation, figure p. An easier shortcut to get the result is to equate the time dilation effect seen in the earth's frame to the length contraction factor seen in the muon's. If we let $\mathcal{I}^2 = t^2 - d^2$, set $d = vt$, and solve for t , we find that the time dilation factor t/\mathcal{I} is

$$\frac{1}{\sqrt{1 - v^2}}.$$

The standard notation for this quantity is γ (Greek letter gamma, which makes the "g" sound). We therefore conclude that the length contraction is by the same factor γ , i.e., that if a meter-stick has length L in its own frame of reference, then in a frame moving relative to it at speed v , its length is reduced to $L/\gamma = L\sqrt{1 - v^2}$.

self-check C

What happens to the equation for length contraction at $v = 0$? Why does this make sense? ▷ Answer, p. 455



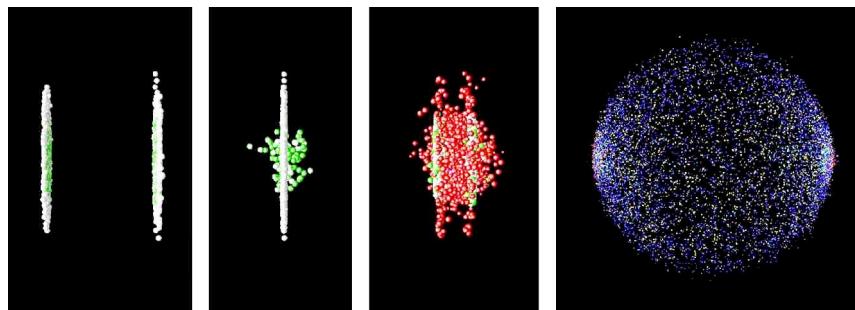
p / The ruler is shorter in the frame of the square graph paper, 1, than in its own frame 2. This happens because the observer in frame 1 judges the distance between two events that they say are simultaneous, but in the ruler's frame, the event at the tail end of the ruler is later, and therefore has had longer to travel and lies farther to the right.

Discussion question

A On a spaceship moving at relativistic speeds, would a lecture seem even longer and more boring than normal?

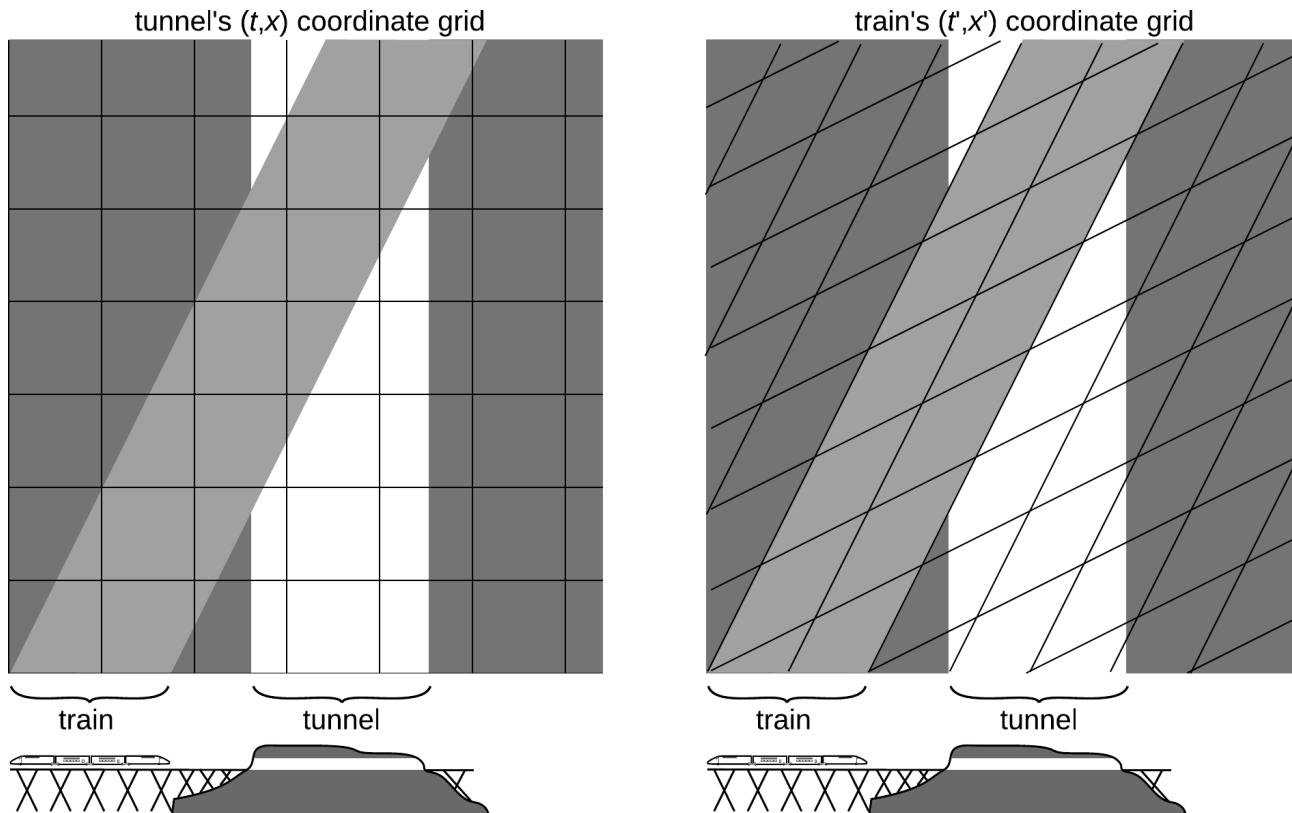
B Figure q shows an artist's rendering of the length contraction for the collision of two gold nuclei at relativistic speeds in the RHIC accelerator in Long Island, New York. The gold nuclei would appear nearly spherical (or just slightly lengthened like an American football) in frames moving along with them, but in the laboratory's frame, they both appear drastically foreshortened as they approach the point of collision. The later pictures show the nuclei merging to form a hot soup, observed at RHIC in 2010, in which the quarks are no longer confined inside the protons and neutrons.

What would the shapes of the two nuclei look like to a microscopic observer riding on the left-hand nucleus? To an observer riding on the right-hand one? Can they agree on what is happening? If not, why not — after all, shouldn't they see the same thing if they both compare the two nuclei side-by-side at the same instant in time?



q / Discussion question B: colliding nuclei show relativistic length contraction.

C Figure r shows a spacetime diagram for a train passing through a tunnel at half the speed of light. Rather than trying to show two coordinate grids on the same diagram, I've drawn two versions of the figure. Suppose we want to know whether the entire train ever fits inside the tunnel, i.e., whether the tail end enters the tunnel before the front exits. Identify the relevant events on the diagram. Compare the results as described in the two frames of reference.



r / Discussion question C.

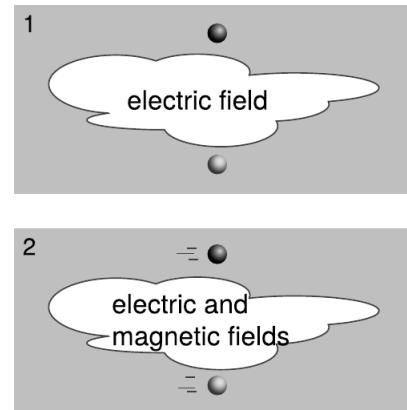
4.6 Magnetism as a relativistic effect

Magnetism is an interaction between moving charges and moving charges. But how can that be? Relativity tells us that motion is a matter of opinion. Consider figure s. In this figure and in figure t, the dark and light coloring of the particles represents the fact that one particle has positive charge and the other negative. Observer s/2 sees the two particles as flying through space side by side, so they would interact both electrically (simply because they're charged) and magnetically (because they're charges in motion). But an observer moving along with them, s/1, would say they were both at rest, and would expect only an electrical interaction. This seems like a paradox. Magnetism, however, comes not to destroy relativity but to fulfill it. Magnetic interactions *must* exist according to the theory of relativity. To understand how this can be, consider how time and space behave in relativity. Observers in different frames of reference disagree about the lengths of measuring sticks and the speeds of clocks, but the laws of physics are valid and self-consistent in either frame of reference. Similarly, observers in different frames of reference disagree about what electric and magnetic fields there are, but they agree about concrete physical events. An observer in frame of reference s/1 says there are electric fields around the particles, and predicts that as time goes on, the particles will begin to accelerate towards one another, eventually colliding. She explains the collision as being due to the electrical attraction between the particles. A different observer, s/2, says the particles are moving. This observer also predicts that the particles will collide, but explains their motion in terms of both an electric field and a magnetic field. The magnetic field is *required* in order to maintain consistency between the predictions made in the two frames of reference.

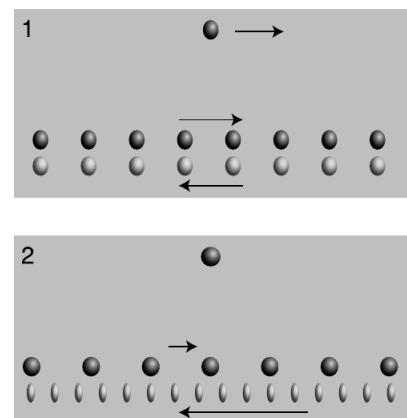
To see how this really works out, consider figure t. We have two long lines of charges moving in opposite directions. Note that the currents of the two lines of charges do not cancel out. The dark and light balls represent particles with opposite charges. Because of this, the total current in the “wire” is double what it would be if we took away one line.

What electrical force does the lone particle in figure t/1 feel? Since the density of “traffic” on the two sides of the “road” is equal, there is zero overall electrical force on the lone particle. Each “car” that attracts the lone particle is paired with a partner on the other side of the road that repels it. If we didn't know about magnetism, we'd think this was the whole story: the lone particle feels no force at all from the wire.

Figure t/2 shows what we'd see if we were observing all this from a frame of reference moving along with the lone charge. Here's where the relativity comes in. Relativity tells us that moving objects appear contracted to an observer who is not moving along with them.



s / One observer sees an electric field, while the other sees both an electric field and a magnetic one.



t / A model of a charged particle and a current-carrying wire, seen in two different frames of reference. The relativistic length contraction is highly exaggerated. The force on the lone particle is purely magnetic in 1, and purely electric in 2.

Both lines of charge are in motion in both frames of reference, but in frame 1 they were moving at equal speeds, so their contractions were equal. In frame 2, however, their speeds are unequal. The dark charges are moving more slowly than in frame 1, so in frame 2 they are less contracted. The light-colored charges are moving more quickly, so their contraction is greater now. The “cars” on the two sides of the “road” are no longer paired off, so the electrical forces on the lone particle no longer cancel out as they did in $t/1$. The lone particle is attracted to the wire, because the particles attracting it are more dense than the ones repelling it. Furthermore, the attraction felt by the lone charge must be purely electrical, since the lone charge is at rest in this frame of reference, and magnetic effects occur only between moving charges and other moving charges.

Now observers in frames 1 and 2 disagree about many things, but they do agree on concrete events. Observer 2 is going to see the lone particle drift toward the wire due to the wire’s electrical attraction, gradually speeding up, and eventually hit the wire. If 2 sees this collision, then 1 must as well. But 1 knows that the total electrical force on the lone particle is exactly zero. There must be some new type of force. She invents a name for this new type of force: magnetism. This was a particularly simple example, because the force was purely magnetic in one frame of reference, and purely electrical in another. In general, an observer in a certain frame of reference will measure a mixture of electric and magnetic fields, while an observer in another frame, in motion with respect to the first, says that the same volume of space contains a different mixture.

We therefore arrive at the conclusion that electric and magnetic phenomena aren’t separate. They’re different sides of the same coin. We refer to electric and magnetic interactions collectively as electromagnetic interactions. Our list of the fundamental interactions of nature now has two items on it instead of three: gravity and electromagnetism.

Discussion questions

A In the situation shown in figure t, is there a frame in which the force \mathbf{F} is a purely electric one, \mathbf{F}_E ? Pure \mathbf{F}_B ?

Is there a frame in which the electromagnetic field is a pure \mathbf{E} ? Pure \mathbf{B} ?

Is there zero net charge in both frames? One? Neither?

What about the current?

B For the situation shown in figure t, draw a spacetime diagram showing the positive charges as black world-lines and the negative as red, in the wire’s rest frame. Use a ruler, and draw the spacing fairly accurately. Interpret this in the frame of the lone charge.

C Resolve the following paradox concerning the argument given in this section. We would expect that at any given time, electrons in a solid would be associated with protons in a definite way. For simplicity, let's imagine that the solid is made out of hydrogen (which actually does become a metal under conditions of very high pressure). A hydrogen atom consists of a single proton and a single electron. Even if the electrons are moving and forming an electric current, we would imagine that this would be like a game of musical chairs, with the protons as chairs and the electrons as people. Each electron has a proton that is its "friend," at least for the moment. This is the situation shown in figure t/1. How, then, can an observer in a different frame see the electrons and protons as not being paired up, as in t/2?

Notes for chapter 4

288 Which direction to slant the x axis in a Lorentz transformation

In a Lorentz transformation, the x and t axes counterrotate like the blades of a pair of scissors.

In the reasoning leading up to the Lorentz transformation, how did we know in figure c to rotate the line of simultaneity counterclockwise, so that the two lines would close on each other like a pair of scissors? One way to see this is that if we rotated *both* lines clockwise, then it would be just like taking the page and rotated it by some angle. But then as we went to higher and higher velocities for Bill relative to Amy, eventually we would get to a 180-degree rotation, and Bill's time axis would be pointing downward, opposite to Amy's. This isn't what we see happening in reality, and leads to time-travel paradoxes, such as Bill's being able to go back into the past and kill Amy's grandfather when he was a child.

289 Conservation of area in a Lorentz transformation

In a Lorentz transformation, area in the x - t plane is conserved.

Suppose that a Lorentz transformation for a velocity v increases the area of a region in the x - t plane by a factor R_v . We want to show that $R_v = 1$. By symmetry, we expect that $R_v = R_{-v}$. We also know that if we do a Lorentz transformation for v and then another for $-v$, they undo each other, so $R_v R_{-v} = 1$. From these two equations we find that $R_v = \pm 1$. Since R_v should be a continuous function of v , and $R_0 = 1$, we must have $R_v = 1$ for all v . Note that this result also holds for the Galilean transformation, since we never employed any assumptions that would have distinguished Galilean relativity from Einstein's relativity.

292 c as a speed limit on cause and effect

We compare two different arguments leading

to stronger and weaker conclusions about c as a speed limit on cause and effect.

On p. 26 we made a simple argument that led to a weaker form of this conclusion. There we argued that if observers A and B are in motion relative to one another, and if all frames of reference are equally valid, then time dilation leads to paradoxes if there is instantaneous communication. If A and B can communicate instantaneously, then they can establish whose time is actually faster, even when they are far apart. This breaks the symmetry between them, and contradicts the assumption that all frames of reference are equally valid.

At that point we had not yet discussed the fact that simultaneity is relative. Knowing this, we can now see that the notion of "instantaneous" communication is not even well defined. Instantaneous communication would mean sending a signal from event P to event Q when P and Q are simultaneous. But different observers do not even agree on whether this is the case. So the argument on p. 26 was really more like an argument that there should be no frame of reference that has special status, such as the frame of the aether. Such a special frame of reference is referred to as a *preferred* frame of reference. In relativity, there is no preferred frame.

The argument on p. 92 leads to a stronger conclusion. There we argued that if the displacement between events P and Q is spacelike, then there will be some frames of reference in which P happens before Q, some in which they are simultaneous, and some in which Q happens before P. Therefore we get paradoxes about cause and effect if P causes Q or Q causes P. This is a stronger conclusion because it prohibits not just instantaneous cause-and-effect at a distance, but any propagation of cause and effect at a speed greater than c .

294 Combination of velocities

We present some shortcuts for combining velocities

When we do two boosts, the diagonal stretch

factors multiply, $D = D_1 D_2$. This implicitly relates v_1 and v_2 to the combined velocity v . Usually we prefer to deal with additive rather than multiplicative quantities, so it can be useful to define the *rapidity* $\eta = \ln D$. (The symbol η is Greek letter “eta”). For small velocities, $\eta \approx v$, so velocities are approximately additive.

Sometimes we would rather just work with velocities rather than with D 's and η 's. One can then show that

$$v = \frac{v_1 + v_2}{1 + v_1 v_2}.$$

295 The Doppler shift is D

We prove that the diagonal stretch factor D is the same as the relativistic Doppler shift.

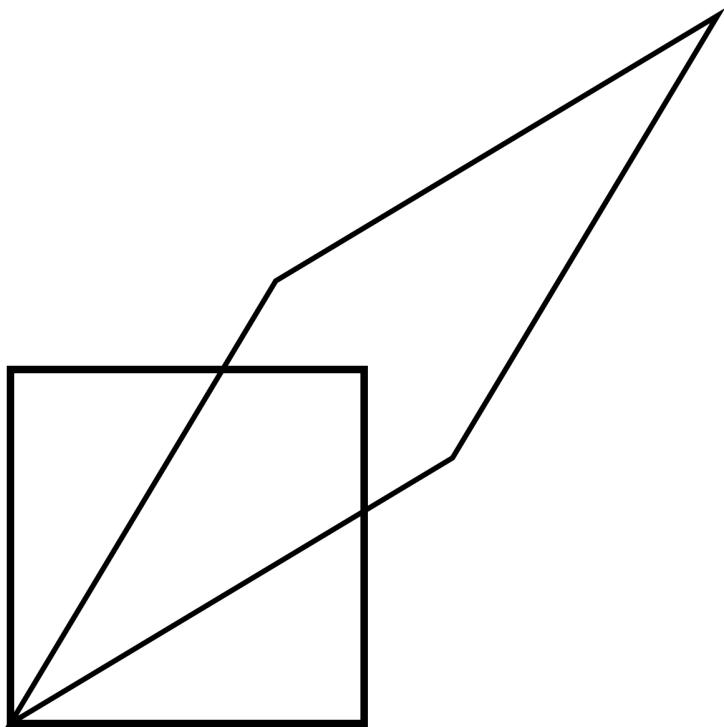
Suppose A emits a wave with a certain frequency, which is then received by both B and C. A, B, and C can all be in motion relative to one another. Since the aether doesn't exist, the only velocities that can determine the relevant Doppler shifts are the velocities of A, B, and C *relative to each other*. Suppose, for clarity of exposition, that the wave passes by B first and then later gets to C. That is, the situation is just as if B received the wave and then retransmitted it. Then the Doppler shifts for B relative to A and C relative to B must multiply. Now we also know that the D factors behave multiplicatively, so we must have (Doppler shift) = D^α for some exponent α . But we need the relativistic Doppler shifts to match up with the nonrelativistic equations for $v \ll c$, so this fixes α . A comparison of the Taylor series shows that in order to get agreement to first order in v , we need $\alpha = 1$.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 The figure illustrates a Lorentz transformation using the conventions employed in section 4.2. For simplicity, the transformation chosen is one that lengthens one diagonal by a factor of 2. Since Lorentz transformations preserve area, the other diagonal is shortened by a factor of 2. Let the original frame of reference, depicted with the square, be A, and the new one B. (a) By measuring with a ruler on the figure, show that the velocity of frame B relative to frame A is $0.6c$. (b) Print out a copy of the page. With a ruler, draw a third parallelogram that represents a second successive Lorentz transformation, one that lengthens the long diagonal by another factor of 2. Call this third frame C. Use measurements with a ruler to determine frame C's velocity relative to frame A. Does it equal double the velocity found in part a? Explain why it should be expected to turn out the way it does. ✓



2 In example 5, p. 96, and section 4.4.2, p. 94, we saw the usefulness of being able to find v in terms of D . Derive this equation.

✓

3 Prove the approximation $\gamma \approx D/2$ for $v \approx 1$.

4 The relativistic factor

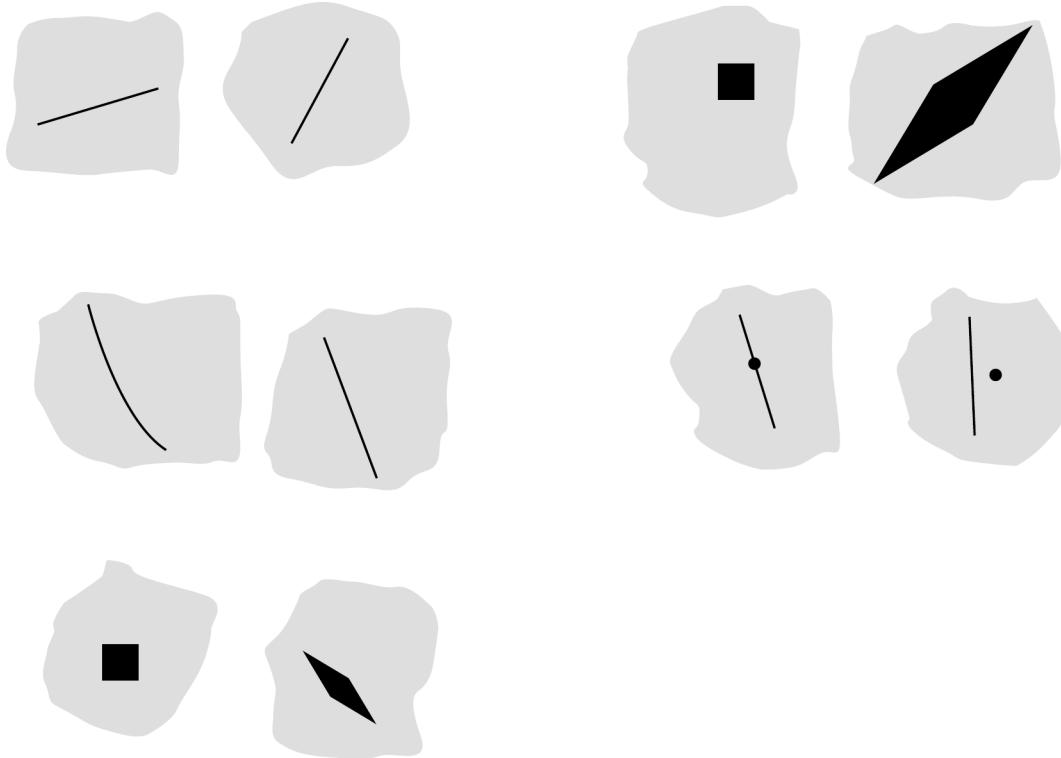
$$\gamma = \frac{1}{\sqrt{1 - v^2}},$$

introduced on p. 98, gives the amount of length contraction and time dilation.

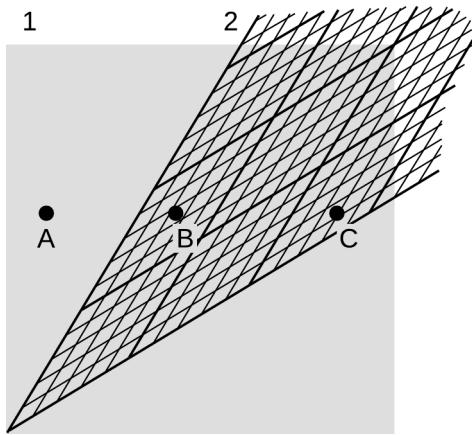
- (a) Find the first two nonvanishing terms in its Taylor series, and show that the first non-constant term is of order v^2 . Why would it not make sense physically to have a term of order v ? ✓
- (b) Take your expression from part a and insert factors of c in the appropriate places in order to make it valid in SI units. ✓
- (c) The earth is orbiting the sun, and therefore is contracted relativistically in the direction of its motion. Compute the amount by which its diameter shrinks in this direction. ✓

Exercise 4A: The Lorentz transformation

1. These are five pairs of spacetime diagrams, organized so that there is some resemblance between the left and right of each pair. In each case, decide whether the two could actually be the same thing represented in a different inertial frame of reference. Write yes or no, and explain why.

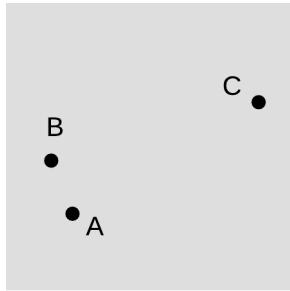


2. The diagram shows three events and two frames of reference. Describe the time-ordering of the events in these frames. Is any other time-ordering possible in any other frame?



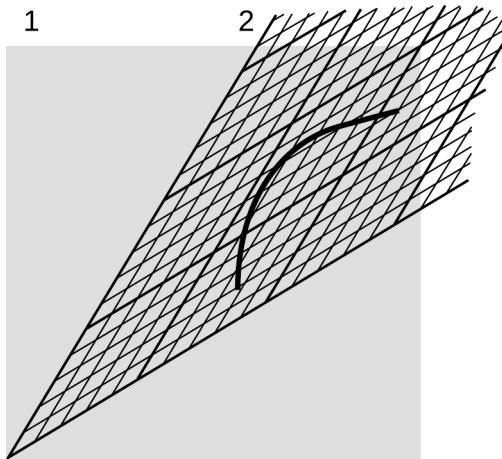
Turn the page.

3. (a) The figure shows three events and a square representing the t and x axes of a frame of reference. In this frame, the time-ordering of the events is ABC. If we switch to another frame, what other orderings, if any, are possible?



- (b) What can you say about \mathcal{I}_{AB}^2 ? Relate this to the orderings that are possible and not possible.

4. The figure shows the motion of an object, with two different frames of reference superimposed. Is anything strange going on?





A diver photographed this fish, and its reflection, from underwater. The reflection is the one on top, and is formed by light waves that went up to the surface of the water, but were then reflected back down into the water.

Chapter 5

Waves done medium well

So far we've implicitly assumed that our waves lived in a medium that was as uniform and boring as a housing tract in South Orange County, California. We don't usually encounter media with such simple properties. Nature doesn't make such a medium common. She makes such a medium rare. Instead of describing a wave's environment in such a bad and oversimplified way, we will now try to describe our medium well.

As we consider a variety of waves and their media, we will find that many of the same phenomena pop up over and over again in similar ways. This happens not just for mechanical waves but also for electromagnetic waves, which might seem strange because an electromagnetic wave can exist in a vacuum, and is not a vibration of any material substance such as the mythical aether. But light waves do pass through substances such as air, water, or the lens of your eye. We can consider these as media for the electromagnetic wave, with a vacuum being just one more "medium" — the one that is there in the case where nothing is actually there. The speed of an electromagnetic wave in a medium is less than c , so when we characterize c as "the speed of light," we really mean the speed of

light in a vacuum.

5.1 Measures of amplitude

We would use completely different units to measure the amplitude of a light wave than the ones we would use for a sound wave. Even when we fix our attention on a particular type of wave and a particular medium, we very commonly find that there are exactly two different measures of amplitude that would both seem equally sensible. Here are some examples:

sound waves	pressure	flow rate
waves on a string	y (transverse displ.)	v_y (transverse vel.)
light waves	E	B

The reason we see this situation repeatedly seems to be that energy has to be conserved when waves superpose, as in the example of sec. 3.2.2, p. 68, where two electromagnetic waves collide head-on. If there were only one amplitude to talk about, and only one form of energy, then the energy would have to vanish when the waves canceled. Instead, we have waves with two types of energy, such as electrical and magnetic energy, or potential and kinetic energy. Just as we never see plants and animals in nature that have no way of reproducing, we never see waves that lack some such mechanism of conserving energy when they superpose.

5.2 Impedance

When we have more than one measure of the amplitude of a wave, they are not totally independent things. For example, the electric and magnetic fields of a sinusoidal plane wave, in vacuum, have essentially the *same* amplitude, assuming we ignore a factor of c , which is there only in a system of units like the SI where $c \neq 1$. If we take the same wave and transplant it into some other medium, such as glass, then the ratio of the fields will still be fixed, but it will have some other value. We call such a ratio the *impedance* Z of the medium. The word and symbol are recycled from the study of electric circuits, but we apply them in cases like sound where there is nothing electrical going on. In the case of sound waves, the implied analogy is physically appealing: pressure is like voltage, and flow rate is like current. As in the electrical case, the wave impedance only makes sense when we talk about sine waves with a specific frequency.

If we have two different measures of amplitude, A_1 and A_2 , then there will be two different ways we can flip the ratio that defines the impedance: A_1/A_2 or A_2/A_1 . Most of the interesting physics doesn't depend on this choice, and I don't have it memorized for the different types of waves.

When a wave is transporting energy, we can define an energy flux, such as the Poynting vector \mathbf{S} introduced on p. 73 for electromagnetic waves. A loud sound or a big ocean wave has a big S . The energy *density* of the wave is proportional to the square of the wave's amplitude, with some constant of proportionality that depends on the medium. For example, a one-meter-high wave in a pool of water will have less energy than a wave of the same height in a pool of mercury. The energy flux is proportional to the energy density, but also to the wave's speed. Putting all of these factors together, we generally get something that can be expressed like either

$$S = (\text{const.})ZA_1^2$$

or

$$S = (\text{const.})\frac{1}{Z}A_2^2,$$

where A_1 and A_2 are our two different measures of amplitude, $Z = A_2/A_1$, and the “const.” factors are the same in the two cases.

The Poynting vector in a vacuum

example 1

The magnitude of the Poynting vector for an electromagnetic plane wave in vacuum is

$$\frac{c^2}{4\pi k}EB,$$

and since $B = E/c$, this can be expressed as either

$$(\text{const.})cB^2$$

or

$$(\text{const.})\frac{1}{c}E^2.$$

Thus if we take our two measures of amplitude to be E and B , then we can define the impedance of the vacuum to be either the speed of light or one over the speed of light.

In example 1, the impedance basically works out to be the speed (or its inverse), and this is very common. In many cases we can treat “impedance” and “wave speed” as synonyms for describing the properties of a medium. As an example where this fails, it will not work when describing an electromagnetic wave in a medium such as iron that has magnetic properties.

The table on p. 112 listed some examples of measures of amplitude that are familiar and easy to measure. If we want to get the most full and detailed possible use of the system of analogies for the different types of waves, then in some cases we need to use variables that differ from that list. Since this kind of thing will probably not be of interest to most readers, I've relegated it to a note, [Z139](#).

5.3 More about reflection and transmission

5.3.1 Why reflection happens

We frequently observe reflection of waves in everyday life. Examples include echoes of sound and the reflection of light from a lake. In this section we'll go into a little more mathematical detail on these phenomena, which were only briefly described in sec. 2.6. Let's consider the simplest geometry, where the wave comes in along a line perpendicular to the surface. For concreteness, let's say that this is a light wave of sunlight coming straight down and hitting the surface of some water. The wave starts in one medium, air, and then enters another one, water. To keep the equations simple, we'll say that the amplitude of the incident wave, arriving through the air, is just a unitless 1, i.e., all other amplitudes will be expressed relative to this one.

Why do we get reflection? Let's see what goes wrong if we assume there is none. The wave is simply transmitted into the water, where it has some amplitude T . In almost all cases, we have some measure of amplitude that we expect to be a continuous function at the boundary. In this particular example, it's not hard to show using Faraday's law that the electric field must be continuous, so let's use that as our measure of amplitude. (A simpler example would be a wave on a string, for which clearly the wave's height must be continuous, if the string hasn't broken.) Continuity requires that we have

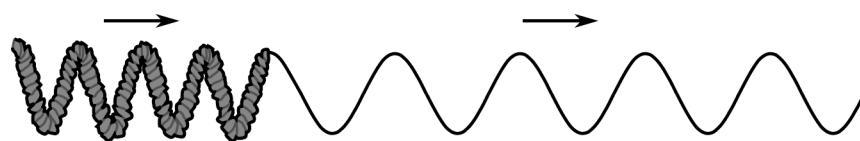
$$1 = T. \quad [\text{wrong}]$$

a / This is impossible. A sine wave is completely transmitted from the heavy rope into the skinny one. The skinny rope has $1/4$ the mass per unit length. The amplitude of the transverse velocity is the same in both ropes, because otherwise a discontinuity would develop at the boundary. This means that the kinetic energy density is $1/4$ as much. The same is true for the potential energy density. There is a change in the energy flux by a factor of $1/2$, since $\text{energy}/\text{time} = (\text{energy}/\text{distance})(\text{distance}/\text{time})$. The result is that half the energy has disappeared, violating conservation of energy.

We also need conservation of energy. If there is no reflection, then 100% of the energy of this wave must pass on into the water, i.e., the energy flux must be equal on both sides of the boundary. But this is a different condition than $1 = T$, since the energy flux depends not just on the squared amplitude but also on the medium. Since water and air are nonmagnetic substances, the impedance here is really the same as the speed of the wave. Let subscript 1 stand for air and 2 for water, and $\alpha = v_1/v_2 = Z_1/Z_2$. Then conservation of energy requires

$$1 = \alpha T^2. \quad [\text{wrong}]$$

It is impossible to satisfy these two equations simultaneously, because waves with the same amplitude have *different* energy fluxes.



5.3.2 How much reflection?

Only by putting in a partial reflection of the wave's energy do we get equations that have an actual solution. Let the amplitude of the reflected wave be R . Then the reflected and incident waves superpose in the air just above the water's surface, and continuity gives

$$1 + R = T.$$

Conservation of energy now reads as

$$1 = \alpha T^2 + R^2,$$

where the left-hand side is the energy arriving at the surface and the right-hand side is the energy departing in the upward and downward directions. We now have two equations in two unknowns, which means we should expect a solution to exist. The result turns out to be

$$R = \frac{1 - \alpha}{1 + \alpha} \quad \text{and} \quad T = \frac{2}{\alpha + 1}.$$

self-check A

What happens in these equations when $\alpha = 1$, and why does this make sense physically?

▷ Answer, p. 455

Fish have internal ears.

example 2

Why don't fish have ear-holes? Body tissues of animals are mostly water, so the impedance of body tissues for sound waves is generally about the same as for water. For this reason, sound waves are not strongly reflected from a fish's skin. They pass right through its body, so fish can have internal ears.

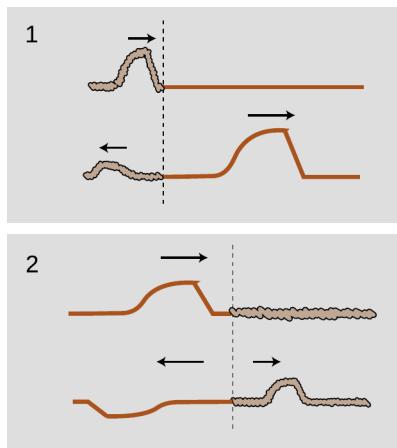
Duality: Suppose the impedance ratio α has some value like 3. There are then two different physical reasons why it may be of interest to consider as well what happens when α is $1/3$.

1. Interchanging the roles of the two media changes α to $1/\alpha$.
2. Depending on the choice of which amplitude we mean by " A_1 " and which one we mean by " A_2 ," we can also invert α .

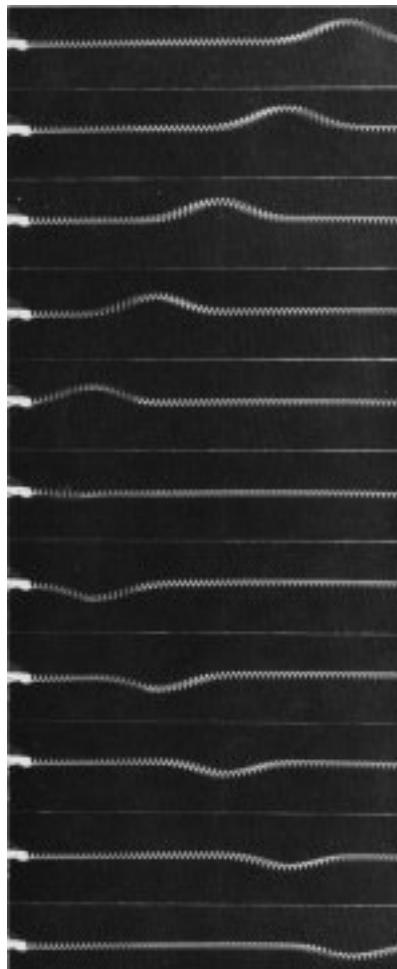
As an example of number 1, the underwater photo on p. 111 shows reflections visible from underwater, which result from the partial reflection of light back into the water. But we could also view the surface of the water, as we humans normally do, from above, in which case we see reflections of light back into the air.

Case 2 occurs because we usually prefer definitions of amplitude A_1 and A_2 such that the energy flux comes out looking like $S = (\text{const.})ZA_1^2 = (\text{const.})\frac{1}{Z}A_2^2$ and $Z = A_2/A_1$, but interchanging our definitions of A_1 and A_2 then flips $\alpha = Z_1/Z_2$.

There doesn't seem to be any standard name for these ideas in this specific case, but more generally this sort of relationship is called a *duality*, so I'll use that term as a shorthand label.



b / 1. An uninverted reflection. 2. An inverted reflection.



c / A wave on a spring, initially traveling to the left, is reflected from the fixed end.

5.3.3 Inverting and uninverting reflections

When $\alpha > 1$, we get $R < 0$. This is called an inverting reflection. The reflected wave comes back with its amplitude reversed in sign, i.e., the wave is upside-down. A good example is provided if you walk up to a taut rope such as a clothesline and give it a karate chop at one end. The pulse travels to the other end and then returns upside-down. This reflection has to be inverting because the rope is attached at the ends, so it can't move there. The reflected wave superposes with the incident wave and cancels it at that point, giving zero motion.

Figure b shows idealized cartoons of how this plays out for a wave on a string, which is easy to visualize. Figure d shows snapshots of a real-world experiment.

self-check B

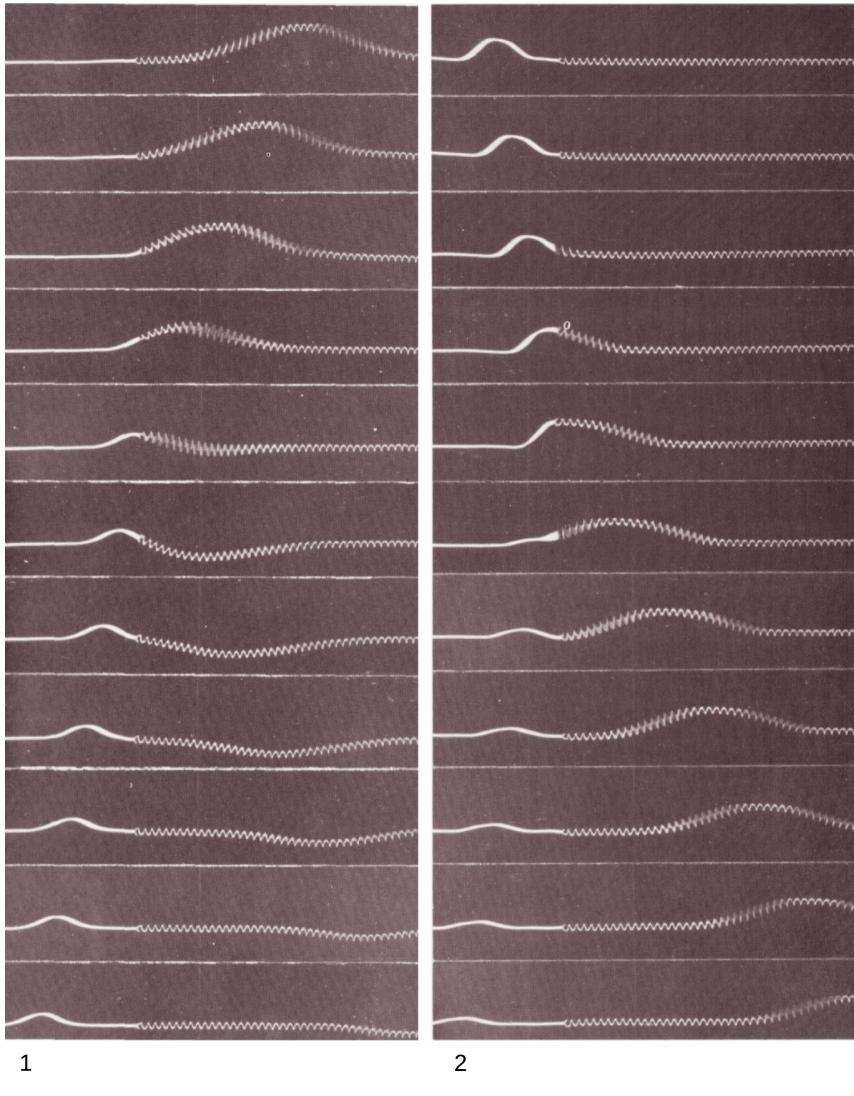
In figure c, the reflected pulse is upside-down, but its depth is just as big as the original pulse's height. How does the energy of the reflected pulse compare with that of the original? ▷ Answer, p. 455

When we flip α to make $1/\alpha$ (“duality,” p. 115), the value of R changes (problem 2, p. 140). In our example of the light wave hitting the lake, the speed of light in the water is lower than in the air, so $\alpha > 1$ and the electric field is not inverted. But if we had picked the magnetic field as our measure, then we would have had $\alpha = v_2/v_1 = Z_2/Z_1 < 1$ and $R < 0$. This tells us that in this example, the reflected light wave has an uninverted electric field and an inverted magnetic field. This makes sense, because the wave propagates in the direction of the Poynting vector $\mathbf{S} \propto \mathbf{E} \times \mathbf{B}$. The reflected wave, rising back up toward the sun, needs to have an upward \mathbf{S} , i.e., it needs to be flipped relative to the \mathbf{S} of the wave that came down originally. This does happen because $\mathbf{E} \times (-\mathbf{B}) = -(\mathbf{E} \times \mathbf{B}) =$.

So in summary, we can have reflections back into the higher-impedance medium or back into the lower-impedance medium. Once we pick a definition of amplitude, one of these will be inverting and one noninverting. But because of the duality between different measures of amplitude, it is not worthwhile to memorize which of these is inverting and which is noninverting. What is worth knowing is that one is and one isn't.

5.3.4 Total reflection

In the limit of very dissimilar media, we get total reflection. But there are actually two different versions of this limit: the one where $\alpha = Z_2/Z_1 \rightarrow 0$ and the one where $\alpha \rightarrow \infty$. As an example, sound waves in steel experience an impedance that is greater than the impedance of air by a factor of about 10^5 , so when we calculate the reflection coefficient $R = (\alpha - 1)/(\alpha + 1)$ for sound waves at a



- d / 1. A wave in the lighter spring, where the wave speed is greater, travels to the left and is then partly reflected and partly transmitted at the boundary with the heavier coil spring, which has a lower wave speed. The reflection is inverted.
 2. A wave moving to the right in the heavier spring is partly reflected at the boundary with the lighter spring. The reflection is uninverted.

boundary between air and steel, we will get something either very close to 1 or very close to -1 . $R^2 = 1$, so all of the energy is reflected.

In the case of $R = -1$, the reflected wave will exactly cancel the incident wave at the boundary. This happens, for example, at the end of a guitar string if we use the displacement y as a measure of amplitude. The end of the string is tied down, so it can't move.

In the case of $R = 1$, the amplitude achieves a *maximum* at the boundary. To see this, consider a wave propagating along the x axis, with amplitude y . The reflection flips the function $y(x)$ in the left-right direction, which causes the derivative y' to flip its sign. The derivatives of the incident and reflected wave cancel, so $y' = 0$. We know from calculus that this is the condition for y to be at an extremum.

When a standing-wave pattern has a point that doesn't vibrate,

like the ends of the guitar string, we call that point a *node*. Harmonics above the first have nodes at additional points in the middle of the string, so that, e.g., the pattern $\bigcirc\!\!\!\bigcirc$ has a total of three. An extremum of the vibration is called an *antinode*, and the pattern we were just discussing has two of these.

5.4 Symmetric and asymmetric standing wave patterns

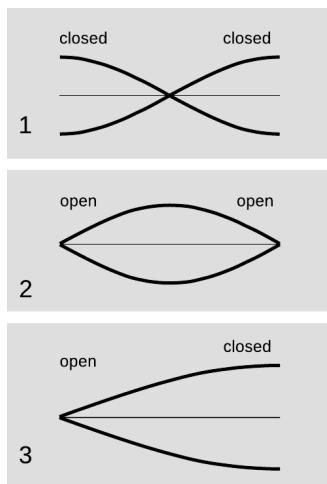
We discussed standing waves in sec. 2.6, p. 49, but there we only considered the case where $R = -1$ at both ends. Actually we can have standing waves with all the different combinations: $R = -1$ at both ends, $R = 1$ at both ends, or the asymmetric case with $R = 1$ at one end and $R = -1$ at the other. An important application is to the physics of wind instruments.

Some organ pipes are closed at both ends. The impedance is very different in metal than in air, so there is a strong reflection at the closed ends, and we can have standing waves. These reflections are both pressure-noninverting, so we get symmetric standing-wave patterns, such as the one shown in figure e/1. This is the longest possible wavelength, having the lowest possible frequency f_o .

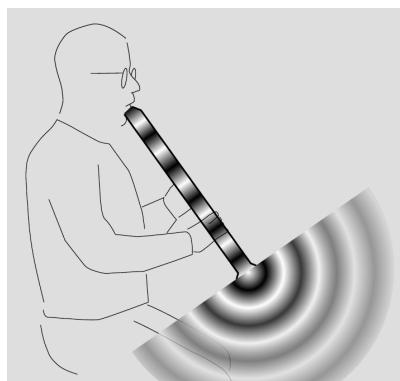
Figure f shows the sound waves in and around a bamboo Japanese flute called a shakuhachi, which is *open* at both ends of the air column. We can only have a standing wave pattern if there are reflections at the ends, but that is very counterintuitive — why is there any reflection at all, if the sound wave is free to emerge into open space, and there is no change in medium? The wave is adjusting from being a plane wave to being a spherical wave, and the impedance is unequal for these two cases, because the amount of flow for a given pressure is not the same for the tube as for open space. The reflections at the open ends are pressure-inverting, e/2, so the wave pattern is pinched off at the ends. As we saw in sec. 2.6, there will be higher harmonics like $\bigcirc\!\!\!\bigcirc$ $\bigcirc\!\!\!\bigcirc\!\!\!\bigcirc$, with successively shorter wavelengths and higher frequencies. The frequencies form the pattern $f_o, 2f_o, 3f_o, \dots$

Comparing panels 1 and 2 of the figure, we see that although the wave patterns are different, in both cases the wavelength is the same: in the lowest-frequency standing wave, half a wavelength fits inside the tube. In fact, 1 and 2 could be two different representations of the same wave, if we used two different measures of amplitude. This just interchanges the nodes and antinodes, without changing the sequence of wavelengths or the pattern of the frequencies $f_o, 2f_o, 3f_o, \dots$

Finally, we can have an asymmetric tube: closed at one end and open at the other. A common example is pan pipes, which are closed



e / Graphs of excess pressure versus position for the lowest-frequency standing waves of three types of air columns. Points on the axis have normal air pressure.



f / Surprisingly, sound waves undergo partial reflection at the open ends of tubes as well as closed ones.

at the bottom and open at the top. The standing wave with the lowest frequency is therefore one in which $1/4$ of a wavelength fits along the length of the tube, as shown in figure e/3. The standing wave patterns for the asymmetric tube are *not* at all equivalent to the symmetric case, nor is the pattern of frequencies the same. This is left as an exercise in self-check C and problem 5.

Sometimes an instrument's physical appearance can be misleading. A concert flute, g, is closed at the mouth end and open at the other, so we would expect it to behave like an asymmetric air column; in reality, it behaves like a symmetric air column open at both ends, because the embouchure hole (the hole the player blows over) acts like an open end. The clarinet and the saxophone look similar, having a mouthpiece and reed at one end and an open end at the other, but they act different. In fact the clarinet's air column has patterns of vibration that are asymmetric, the saxophone symmetric. The discrepancy comes from the difference between the conical tube of the sax and the cylindrical tube of the clarinet. The adjustment of the wave pattern from a plane wave to a spherical wave is more gradual at the flaring bell of the saxophone.

self-check C

Draw a graph of pressure versus position for the second harmonic of the air column in a tube open at one end and closed at the other. This will be the next-to-longest possible wavelength that allows for a point of maximum vibration at one end and a point of no vibration at the other. How many times shorter will its wavelength be compared to the wavelength of the lowest-frequency standing wave, shown in figure e/3? Based on this, how many times greater will its frequency be? ➤ Answer, p. 455

5.5 ★ Musical consonance and dissonance

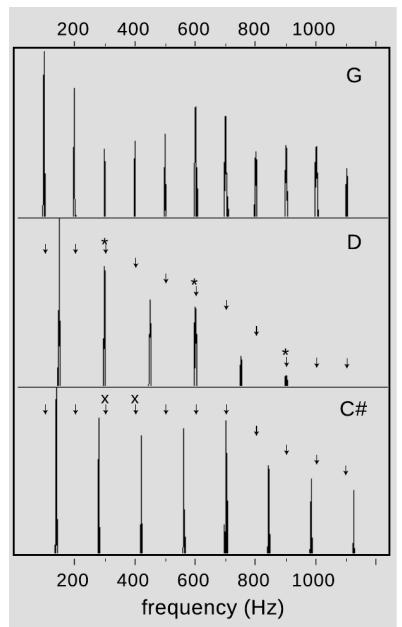
Many musicians claim to be able to pick out by ear several of the frequencies $2f_o$, $3f_o$, ..., called overtones or *harmonics* of the fundamental f_o , but they are kidding themselves. In reality, the overtone series has two important roles in music, neither of which depends on this fictitious ability to “hear out” the individual overtones.

First, the relative strengths of the overtones is an important part of the personality of a sound, called its timbre (rhymes with “amber”). The characteristic tone of the brass instruments, for example, is a sound that starts out with a very strong harmonic series extending up to very high frequencies, but whose higher harmonics die down drastically as the attack changes to the sustained portion of the note.

Second, although the ear cannot separate the individual harmonics of a single musical tone, it is very sensitive to clashes between the overtones of notes played simultaneously, i.e., in harmony. We tend to perceive a combination of notes as being dissonant if they



g / A concert flute looks like an asymmetric air column, open at the mouth end and closed at the other. However, its patterns of vibration are symmetric, because the embouchure hole acts like an open end.



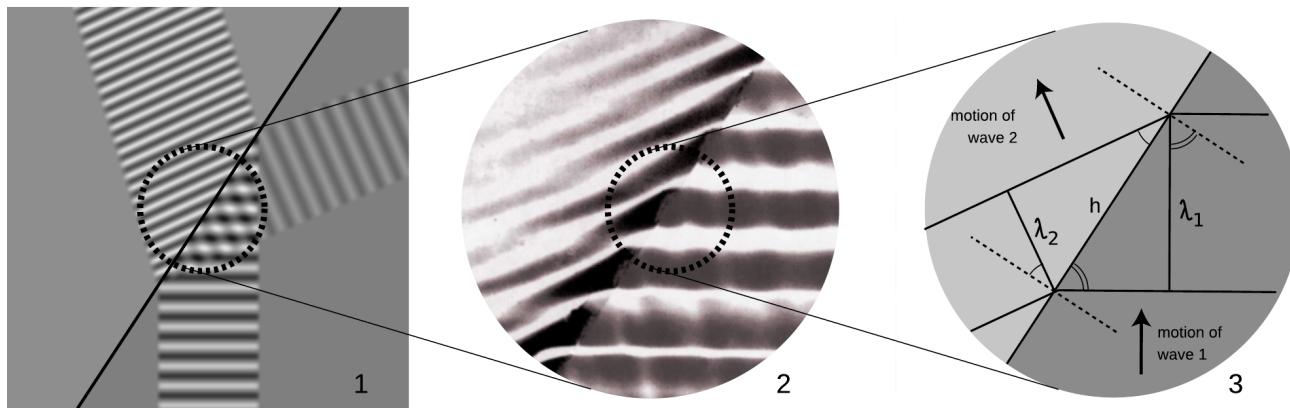
h / Graphs of loudness versus frequency for the vowel “ah,” sung as three different musical notes. G is consonant with D, since every overtone of G that is close to an overtone of D (marked “**”) is at exactly the same frequency. G and C \sharp are dissonant together, since some of the overtones of G (marked “x”) are close to, but not right on top of, those of C \sharp .

have overtones that are close but not the same. Roughly speaking, strong overtones whose frequencies differ by more than 1% and less than 10% cause the notes to sound dissonant. It is important to realize that the term “dissonance” is not a negative one in music. No matter how long you search the radio dial, you will never hear more than three seconds of music without at least one dissonant combination of notes. Dissonance is a necessary ingredient in the creation of a musical cycle of tension and release. Musically knowledgeable people do not usually use the word “dissonant” as a criticism of music, and if they do, what they are really saying is that the dissonance has been used in a clumsy way, or without providing any contrast between dissonance and consonance.

5.6 Refraction and reflection in two dimensions

5.6.1 Snell's law

Up until now we've been considering only the case in which a wave approaches a boundary along the direction perpendicular to it. A line perpendicular to the boundary is called a normal (“normal” being just a fancy word for “perpendicular,” as in the normal force from first-semester mechanics). For normal incidence, the problem of describing reflection and transmission is essentially a one-dimensional one. It's natural to consider what happens if the incident wave instead comes in at some nonzero angle θ_1 with respect to the normal. Figure i shows an example with water waves.



i / A derivation of Snell's law. The figure shows a series of successive zooms. 1. A simulation of the incident, reflected, and transmitted waves. 2. A close-up near the boundary. 3. A further zoom in, showing a single wavelength.

The picture breaks down into four quadrants , with the incident wave I coming in from one quadrant and the transmitted wave

T and reflected wave R heading out into two others, .

The reflected wave is faint and hard to see in this picture, which is all right because we'll concentrate on the transmitted one. We just note in passing that the direction of propagation of the reflected wave is at an angle with respect to the normal that is equal to θ_1 , but on the other side of the normal.

Concentrating on the transmitted wave, we see that its wavefronts are crowded closer together, i.e., its wavelength is shortened. This is because the velocity v_2 of the waves in the new medium is lower than the original velocity v_1 . Since the frequency doesn't change, $\lambda \propto v$. We can also see that the wavefronts are *bent* at the boundary. They have to be, because the transmitted wavefronts have to join up continuously with the incident ones, whose vibration is after all what is causing them. This bending is called *refraction*, from the same root as the word "fracture," as in a fractured bone.

Usually we're actually interested in the direction of propagation of the wave, which is perpendicular to the wavefronts. As shown by the arrows in the close-up in figure i, there is also a refraction of this direction — it twists by the same amount as the wavefronts. We now want to find the direction θ_2 of the refracted ray. In the close-up view, the dashed lines are normals to the interface. The two marked angles on the right side are both equal to θ_1 , and the two on the left to θ_2 . Trigonometry gives

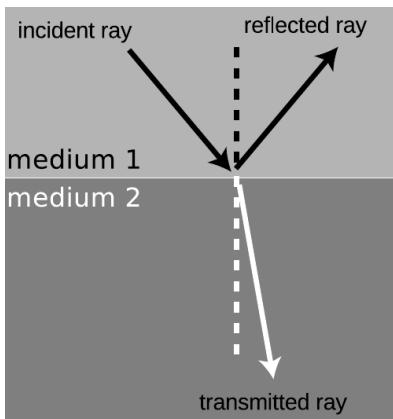
$$\begin{aligned}\sin \theta_1 &= \lambda_1/h && \text{and} \\ \sin \theta_2 &= \lambda_2/h,\end{aligned}$$

and combining this with $\lambda \propto v$ gives

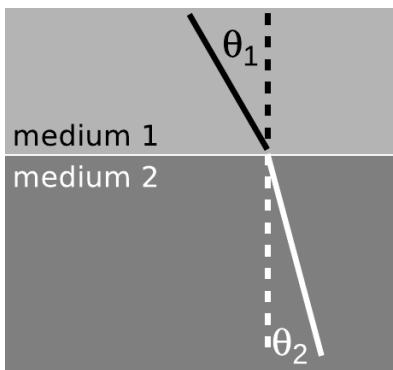
$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{v_1}{v_2},$$

which is known as Snell's law.

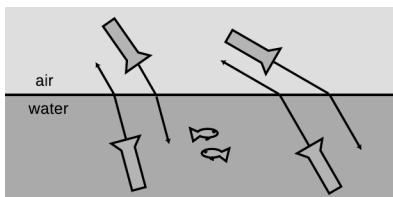
<p><i>Ocean waves near and far from shore</i></p> <p>Ocean waves are formed by winds, typically on the open sea, and the wavefronts are perpendicular to the direction of the wind that formed them. At the beach, however, you have undoubtedly observed that waves tend come in with their wavefronts very nearly (but not exactly) parallel to the shoreline. This is because the speed of water waves in shallow water depends on depth: the shallower the water, the slower the wave. Although the change from the fast-wave region to the slow-wave region is gradual rather than abrupt, there is still refraction, and the wave motion is nearly perpendicular to the normal in the slow region.</p>	<p><i>example 3</i></p>
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------



j / The incident, reflected, and transmitted (refracted) rays all lie in a plane that includes the normal (dashed line).



k / The angles θ_1 and θ_2 are related to each other, and also depend on the properties of the two media. Because refraction is time-reversal symmetric, there is no need to label the rays with arrowheads.



l / Refraction has time-reversal symmetry. Regardless of whether the light is going into or out of the water, the relationship between the two angles is the same, and the ray is closer to the normal while in the water.

Refraction occurs with any wave, but refraction of light waves is particularly important, since it is the physical principle behind the lens. You're reading this book through the lenses of your eyes. This kind of optical application is inherently three-dimensional, unlike the case of surface waves on water. Usually in optics we can ignore the wave nature of light and model it instead as a ray, like the beam of a pen-pointer laser. Taking the emission of light from a candle flame as an example, the ray model would describe it not as concentric spherical wavefronts but rather as a porcupine of rays spreading out in all directions from the source. Considering refraction in the ray model, in three dimensions, we have a plane of incidence, j , which contains the incident ray and the normal. The reflected and transmitted rays also lie in this plane. Figure k shows how the angles θ_1 and θ_2 occur in diagonally opposite quadrants of this plane. Refraction is time-reversal symmetric, l, so that it doesn't matter which medium we take to be which in Snell's law.

5.6.2 The index of refraction

During Snell's lifetime (1580-1626), it had not even been established that light was a wave or that it had a finite speed. It's impressive with historical hindsight that anything at all could have been done to probe something as fast as the motion of light, decades before anyone even invented the pendulum clock. At the time, people like Snell simply used data on refraction to characterize different substances, and they did it using a variable that we would today define as $n = c/v$. By this definition, we have $n = 1$ for a vacuum, and larger values of n in matter. Usually (not always), substances with a higher mass density have higher values of n , so can we refer to n as the optical density of the medium. It is also called the index of refraction. The index of refraction tells us about the (inverse) speed of light in a medium, but only in relative terms, not in any units such as s/m. The index of refraction of air at normal atmospheric pressure is 1.0003, so for most purposes it is a good approximation to assume that air has $n = 1$.

When we write Snell's law in terms of the indices of refraction, we usually express it as

$$n_1 \sin \theta_1 = n_2 \sin \theta_2,$$

which makes us less likely to get the 1's and 2's mixed up.

The mechanical model shown in figure m is a helpful way to remember the qualitative result of Snell's law. Suppose medium 2 is thick, sticky mud, which slows down the car. The car's right wheel hits the mud first, causing the right side of the car to slow down. This will cause the car to turn to the right until it moves far enough forward for the left wheel to cross into the mud. After that, the two sides of the car will once again be moving at the same speed, and the car will go straight. The result of this model is always the same

as the result of Snell's law: the ray is closer to the normal in the more optically dense medium.

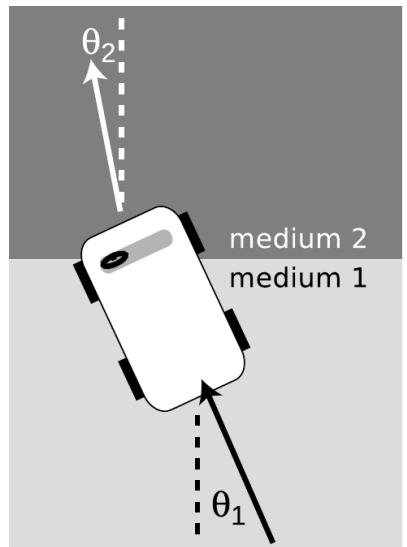
Finding an angle using Snell's law

example 4

- ▷ A submarine shines its searchlight up toward the surface of the water. What is the angle α shown in the figure?

▷ The tricky part is that Snell's law refers to the angles with respect to the normal. Forgetting this is a very common mistake. The beam is at an angle of 30° with respect to the normal in the water. Let's refer to the air as medium 1 and the water as 2. Solving Snell's law for θ_1 , we find

$$\theta_1 = \sin^{-1} \left(\frac{n_2}{n_1} \sin \theta_2 \right).$$



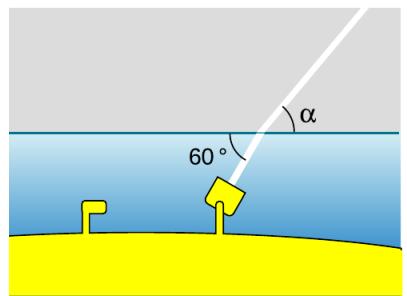
m / A mechanical model of refraction.

As mentioned above, air has an index of refraction very close to 1, and water's is about 1.3, so we find $\theta_1 = 40^\circ$. The angle α is therefore 50° .

For nonmagnetic substances, the impedance is the same as the speed (or its inverse, depending on what you call A_1 and what you call A_2). Therefore the ratio of the indices of refraction can be used to find the impedance ratio α . The relations like $R = (\alpha - 1)/(\alpha + 1)$ derived in section 5.3.2 apply only for normal incidence, but they do correctly predict most trends and qualitative facts when the incidence is not normal. The complete expressions for R and T for an electromagnetic wave, with non-normal incidence, are called Fresnel's equations, and we will not study them here. They depend on the polarization of the wave, which is why reflection of a light wave usually produces partial reflection. Such relations for electromagnetic waves apply to insulators — for a perfect conductor, there is always 100% reflection, since electric fields cannot exist in its interior.

Discussion questions

- A** What index of refraction should a fish have in order to be invisible to other fish?
- B** The earth's atmosphere gets thinner and thinner as you go higher in altitude. If a ray of light comes from a star that is below the zenith, what will happen to it as it comes into the earth's atmosphere?
- C** A denser sample of a gas has a higher index of refraction than a less dense sample (i.e., a sample under lower pressure), but why would it not make sense for the index of refraction of a gas to be proportional to density? Hint: What the graph of a proportionality look like?

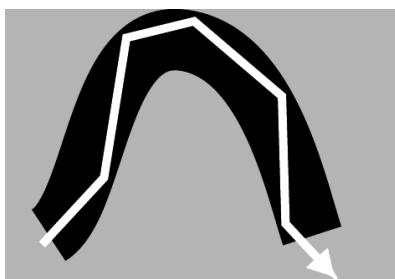


n / Example 4.

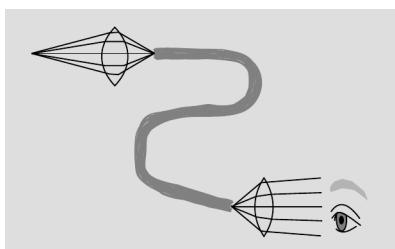
5.6.3 Total internal reflection

For normal incidence, we can get 100% reflection if the ratio of the impedances of our two media is 0 or ∞ , but this is only an idealization. But for non-normal incidence, it is possible to have situations where no possible angle for the refracted ray can satisfy Snell's law. Solving Snell's law for θ_2 , we find

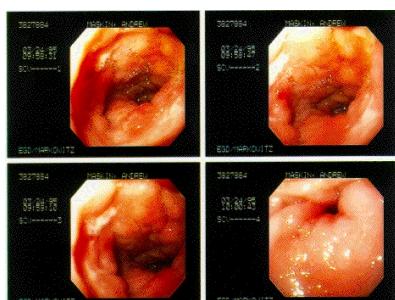
$$\theta_2 = \sin^{-1} \left(\frac{n_1}{n_2} \sin \theta_1 \right),$$



o / Total internal reflection in a fiber-optic cable.



p / A simplified drawing of a surgical endoscope. The first lens forms a real image at one end of a bundle of optical fibers. The light is transmitted through the bundle, and is finally magnified by the eyepiece.



q / Endoscopic images of a duodenal ulcer.

and if n_1 is greater than n_2 , then there will be large values of θ_1 for which the quantity $(n_1/n_2) \sin \theta$ is greater than one, meaning that your calculator will flash an error message at you when you try to take the inverse sine. What can happen physically in such a situation? The answer is that all the light is reflected, so there is no refracted ray. This phenomenon is known as *total internal reflection*, and is used in the fiber-optic cables that nowadays carry almost all long-distance telephone calls. The electrical signals from your phone travel to a switching center, where they are converted from electricity into light. From there, the light is sent across the country in a thin transparent fiber. The light is aimed straight into the end of the fiber, and as long as the fiber never goes through any turns that are too sharp, the light will always encounter the edge of the fiber at an angle sufficiently oblique to give total internal reflection. If the fiber-optic cable is thick enough, one can see an image at one end of whatever the other end is pointed at.

Alternatively, a bundle of cables can be used, since a single thick cable is too hard to bend. This technique for seeing around corners is useful for making surgery less traumatic. Instead of cutting a person wide open, a surgeon can make a small "keyhole" incision and insert a bundle of fiber-optic cable (known as an endoscope) into the body.

Since rays at sufficiently large angles with respect to the normal may be completely reflected, it is not surprising that the relative amount of reflection changes depending on the angle of incidence, and is greatest for large angles of incidence.

Discussion questions

A Does a surgeon using an endoscope need a source of light inside the body cavity? If so, how could this be done without inserting a light bulb through the incision?

B Does total internal reflection occur when light in a denser medium encounters a less dense medium, or the other way around? Or can it occur in either case?

5.7 Review of resonance and complex numbers

Light isn't a vibration of a physical medium at all. Why, then, does it have a different speed in different media? Explaining this in a satisfying way turns out to be an excellent application of concepts about resonance and complex numbers.

5.7.1 Physical motivation for use of complex numbers: feedback systems

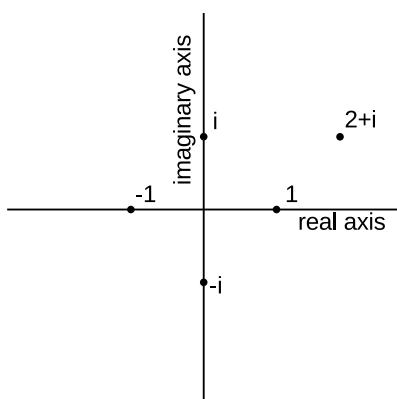
For most of us the word feedback evokes an image of Jimi Hendrix intentionally creating earsplitting screeches, or of the school principal doing the same inadvertently in the auditorium. In the guitar example, the musician stands in front of the amp and turns it up so high that the sound waves coming from the speaker come back to the guitar string and make it shake harder. This is an example of *positive* feedback: the harder the string vibrates, the stronger the sound waves, and the stronger the sound waves, the harder the string vibrates. The only limit is the power-handling ability of the amplifier.

Negative feedback is equally important. Your thermostat, for example, provides negative feedback by kicking the heater off when the house gets warm enough, and by firing it up again when it gets too cold. This causes the house's temperature to oscillate back and forth within a certain range. Just as out-of-control exponential freak-outs are a characteristic behavior of positive-feedback systems, oscillation is typical in cases of negative feedback.

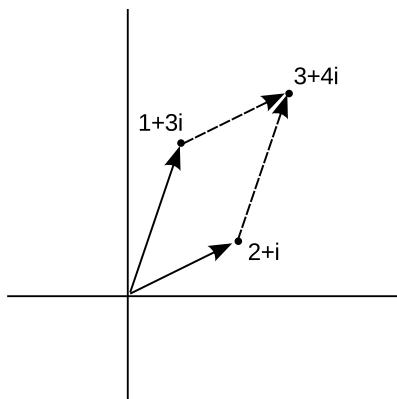
In mechanics, a pendulum is the classic example of negative feedback, while a good example of positive feedback is the unstable equilibrium we get when we try to balance a pencil on its tip. In electromagnetism, a series LC circuit oscillates because of feedback that occurs because charge flowing through the inductor accumulates on the plates of the capacitor, while the inductor acts back on the circuit through induced electric fields.

5.7.2 Complex numbers

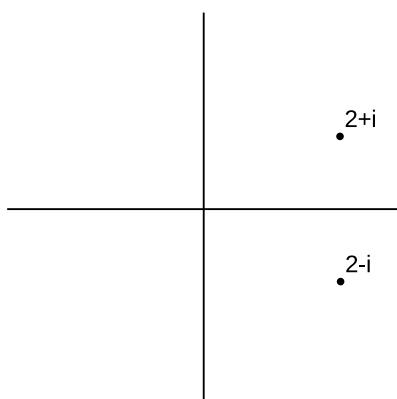
The complex number system gives us an efficient way of describing both positive and negative feedback. Complex numbers are a vital part of the day-to-day toolbox of the working engineer. We will also use them extensively in our study of quantum physics later in this course. This discussion assumes that you've had some previous exposure to the complex number system, but you may find that the techniques and applications laid out here are very different than the ones in a typical presentation in a trigonometry or precalculus class. For a more detailed treatment of complex numbers, see ch. 3 of James Nearing's free book at physics.miami.edu/~nearing/mathmethods.



r / Visualizing complex numbers as points in a plane.



s / Addition of complex numbers is just like addition of vectors, although the real and imaginary axes don't actually represent directions in space.



t / A complex number and its conjugate.

We assume there is a number, i , such that $i^2 = -1$. The square roots of -1 are then i and $-i$. (In electrical engineering work, where i stands for current, j is sometimes used instead.) This gives rise to a number system, called the complex numbers, which contain the real numbers as a subset.

If we calculate successive powers of i , we get the following:

$$\begin{aligned} i^0 &= 1 && [\text{true for any base}] \\ i^1 &= i \\ i^2 &= -1 && [\text{definition of } i] \\ i^3 &= -i \\ i^4 &= 1. \end{aligned}$$

By repeatedly multiplying i by itself, we have wrapped around, returning to 1 after four iterations. If we keep going like this, we'll keep cycling around. This is how the complex number system models oscillations, which result from negative feedback. For a mechanical system such as a mass on a spring, the farther the mass gets to the right of equilibrium

To model exponential behavior in the complex number system, we also use repeated multiplication. For example, if interest payments on your credit card debt cause it to double every decade (a positive feedback cycle), then your debt goes like $2^0 = 1$, $2^1 = 2$, $2^2 = 4$, and so on.

Any complex number z can be written in the form $z = a + bi$, where a and b are real, and a and b are then referred to as the real and imaginary parts of z . A number with a zero real part is called an imaginary number. The complex numbers can be visualized as a plane, with the real number line placed horizontally like the x axis of the familiar $x - y$ plane, and the imaginary numbers running along the y axis. The complex numbers are complete in a way that the real numbers aren't: every nonzero complex number has two square roots. For example, 1 is a real number, so it is also a member of the complex numbers, and its square roots are -1 and 1 . Likewise, -1 has square roots i and $-i$, and the number i has square roots $1/\sqrt{2} + i/\sqrt{2}$ and $-1/\sqrt{2} - i/\sqrt{2}$.

Complex numbers can be added and subtracted by adding or subtracting their real and imaginary parts. Geometrically, this is the same as vector addition.

The complex numbers $a + bi$ and $a - bi$, lying at equal distances above and below the real axis, are called complex conjugates. The results of the quadratic formula are either both real, or complex conjugates of each other. The complex conjugate of a number z is denoted as \bar{z} or z^* .

The complex numbers obey all the same rules of arithmetic as the reals, except that they can't be ordered along a single line. That

is, it's not possible to say whether one complex number is greater than another. We can compare them in terms of their magnitudes (their distances from the origin), but two distinct complex numbers may have the same magnitude, so, for example, we can't say whether 1 is greater than i or i is greater than 1.

A square root of i

example 5

▷ Prove that $1/\sqrt{2} + i/\sqrt{2}$ is a square root of i .

▷ Our proof can use any ordinary rules of arithmetic, except for ordering.

$$\begin{aligned} \left(\frac{1}{\sqrt{2}} + \frac{i}{\sqrt{2}}\right)^2 &= \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cdot \frac{i}{\sqrt{2}} + \frac{i}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + \frac{i}{\sqrt{2}} \cdot \frac{i}{\sqrt{2}} \\ &= \frac{1}{2}(1 + i + i - 1) \\ &= i \end{aligned}$$

Example 5 showed one method of multiplying complex numbers. However, there is another nice interpretation of complex multiplication. We define the argument of a complex number as its angle in the complex plane, measured counterclockwise from the positive real axis. Multiplying two complex numbers then corresponds to multiplying their magnitudes, and adding their arguments.

self-check D

Using this interpretation of multiplication, how could you find the square roots of a complex number?

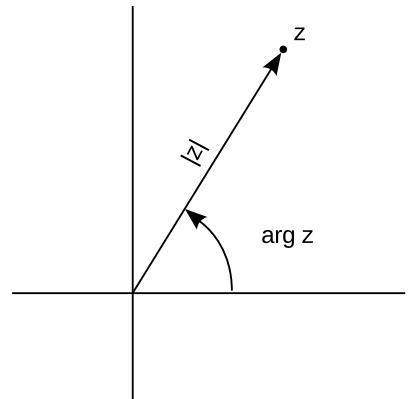
▷ Answer, p. 455

An identity

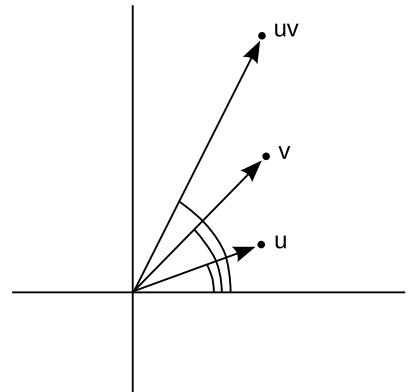
example 6

The magnitude $|z|$ of a complex number z obeys the identity $|z|^2 = z\bar{z}$. To prove this, we first note that \bar{z} has the same magnitude as z , since flipping it to the other side of the real axis doesn't change its distance from the origin. Multiplying z by \bar{z} gives a result whose magnitude is found by multiplying their magnitudes, so the magnitude of $z\bar{z}$ must therefore equal $|z|^2$. Now we just have to prove that $z\bar{z}$ is a positive real number. But if, for example, z lies counterclockwise from the real axis, then \bar{z} lies clockwise from it. If z has a positive argument, then \bar{z} has a negative one, or vice-versa. The sum of their arguments is therefore zero, so the result has an argument of zero, and is on the positive real axis.

1



u / A complex number can be described in terms of its magnitude and argument.



v / The argument of uv is the sum of the arguments of u and v .

¹I cheated a little. If z 's argument is 30 degrees, then we could say \bar{z} 's was -30, but we could also call it 330. That's OK, because $330+30$ gives 360, and an argument of 360 is the same as an argument of zero.

This whole system was built up in order to make every number have square roots. What about cube roots, fourth roots, and so on? Does it get even more weird when you want to do those as well? No. The complex number system we've already discussed is sufficient to handle all of them. The nicest way of thinking about it is in terms of roots of polynomials. In the real number system, the polynomial $x^2 - 1$ has two roots, i.e., two values of x (plus and minus one) that we can plug in to the polynomial and get zero. Because it has these two real roots, we can rewrite the polynomial as $(x - 1)(x + 1)$. However, the polynomial $x^2 + 1$ has no real roots. It's ugly that in the real number system, some second-order polynomials have two roots, and can be factored, while others can't. In the complex number system, they all can. For instance, $x^2 + 1$ has roots i and $-i$, and can be factored as $(x - i)(x + i)$. In general, the fundamental theorem of algebra states that in the complex number system, any n th-order polynomial can be factored completely into n linear factors, and we can also say that it has n complex roots, with the understanding that some of the roots may be the same. For instance, the fourth-order polynomial $x^4 + x^2$ can be factored as $(x - i)(x + i)(x - 0)(x - 0)$, and we say that it has four roots, i , $-i$, 0, and 0, two of which happen to be the same. This is a sensible way to think about it, because in real life, numbers are always approximations anyway, and if we make tiny, random changes to the coefficients of this polynomial, it will have four distinct roots, of which two just happen to be very close to zero.

Discussion questions

- A** Find $\arg i$, $\arg(-i)$, and $\arg 37$, where $\arg z$ denotes the argument of the complex number z .
- B** Visualize the following multiplications in the complex plane using the interpretation of multiplication in terms of multiplying magnitudes and adding arguments: $(i)(i) = -1$, $(i)(-i) = 1$, $(-i)(-i) = -1$.
- C** If we visualize z as a point in the complex plane, how should we visualize $-z$? What does this mean in terms of arguments? Give similar interpretations for z^2 and \sqrt{z} .
- D** Find four different complex numbers z such that $z^4 = 1$.
- E** Compute the following. For the final two, use the magnitude and argument, not the real and imaginary parts.

$$|1+i| , \quad \arg(1+i) , \quad \left| \frac{1}{1+i} \right| , \quad \arg \left(\frac{1}{1+i} \right) ,$$

From these, find the real and imaginary parts of $1/(1+i)$.

5.7.3 Euler's formula

Having expanded our horizons to include the complex numbers, it's natural to want to extend functions we knew and loved from the world of real numbers so that they can also operate on complex numbers. The only really natural way to do this in general is to use Taylor series. A particularly beautiful thing happens with the functions e^x , $\sin x$, and $\cos x$:

$$\begin{aligned} e^x &= 1 + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots \\ \cos x &= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots \\ \sin x &= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots \end{aligned}$$

If $x = i\phi$ is an imaginary number, we have

$$e^{i\phi} = \cos \phi + i \sin \phi,$$

a result known as Euler's formula. The geometrical interpretation in the complex plane is shown in figure w.

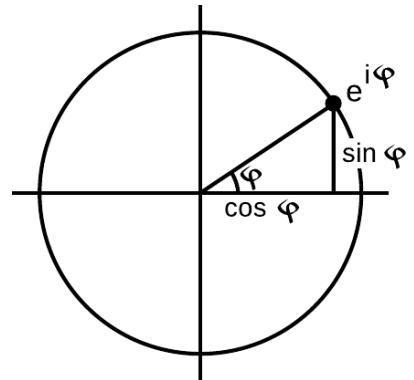
Although the result may seem like something out of a freak show at first, applying the definition of the exponential function makes it clear how natural it is:

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n.$$

When $x = i\phi$ is imaginary, the quantity $(1 + i\phi/n)$ represents a number lying just above 1 in the complex plane. For large n , $(1 + i\phi/n)$ becomes very close to the unit circle, and its argument is the small angle ϕ/n . Raising this number to the n th power multiplies its argument by n , giving a number with an argument of ϕ .

Euler's formula is used frequently in physics and engineering.

-
- Trig functions in terms of complex exponentials* example 7
- ▷ Write the sine and cosine functions in terms of exponentials.
 - ▷ Euler's formula for $x = -i\phi$ gives $\cos \phi - i \sin \phi$, since $\cos(-\theta) = \cos \theta$, and $\sin(-\theta) = -\sin \theta$.



w / The complex number $e^{i\phi}$ lies on the unit circle.



x / Leonhard Euler (1707-1783)

$$\begin{aligned} \cos x &= \frac{e^{ix} + e^{-ix}}{2} \\ \sin x &= \frac{e^{ix} - e^{-ix}}{2i} \end{aligned}$$

▷ Evaluate

$$\int e^x \cos x \, dx$$

▷ This seemingly impossible integral becomes easy if we rewrite the cosine in terms of exponentials:

$$\begin{aligned} & \int e^x \cos x \, dx \\ &= \int e^x \left(\frac{e^{ix} + e^{-ix}}{2} \right) \, dx \\ &= \frac{1}{2} \int (e^{(1+i)x} + e^{(1-i)x}) \, dx \\ &= \frac{1}{2} \left(\frac{e^{(1+i)x}}{1+i} + \frac{e^{(1-i)x}}{1-i} \right) + C \end{aligned}$$

Since this result is the integral of a real-valued function, we'd like it to be real, and in fact it is, since the first and second terms are complex conjugates of one another. If we wanted to, we could use Euler's theorem to convert it back to a manifestly real result.²

5.7.4 Simple harmonic motion and the LC circuit

The simple harmonic oscillator and the LC circuit should already be familiar to you. Here we show how complex numbers apply to these topics.

Figure y/1 shows a mass vibrating on a spring. If there is no friction, then the mass vibrates forever, and energy is transferred repeatedly back and forth between kinetic energy in the mass and potential energy in the spring. If we add friction, then the oscillations will dissipate these forms of energy into heat over time.

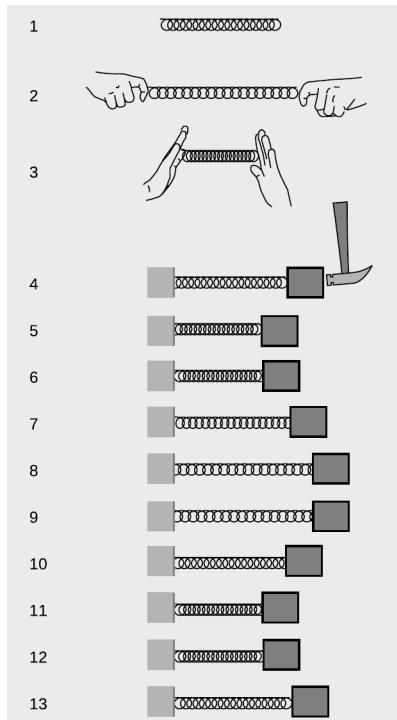
Figure y/2 shows the electrical analog. Energy cycles back and forth between electrical energy in the field of the capacitor and magnetic energy in the field of the coil. The electrical analog of friction is the resistance, which dissipates the energy of the oscillations into heat.

The dissipation of energy into heat is referred to as damping. This section covers simple harmonic motion, which is the case without damping. We consider the more general damped case in sec. 5.7.5.

The system of analogous variables for these two systems is as follows:

mechanical	electrical
$x = \text{position}$	$q = \text{charge on one plate}$
$v = x' = \text{velocity}$	$I = q' = \text{current}$
$a = x'' = \text{acceleration}$	$I' = \text{rate of change of current}$

²In general, the use of complex number techniques to do an integral could result in a complex number, but that complex number would be a constant, which could be subsumed within the usual constant of integration.



y / 1. A mass vibrating on a spring. 2. A circuit displaying analogous electrical oscillations.

Since this system of analogies is perfect, we'll discuss the behavior of the more concrete mechanical system. The mass is acted on by a force $-kx$ from the spring. Newton's second law can be written as

$$mx'' + kx = 0.$$

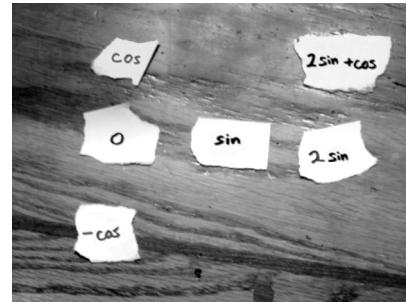
An equation like this, which relates a function to its own derivatives, is called a *differential equation*. This one is a *linear* differential equation, meaning that if $x_1(t)$ and $x_2(t)$ are both solutions, then so is any linear combination of them, $c_1x_1(t) + c_2x_2(t)$. It's not hard to guess what the solutions are: sines and cosines work as solutions, because the sine and cosine are functions whose second derivative is the same as the original function, except for a sign flip. The most general solution is of the form

$$c_1 \sin \omega t + c_2 \cos \omega t,$$

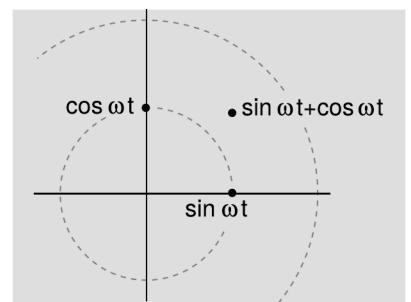
where the frequency is $\omega = \sqrt{k/m}$. It makes sense that there are two adjustable constants, because if we're given some the initial position and velocity of the mass, those are two numbers that we want to produce, and typically two equations in two unknowns will have a solution. Mathematically, this happens because the highest derivative in the differential equation is a second derivative.

But we would like to have some more specific and convenient way of organizing our thoughts about the physical interpretation of the constants c_1 and c_2 . Suppose we write down some examples of solutions on scraps of papers and then put them on a table and shuffle them around to try to see them in an organized way. As a loose analogy, this was how Mendeleev came up with the periodic table of the elements. Figure z shows what we might come up with for our "periodic table of the sine waves." What we've created is a system in which the solution $c_1 \sin \omega t + c_2 \cos \omega t$ is represented as a square on a checkerboard or, more generally, a point in the plane. Beautiful things happen if we think of this plane as being the complex plane, as laid out in the following table of exact mathematical analogies.

<i>sine waves</i>	<i>complex plane</i>
amplitude	magnitude
phase	argument
addition	addition
differentiation	multiplication by $i\omega$



z / Organizing some solutions to the equations of motion for simple harmonic motion.



aa / Example 10.

Sine compared to cosine

example 9

The sine function is the same as a cosine that has been delayed in phase by a quarter of a cycle, or 90 degrees. The two functions correspond to the complex numbers 1 and i , which have the same magnitude but differ by 90 degrees in their arguments.

Adding two sine waves

example 10

The trigonometric fact $\sin \omega t + \cos \omega t = \sqrt{2} \sin(\omega t + \pi/4)$ is visualized in figure aa.

A function's first and second derivative

example 11

Differentiating $\sin 3x$ gives $3 \cos 3x$. In terms of the complex plane, the function $\sin 3x$ is represented by 1. Differentiating it corresponds to multiplying this complex number by $3i$, which gives $3i$, and $3i$ represents the function $3 \cos 3x$ in our system.

Differentiating a second time gives $(\sin 3x)'' = -9 \sin 3x$. In terms of complex numbers, this is $1(3i)(3i) = -9$.

self-check E

Which of the following functions can be represented in this way? $\cos(6t - 4)$, $\cos^2 t$, $\tan t$

▷ Answer, p. 455

If we apply this system of analogies to the equation of motion $mx'' + kx = 0$, for a solution with amplitude A , we get $(-m\omega^2 + k)A = 0$, and if A is nonzero, this means that

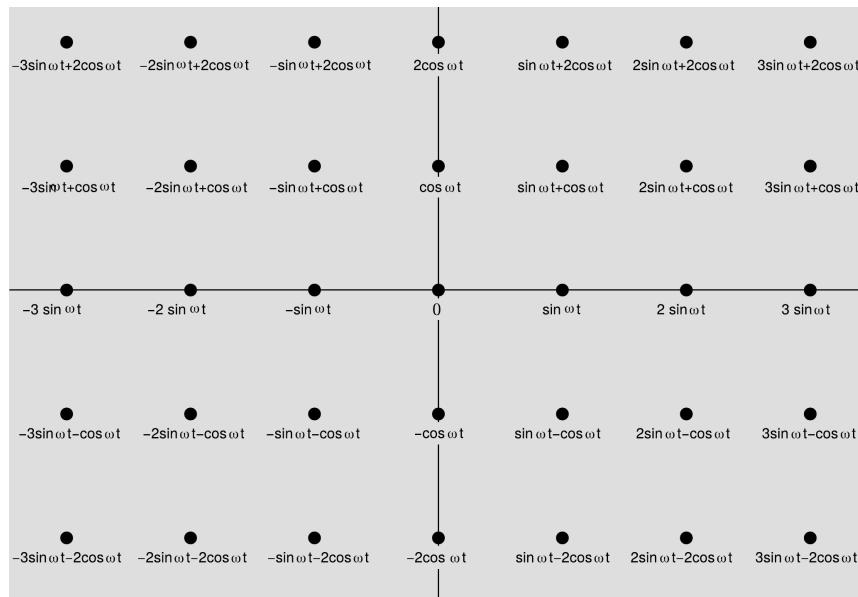
$$-m\omega^2 + k = 0.$$

This is a big win, because now instead of solving a differential equation, we just have to analyze an equation using algebra. If A is nonzero, then the factor in parentheses has to be zero, and that gives $\omega = \sqrt{k/m}$. (We could use the negative square root, but that doesn't actually give different solutions.)

Discussion question

A Interpret the following math facts visually using the figure below.

$$\begin{aligned}
 (\sin t)' &= \cos t \\
 (\cos t)' &= -\sin t \\
 (-\sin t)' &= -\cos t \\
 (-\cos t)' &= \sin t \\
 (2\sin t)' &= 2\cos t \\
 0' &= 0 \\
 (\cos 2t)' &= -2\sin 2t \\
 (\cos 3t)' &= -3\sin 3t
 \end{aligned}$$



5.7.5 Damped oscillations

We now extend the discussion to include damping. If you haven't learned about damped oscillations before, you may want to look first at a treatment that doesn't use complex numbers, such as the one in ch. 15 of OpenStax University Physics, volume 1, which is free online.

In the mechanical case, we will assume for mathematical convenience that the frictional force is proportional to velocity. Although this is not realistic for the friction of a solid rubbing against a solid, it is a reasonable approximation for some forms of friction, and anyhow it has the advantage of making the mechanical and electrical systems in figure y exactly analogous mathematically.

With this assumption, we add in to Newton's second law a frictional force $-bv$, where b is a constant. The equation of motion is

now

$$mx'' + bx' + kx = 0.$$

Applying the trick with the complex-number analogy, this becomes $-m\omega^2 + i\omega b + k = 0$, which says that ω is a root of a polynomial. Since we're used to dealing with polynomials that have real coefficients, it's helpful to switch to the variable $s = i\omega$, which means that we're looking for solutions of the form Ae^{st} . In terms of this variable,

$$ms^2 + bs + k = 0.$$

The most common case is one where b is fairly small, so that the quadratic formula produces two solutions for r that are complex conjugates of each other. As a simple example without units, let's say that these two roots are $s_1 = -1+i$ and $s_2 = -1-i$. Then if $A = 1$, our solution corresponding to s_1 is $x_1 = e^{(-1+i)t} = e^{-t}e^{it}$. The e^{it} factor spins in the complex plane, representing an oscillation, while the e^{-t} makes it die out exponentially due to friction. In reality, our solution should be a real number, and if we like, we can make this happen by adding up combinations, e.g., $x_1 + x_2 = 2e^{-t} \cos t$, but it's usually easier just to write down the x_1 solution and interpret it as a decaying oscillation. Figure ac shows an example.



ac / The amplitude is halved with each cycle.

self-check F

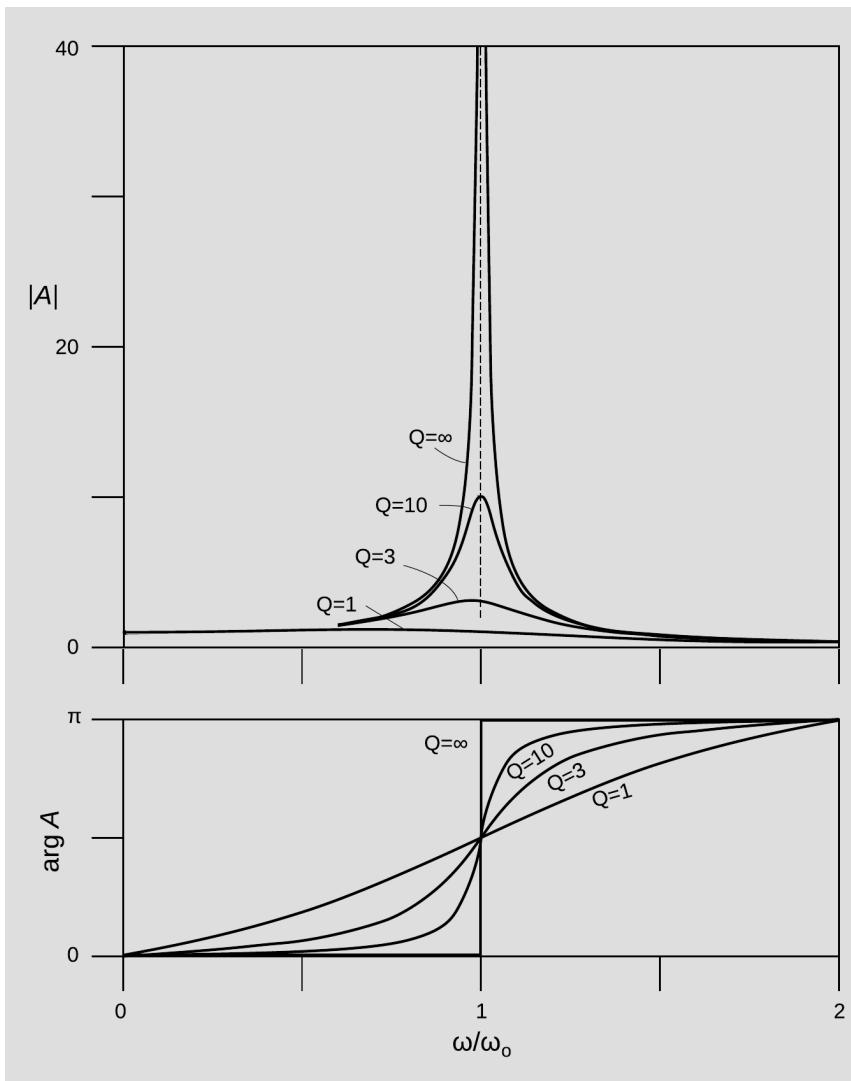
Figure ac shows an x-t graph for a strongly damped vibration, which loses half of its amplitude with every cycle. What fraction of the energy is lost in each cycle? ▷ Answer, p. 456

It is often convenient to describe the amount of damping in terms of the unitless *quality factor* $Q = \sqrt{km}/b$, which can be interpreted as the number of oscillations required for the energy to fall off by a factor of $e^{2\pi} \approx 535$.

5.7.6 Resonance

When a sinusoidally oscillating external driving force is applied to our system, it will respond by settling into a pattern of vibration in which it oscillates at the driving frequency. A mother pushing her kid on a playground swing is a mechanical example (not quite a rigorous one, since her force as a function of time is not a sine wave). An electrical example is a radio receiver driven by a signal picked up from the antenna. In both of these examples, it matters whether we pick the right driving force. In the example of the playground swing, Mom needs to push in rhythm with the swing's pendulum frequency. In the radio receiver, we tune in a specific frequency and reject others. These are examples of resonance: the system responds most strongly to driving at its natural frequency of oscillation. If you haven't had a previous introduction to resonance in the mechanical context, this review will not be adequate, and you will first want to look at another book, such as OpenStax University Physics.

ad / Dependence of the amplitude and phase angle on the driving frequency. The undamped case is $Q = \infty$, and the other curves represent $Q=1, 3$, and 10 . \tilde{F} , m , and $\omega_0 = \sqrt{k/m}$ are all set to 1.



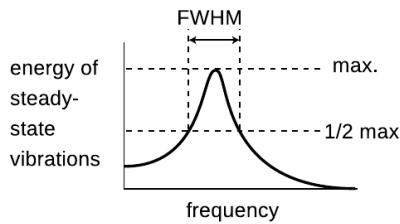
With the addition of a driving force F , the equation of motion for the damped oscillator becomes

$$mx'' + bx' + kx = F,$$

where F is a function of time. In terms of complex amplitudes, this is $(-\omega^2m + i\omega b + k)A = \tilde{F}$. Here we introduce the notation \tilde{F} , which looks like a little sine wave above the F , to mean the complex number representing F 's amplitude. The result for the steady-state response of the oscillator is

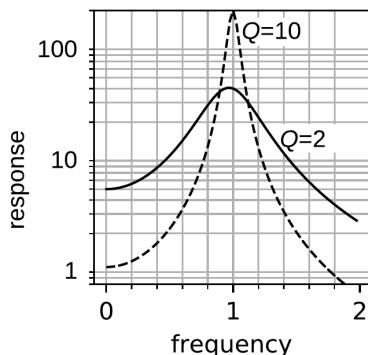
$$A = \frac{\tilde{F}}{-\omega^2m + i\omega b + k}.$$

To see that this makes sense, consider the case where $b = 0$. Then by setting ω equal to the natural frequency $\sqrt{k/m}$ we can make A blow up to infinity. This is exactly what would happen if Mom pushed Baby on the swing and there was no friction to keep the oscillations from building up indefinitely.



af / Definition of the FWHM of the resonance peak.

Figure ad shows how the response depends on the driving frequency. The peak in the graph of $|A|$ demonstrates that there is a resonance. Increasing Q , i.e., decreasing damping, makes the response at resonance greater, which is intuitively reasonable. What is a little more surprising is that it also changes the *shape* of the resonance peak, making it narrower and spikier, as shown in figure ae. The width of the resonance peak is often described using the full width at half-maximum, or FWHM, defined in figure af. The FWHM is approximately equal to $1/Q$ times the resonant frequency, the approximation being a good one when Q is large.



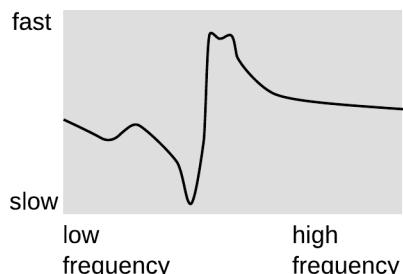
ae / Increasing Q increases the response and makes the peak narrower. In this graph, frequencies are in units of the natural frequency, and the response is the energy of the steady state, on an arbitrary scale. To make the comparison more visually clear, the curve for $Q = 2$ is multiplied by 5. Without this boost in scale, the $Q = 2$ curve would always lie below the one for $Q = 10$.

5.7.7 Dispersion

A surprising and cool application is the explanation of why electromagnetic waves traveling through matter are *dispersive* (section 3.3), i.e., their speed depends on their frequency. We take glass as an example of our medium. Figure ag shows a typical observation, in which clearly something special is happening at a certain frequency. This is a resonance of the charged particles in the glass, which vibrate in response to the electric field of the incoming wave.

To see how this works out, let's say that the incident wave has an electric field with a certain amplitude and phase. Ignoring units for convenience, let's arbitrarily take it to be $\sin \omega t$, so that in our complex-number setup, we represent it as

$$\text{original wave} = 1.$$



ag / The speed of light waves in silica glass (c/v running from 3 to 0) is graphed for increasing frequency and decreasing wavelength (λ from 15 μm to 1 μm).

This causes a charged particle in the glass to oscillate. Its position as a function of time is some other sinusoidal wave with some phase and amplitude, represented by

$$\text{displacement of particle} = A.$$

This A will be a complex number, with magnitude and phase behaving as in figure ad. The motion of these charges produces a current. Their velocity is the time derivative of their position, and we've seen that taking a time derivative can be represented in terms of complex numbers as multiplication by $i\omega$. For our present purposes it would be too much of a distraction to keep track of all the real-valued factors, such as ω , the number of charges, and so on. Omitting all of those, we have

$$\text{current} = iA.$$

Currents create magnetic fields, and this oscillating current will create an oscillating magnetic field, which will be part of a reemitted secondary wave, also traveling to the right,

$$\text{secondary wave} = -iA,$$

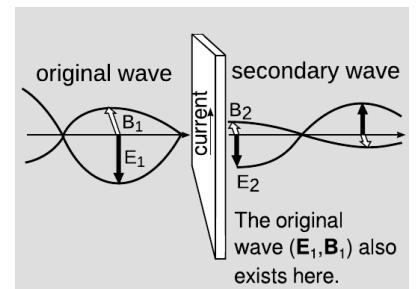
where the extra minus sign comes from Faraday's law. On the right side of the glass, we observe the superposition of the original wave and the secondary wave,

$$\text{transmitted wave} = 1 - iA.$$

Consulting figure ad, we see that for frequencies somewhat below the resonance, A is small and its phase approximately real-positive. Therefore $1 - iA$ is in the fourth quadrant, somewhat below the real axis. This represents a transmitted wave that is behind the original wave in terms of phase. The effect is as if the wave were arriving late, i.e., traveling at lower than normal speed.

Increasing the frequency, we expect that as we hit resonance, A will be large and positive-imaginary. Now the quantity $1 - iA$ becomes positive and real, the real phase indicating that the transmitted wave neither leads nor lags the original wave. This is the point in the middle of the graph where the velocity is back to normal.

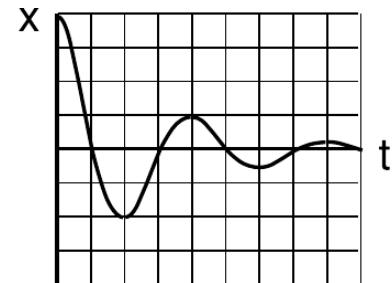
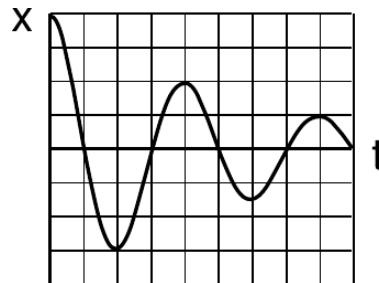
Farther to the right, at frequencies above resonance, A is near the negative real axis, $1 - iA$ is above the real axis, and the transmitted wave leads the original one. The velocity is faster than normal — in fact, it can be faster than c ! Unfortunately this does not give us a way of violating relativity. Our calculations of A were all calculations of the *steady state* response of the resonator. If we turn on our incident wave at some point in time, there will be a delay before the steady-state response is achieved, and this is more than enough to reduce the actual speed of propagation of the signal, called the *group velocity*, to less than c . The velocity that is greater than c is the *phase velocity*, and is not the speed at which energy and information propagate.



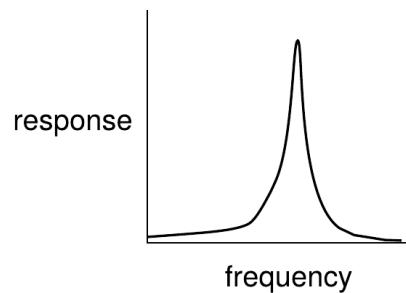
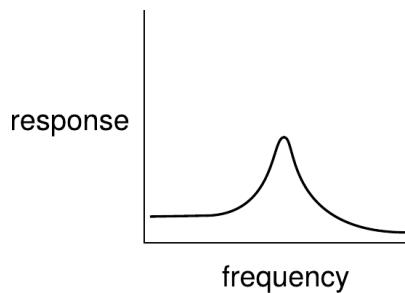
ah / A physical explanation in which the original light wave excites the charges in the glass, which reemit a secondary wave. The secondary wave is observed superposed with the original one. Redrawn from Kitamura, Pilon, and Jonasz, Applied Optics 46 (2007) 8118, reprinted online at <http://www.seas.ucla.edu/~pilon/Publications/AO2007-1.pdf>.

Discussion questions

A Compare the Q values of the two oscillators in the figures below.



B Match the x - t graphs in discussion question A with the amplitude-frequency graphs below.



Notes for chapter 5

213 Definitions relating to impedance

The table on p. 112 listed some examples of measures of amplitude that are familiar and easy to measure. If we want to get the most full and detailed possible use of the system of analogies for the different types of waves, then in some cases we need to use variables that differ from that list. Here we consider the relevant definitions for the three types of waves used as examples in the main text: the electromagnetic wave, the wave on a string, and the sound wave.

There is quite a bit of freedom in defining what variables we want to consider. Often, as in section 5.3, we're interested in the situation where a wave encounters a sharp boundary between two different media. In these cases, it's convenient to consider variables that remain continuous across the boundary.

In the case of an *electromagnetic wave*, this suggests not using the magnetic field \mathbf{B} but rather the auxiliary field $\mathbf{H} = \mathbf{B}/\mu$, where μ is the magnetic permeability. For a wave coming in along the normal, the fields are parallel to the boundary, and one can then show from Faraday's law that \mathbf{H} is what is continuous, not \mathbf{B} . For electromagnetic waves, we therefore define our two measures of amplitude as E and H . We then have the choice of defining the impedance as E/H or H/E . Defining $Z = E/H$ is nice because it gives Z units of ohms. In a vacuum, $Z = 377 \Omega$. This does *not* mean that a vacuum acts like a resistor with this resistance. The energy density is $(1/2)\epsilon E^2 + (1/2)\mu H^2$, where ϵ is the permittivity. The energy flux per unit area is the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$.

For a wave on a string, we take our two measures of amplitude to be the transverse velocity and transverse force v_y and F_y . Because Newton's second law relates force to acceleration, not velocity, there is in general no relationship between these two variables. But in the case of a sinusoidal wave, their ampli-

tudes \tilde{F} and \tilde{v} are related by $Z = \tilde{F}_y/\tilde{v}_y$, where $Z = \sqrt{\mu T}$, μ is the mass per unit length, and T is the tension. The variable v_y must be continuous at a boundary such as the one shown in figure a, p. 114, since otherwise the function $y(x)$ would develop a discontinuity, which would describe a break in the string. We also expect F_y to be continuous, because otherwise the infinitesimal mass element at the boundary would have an infinite acceleration in the y direction. The energy flux is $S = F_y v_y$; because the string is one-dimensional, this is really just a measure of power, with units of watts.

In the case of sound waves, we take our measures of amplitude to be the pressure p and volume flow rate u . The latter has units of cubic meters per second. These quantities can also be defined on a per-unit-area basis. The acoustic impedance is $Z = p/u$, and equals ρv , where ρ is the density of the medium and v is the speed of the waves. The energy flux per unit area for a plane wave is $S = Pu$.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

- 1** (a) Compute the amplitude of light that is reflected back into air at an air-water interface, relative to the amplitude of the incident wave. Assume that the light arrives in the direction directly perpendicular to the surface. The speeds of light in air and water are 3.0×10^8 and 2.2×10^8 m/s, respectively.
(b) Find the energy of the reflected wave as a fraction of the incident energy.

▷ Hint, p. 443 ✓

- 2** The expressions for the amplitudes of reflected and transmitted waves depend on the unitless ratio of the impedances α . As described in the text (“duality,” p. 115), the definition of this ratio is generally somewhat arbitrary, and we can replace α with $1/\alpha$. Interchanging the roles of the two media also changes α to $1/\alpha$. (a) Show that changing α to $1/\alpha$ (e.g., by interchanging the roles of the two media) has an effect on the reflected amplitude that can be expressed in a simple way, and discuss what this means in terms of inversion and energy. (b) Find the two values of α for which $|R| = 1/2$.

- 3** (a) A good tenor saxophone player can play all of the following notes without changing her fingering, simply by altering the tightness of her lips: E \flat (150 Hz), E \flat (300 Hz), B \flat (450 Hz), and E \flat (600 Hz). How is this possible? (I’m not asking you to analyze the coupling between the lips, the reed, the mouthpiece, and the air column, which is very complicated.)
(b) Some saxophone players are known for their ability to use this technique to play “freak notes,” i.e., notes above the normal range of the instrument. Why isn’t it possible to play notes below the normal range using this technique?

- 4** A concert flute produces its lowest note, at about 262 Hz, when half of a wavelength fits inside its tube. Compute the length of the flute.

▷ Answer, p. 459

5 Consider a one-dimensional standing-wave pattern that has asymmetric boundary conditions at the ends, i.e., for whatever measure of amplitude we have chosen, the amplitude is zero at one end but has an extremum at the other end. The purpose of this problem is to predict the frequencies of vibration, as we did in problem 12, p. 59, for the symmetric case.

(a) Sketch the first three patterns. Check yourself against figure figure e/3 for the first pattern and the answer to self-check C for the second one.

(b) As a warm-up with concrete numbers, consider the case where the length L is 1 m. Find the wavelengths of the patterns you drew in part a.

(c) Returning to the general case where L is a variable, find the pattern of wavelengths. You can do this either by writing a list ending in “...” that clearly shows the pattern, or by defining an N and writing an equation in terms of that N . There is more than one way to define an N , so if you do that, explain what your definition is and what are the permissible values of N .

(d) Let v be the speed of the waves. Make a prediction of the frequencies in a similar style. Check against the answer to self-check C.

(e) The clarinet acts as an asymmetric air column. Its lowest note is produced by closing all the tone holes. The note produced in this way has a fundamental frequency (lowest harmonic) of 147 Hz, but this tone also contains all the frequencies predicted in part d. Predict the frequency of the third harmonic, i.e., the third-longest wavelength. ✓

6 The table gives the frequencies of the notes that make up the key of F major, starting from middle C and going up through all seven notes.

(a) Calculate the first four or five harmonics of C and G, and determine whether these two notes will be consonant or dissonant. (Recall that harmonics that differ by about 1-10% cause dissonance.)

(b) Do the same for C and B \flat .

C	261.6 Hz
D	293.7
E	329.6
F	349.2
G	392.0
A	440.0
B \flat	466.2

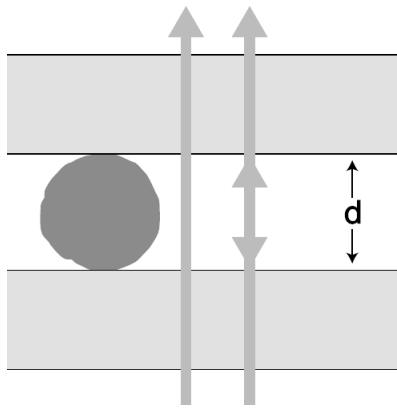
Problem 6.

7 A string hangs vertically, free at the bottom and attached at the top.

(a) Find the velocity of waves on the string as a function of the distance from the bottom. ✓

(b) Find the acceleration of waves on the string. \triangleright Answer, p. 459

(c) Interpret your answers to parts a and b for the case where a pulse comes down and reaches the end of the string. What happens next? Check your answer against experiment and conservation of energy. *



Problem 8.

8 A Fabry-Perot interferometer, shown in the figure being used to measure the diameter of a thin filament, consists of two glass plates with an air gap between them. As the top plate is moved up or down with a screw, the light passing through the plates goes through a cycle of constructive and destructive interference, which is mainly due to interference between rays that pass straight through and those that are reflected twice back into the air gap. (Although the dimensions in this drawing are distorted for legibility, the glass plates would really be much thicker than the length of the wave-trains of light, so no interference effects would be observed due to reflections within the glass.)

- (a) If the top plate is cranked down so that the thickness, d , of the air gap is much less than the wavelength λ of the light, i.e., in the limit $d \rightarrow 0$, what is the phase relationship between the two rays? (Recall that the phase can be inverted by a reflection.) Is the interference constructive, or destructive?
- (b) If d is now slowly increased, what is the first value of d for which the interference is the same as at $d \rightarrow 0$? Express your answer in terms of λ .
- (c) Suppose the apparatus is first set up as shown in the figure. The filament is then removed, and n cycles of brightening and dimming are counted while the top plate is brought down to $d = 0$. What is the thickness of the filament, in terms of n and λ ?

Based on a problem by D.J. Raymond.

9 Diamond has an index of refraction of 2.42, and part of the reason diamonds sparkle is that this encourages a light ray to undergo many total internal reflections before it emerges. (a) Calculate the critical angle at which total internal reflection occurs in diamond. (b) Explain the interpretation of your result: Is it measured from the normal, or from the surface? Is it a minimum angle for total internal reflection, or is it a maximum? How would the critical angle have been different for a substance such as glass or plastic, with a lower index of refraction? ✓

10 (a) A wave pulse moves into a new medium, where its velocity is greater by a factor α . Find an expression for the fraction, f , of the wave energy that is transmitted, in terms of α . Note that, as discussed in the text, you cannot simply find f by squaring the amplitude of the transmitted wave. \triangleright Answer, p. 459

(b) Suppose we wish to transmit a pulse from one medium to another, maximizing the fraction of the wave energy transmitted. To do so, we sandwich another layer in between them, so that the wave moves from the initial medium, where its velocity is v_1 , through the intermediate layer, where it is v_2 , and on into the final layer, where it becomes v_3 . What is the optimal value of v_2 ? (Assume that the middle layer is thicker than the length of the pulse, so there are no interference effects. Also, although there will be later echoes that are transmitted after multiple reflections back and forth across the middle layer, you are only to optimize the strength of the transmitted pulse that is first to emerge. In other words, it's simply a matter of applying your answer from part a twice to find the amount that finally gets through.) \triangleright Answer, p. 459 *

11 (a) Use complex number techniques to rewrite the function $f(t) = 4 \sin \omega t + 3 \cos \omega t$ in the form $A \sin(\omega t + \delta)$. \checkmark

(b) Verify the result using the trigonometric identity $\sin(\alpha + \beta) = \sin \alpha \cos \beta + \sin \beta \cos \alpha$.

12 Use Euler's theorem to derive the addition theorems that express $\sin(a + b)$ and $\cos(a + b)$ in terms of the sines and cosines of a and b . \triangleright Solution, p. 444

13 Find every complex number z such that $z^3 = 1$.

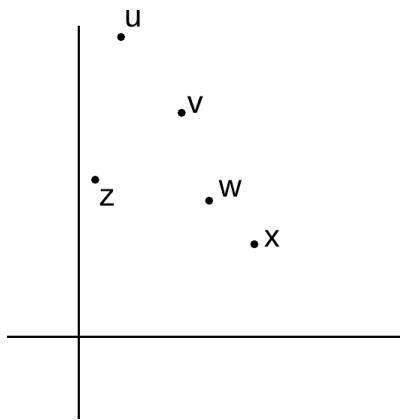
\triangleright Solution, p. 444

14 One solution of the differential equation

$$\frac{d^4 x}{dt^4} - 16x = 0$$

is the function $x(t) = \sin 2t$. Visualize this function as a point in the complex plane, and use the system of analogies on p. 131 to explain why this is a solution to the differential equation.

\triangleright Solution, p. 444



Problem 15.

15 This problem deals with the cubes and cube roots of complex numbers, but the principles involved apply more generally to other exponents besides 3 and 1/3. These examples are designed to be much easier to do using the magnitude-argument representation of complex numbers than with the cartesian representation. If done by the easiest technique, none of these requires more than two or three lines of *simple* math. In the following, the symbols θ , a , and b represent real numbers, and all angles are to be expressed in radians. As often happens with fractional exponents, the cube root of a complex number will typically have more than one possible value. (Cf. $4^{1/2}$, which can be 2 or -2 .) In parts c and d, this ambiguity is resolved explicitly in the instructions, in a way that is meant to make the calculation as easy as possible.

- (a) Calculate $\arg[(e^{i\theta})^3]$. ✓
- (b) Of the points u , v , w , and x shown in the figure, which could be a cube root of z ?
- (c) Calculate $\arg[\sqrt[3]{a+bi}]$. For simplicity, assume that $a+bi$ is in the first quadrant of the complex plane, and compute the answer for a root that also lies in the first quadrant. ✓
- (d) Compute

$$\frac{1+i}{(-2+2i)^{1/3}}.$$

Because there is more than one possible root to use in the denominator, multiple answers are possible in this problem. Use the root that results in the final answer that lies closest to the real line. (This is also the easiest one to find by using the magnitude-argument techniques introduced in the text.)

✓

16 Find the 100th derivative of $e^x \cos x$, evaluated at $x = 0$.
[Based on a problem by T. Needham.] ✓

17 Factor the expression $x^3 - y^3$ into factors of the lowest possible order, using complex coefficients. (Hint: use the result of problem 13.) Then do the same using real coefficients.

18 Calculate the quantity i^i (i.e., find its real and imaginary parts). ▷ Hint, p. 443 ✓ *

19 Many fish have an organ known as a swim bladder, an air-filled cavity whose main purpose is to control the fish's buoyancy and allow it to keep from rising or sinking without having to use its muscles. In some fish, however, the swim bladder (or a small extension of it) is linked to the ear and serves the additional purpose of amplifying sound waves. For a typical fish having such an anatomy, the bladder has a resonant frequency of 300 Hz, the bladder's Q is 3, and the maximum amplification is about a factor of 100 in energy. Over what range of frequencies would the amplification be at least a factor of 50?

✓

20 (a) Given that the argument of a complex number z equals θ , what is the argument of $1/z$?

For the remainder of this problem, let $z = \sqrt{3} + i$. Sketch the location of this point in the complex plane. Find (b) $|z|$, (c) $\arg z$ in degrees, (d) $|1/z|$, (e) $\arg(1/z)$ in degrees, and (f) the imaginary part of $1/z$. Draw $1/z$ on your sketch.

You should be able to do all of these in your head, just by staring at the sketch. Don't do them by manipulating complex numbers in $a + bi$ form, because that's actually harder, and the purpose of this exercise is to get you used to doing it using the magnitude and argument.

✓

21 Simplify $\arg(1/\bar{z})$.

22 In an experiment to measure the unknown index of refraction n of a liquid, you send a laser beam from air into a tank filled with the liquid. Let ϕ be the angle of the beam relative to the normal while in the air, and let θ be the angle in the liquid. You can set ϕ to any value you like by aiming the laser from an appropriate direction, and you measure θ as a result. We wish to plan such an experiment so as to minimize the error dn in the result of the experiment, for a fixed error $d\theta$ in the measurement of the angle in the liquid. We assume that there is no significant contribution to the error from uncertainty in the index of refraction of air (which is very close to 1) or from the angle ϕ . Find dn in terms of $d\theta$, and determine the optimal conditions.

▷ Solution, p. 445

Chapter 6

Relativistic energy and momentum

6.1 Mystery stuff

Over the years, physicists have run into various kinds of mystery stuff, and have always assigned names as soon as possible. They discovered x-rays, and didn't know what they were, so they happily labeled them x-rays — as in *x* for an unknown. They discovered three types of radiation coming from the spontaneous decay of atoms, and since they didn't know what they were, they labeled them A, B, and C — except that they wanted to sound more scientific, so they used the first three letters of the *Greek* alphabet, alpha, beta, and gamma. Today, we have “dark matter,” so called because — you guessed it — it doesn't emit light.

Let's make this into a game with rules. Someone waves a magic wand, and a stream of particles of mystery stuff comes out. As the person playing the game, you then have to figure out something about the stuff. One thing you can do is to let the particles ram into a target. This deposits their energy in the target, heating it up. By measuring the temperature increase, you can indirectly find the energy, and if you then divide by the number of particles you can infer the kinetic energy K of each particle. By similar measurements of the target's recoil from the hit, you can find the momentum p per particle.

According to Newton's version of the laws of physics, the game seems pretty easy to win. Given K and p , you can do a little algebra and find the mass m of the mystery particle. It equals $p^2/2K$. This mass is a built-in property of the particle, which is the kind of thing you wanted to find. Yay, you win!

But your opponent can make you lose by playing a trick on you. They can let the particles come out of the wand with zero velocity. Sadly, you measure $p = 0$ and $K = 0$, which gives you absolutely no information about the particles. You lose.

It may console you to know that when the game is played by Einstein's rules, you can always win. As we saw in sec. 3.6.1, p. 75, mass and energy are equivalent in relativity. Even if your opponent tries to play their dirty trick of releasing the particles at rest, you

can measure their energy E , and also observe that they have $p = 0$. Because $p = 0$, you can tell they're at rest, and therefore the only energy they have is the energy-equivalent of their mass (sometimes called the rest energy). From this you infer their mass, $m = E$. Yay, you win!

self-check A

This is in natural units. Insert the necessary factors of c to get m in terms of E using SI units. ▷ Answer, p. 456

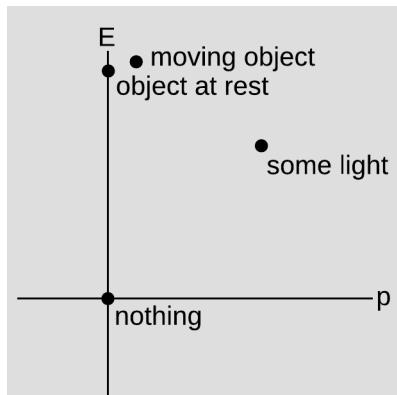
Another way of describing this is that in relativity, if we're given the pair of numbers (E, p) for an object in a certain frame of reference, we can always find the values of these quantities (E', p') in any other frame of reference. This failed in Newtonian physics because $(0, 0)$ didn't give enough information, but we don't have that problem in relativity. In relativity, a $(0, 0)$ is like Dr. Seuss's description of a Wasn't: "A Wasn't just isn't. He just isn't present."

6.2 The energy-momentum vector

The (E, p) pair has other pleasant properties. The two numbers, which have different units in Newtonian physics, have the *same* natural units. Furthermore, we saw in sec. 3.5, p. 73, that for a light wave, $p = E$ (which means in SI units that $p = E/c$, so that the momentum of light is too small to notice in everyday life, since c is big).

The situation ends up looking like the one sketched in figure a. A flash of light lies along one of the diagonals with slopes of ± 1 . A material object at rest lies straight up on the vertical axis.

This looks exactly like something we've seen before. On a spacetime diagram, the world-line a flash of light lies along one of the diagonals. On a spacetime diagram, the world-line of a material object at rest runs straight up and down. The pattern we're seeing here can be expressed in a gratifyingly simple way if we think of a (E, p) pair as the components of a vector lying in the p - E plane. We call this the energy-momentum vector \mathbf{p} . Then: —



a / The E - p plane.

The energy-momentum vector of a thing is always parallel to its world-line.

As a consequence of this, we have the handy fact that $v = p/E$, since the slope on either plot is the same.

self-check B

Figure a includes a dot representing a flash of light. If we had a flash of light twice as bright, where would its dot be? ▷ Answer, p. 456

self-check C

If you couldn't remember whether the rule was $v = p/E$ or $v = E/p$,

how could you tell which one would make sense? ▷ Answer, p. 456

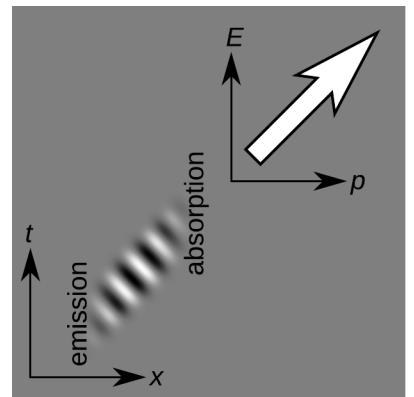
A spacetime diagram of a light wave

example 1

Figure b depicts a light wave on a spacetime diagram. Its world-line is the wavy strip shown in black and white. (Since it has some width to it, we could call it a world-ribbon rather than a world-line.) If a certain pixel on this diagram is white, then the electric field of the wave is very strong at that particular time and place. If the pixel is black, then the electric field is equally strong, but in the opposite direction. (These directions would be in the y - z plane, and therefore not visualizable on this t - x plane.) The world-line has a slope of 1. This makes sense for light, which travels at c .

An arrow representing the wave's energy-momentum vector is shown on the same plot. In component form, this vector is $\mathbf{p} = (E, p)$, and the sizes of the components are equal. This is correct for an electromagnetic wave, which has equal energy and momentum (in natural units).

The energy-momentum vector is parallel to the world-line.



b / A light wave and its energy-momentum vector.

6.3 Four-vectors in general

Up until now we've been confining ourselves almost solely to relativity in 1+1 dimensions, but with all three spatial dimensions present, we have a spacetime displacement vector with components that look like

$$\Delta \mathbf{r} = (\Delta t, \Delta x, \Delta y, \Delta z),$$

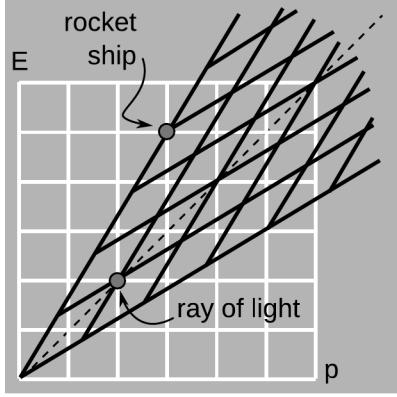
and a momentum vector

$$\mathbf{p} = (E, p_x, p_y, p_z).$$

These can be referred to as four-vectors in order to distinguish them from the familiar three-vectors of Newtonian mechanics. Just as all three-vectors have a consistent set of rules for addition, magnitudes, rotation, and so forth, all four-vectors behave by a consistent set of rules. Let's flesh out some of these analogies.

Lorentz transformation: As far as a mathematician is concerned (see sec. 6.6), we can always create a list of numbers and call it a “vector,” e.g., an expression like (G, d) , where G is the current price of gold and d is the number of dogs in Montana. Physicists are not so casual. For a three-vector, we want the components to behave as expected under a rotation (which (G, d) doesn't, since rotating a gold bar or a dog still leaves it as one bar or one dog). For a four-vector, we would like this behavior under rotations to apply to the three spatial components, and we would also like the components to behave as described by the Lorentz transformation when we do a boost. Experiments do show that the energy-momentum four-vector \mathbf{p} behaves this way. If it didn't, then we would probably violate the principle that \mathbf{p} is always parallel to $\Delta \mathbf{r}$ along the world-line of an

object — the two four-vectors might be parallel in one frame, but if we treated them inconsistently, then after a boost, they would probably not be parallel anymore.



c / Examples 2 and 3.

A Lorentz transformation in the E - p plane

example 2

The figure shows two points in the energy-momentum plane, representing a spaceship and a ray of light. Turning the book sideways, we see that the slope of the black energy axis is -0.6 . Therefore the black frame's velocity relative to the white frame is 0.6 .

In the white frame of reference, the spaceship has $E = 5$ units, which is a combination of its mass and its kinetic energy. Its momentum is $p = 3$ units. In the black frame of reference, which is moving at $v = 0.6$ relative to the white one, the ship has $p = 0$ — this is the ship's rest frame. This tells us that the ship was moving at $v = 0.6$ in the white frame. We can verify this because in the white frame, $v = p/E = 3/5$.

In the white frame, the ray of light has $E = 2$ and $p = 2$. In the black frame, it has been Doppler shifted down to $E = 1$ and $p = 1$.

Magnitudes: The magnitude of a spacetime displacement $\Delta\mathbf{r}$ is what we've been calling the interval \mathcal{I} . We can notate this as $|\Delta\mathbf{r}|^2 = \mathcal{I}^2$. In terms of the components, the rule that we've been notating in 1+1 dimensions as $\mathcal{I}^2 = t^2 - d^2$ can be written as $\mathcal{I}^2 = \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$. In other words, we have an expression like the Pythagorean theorem, except that we have signs that go like $+ - - -$. Reverting to 1+1 dimensions for compactness, $\mathcal{I}^2 = \Delta t^2 - \Delta x^2$. The analogous expression for the momentum vector's squared magnitude is $|\mathbf{p}|^2 = E^2 - p^2$.

6.4 Mass

If we consider an object at rest, as in figure a, then $p = 0$ and the squared magnitude of the energy-momentum vector, $|\mathbf{p}|^2 = E^2 - p^2$, equals the square of its mass. Since the magnitude of a four-vector is invariant, this is equally true in frames of reference where the object is not at rest, and we have the extremely important and useful relation

$$m^2 = E^2 - p^2.$$

We take this to be the *definition* of mass.

self-check D

The relation $m^2 = E^2 - p^2$ is in natural units. Insert the necessary factors of c to express this in SI units. ▷ Answer, p. 456

Because m is invariant, it doesn't change when we accelerate an object. That is, its mass is a permanent property of it. Since the energy-momentum vector is parallel to the object's world-line,

a material object, with $m^2 = |\mathbf{p}|^2 > 0$, must always move along a world-line with $|\Delta\mathbf{r}|^2 > 0$, i.e., it must have a speed less than c . Conversely, a ray of light has zero mass, so it always $|\Delta\mathbf{r}|^2 = 0$ and moves at c — like Lewis Carroll's Red Queen, it *has* to keep moving.

The mass of a spaceship *example 3*

Returning to the spaceship of figure c, we saw previously that the black frame, which is the ship's rest frame, it had $(E, p) = (4, 0)$, and its energy was equal to its rest mass, so $m = 4$.

In the white frame, the ship has $(E, p) = (5, 3)$. We then have $m^2 = 5^2 - 3^2 = 16$, so again $m = 4$. The result is the same in both frames, because the mass is invariant.

A ray of light has zero mass *example 4*

We've already seen that a ray of light or an electromagnetic plane wave has $E = p$, so that its energy-momentum vector in 1+1 dimensions looks like (E, E) . The magnitude of this is

$$E^2 - E^2 = 0,$$

which means that the mass is zero.¹

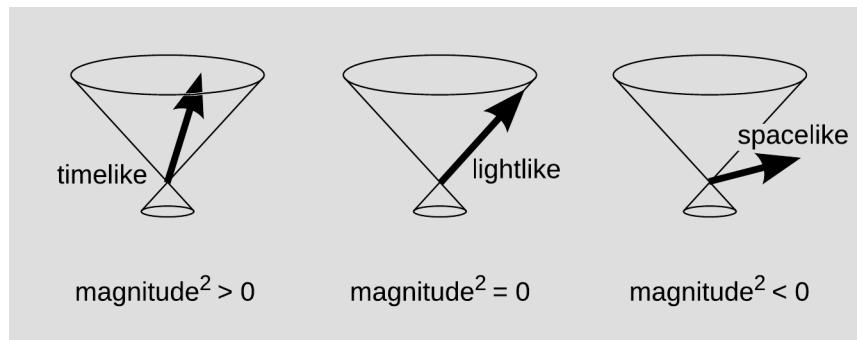
If you look in a math book at a discussion of the magnitude of a vector, whether at the level of high-school trigonometry or a college course in linear algebra, you will usually find a statement of the following property of the magnitude: if \mathbf{A} is a vector, and $|\mathbf{A}| = 0$, then \mathbf{A} is the zero vector, meaning that it has components that are all zero, and adding it to any other vector gives back the same vector, $\mathbf{B} + \mathbf{A} = \mathbf{B}$. In relativity, this is *not* true. In example 4, our ray of light has an energy-momentum vector \mathbf{p} with a magnitude of zero, $|\mathbf{p}| = 0$, but it is *not* true that $\mathbf{p} = 0$ — if the light ray hits you, it warms you up. This is just a difference in terminology. When scientists and engineers communicate with mathematicians, they have to keep in mind that they mean different things by words like “magnitude.”

6.5 Applications

The reason we care about the energy-momentum four-vector is that it's conserved. Figure e shows a collision between two pool balls, with a diagram showing how their energy-momentum vectors add up. As actually happens in this kind of head-on shot (unless you use a lot of spin on the cue-ball), the cue ball stops dead, transferring

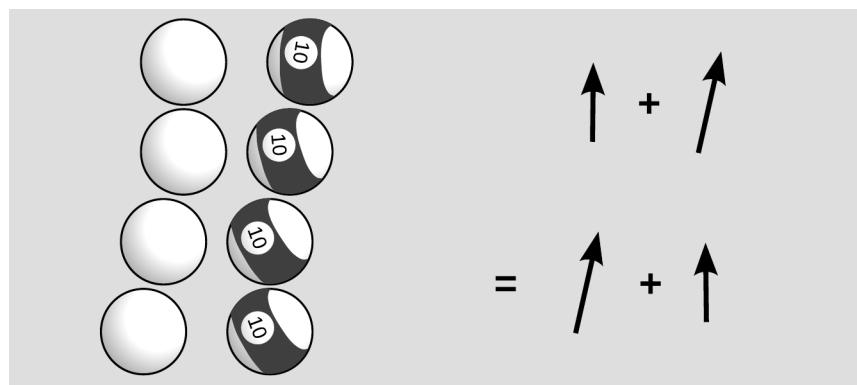
¹A gotcha is that mass is not additive in special relativity. Therefore the result of this example need not be true for other electromagnetic wave patterns. For example, a standing wave could consist of a superposition of two electromagnetic plane waves traveling in opposite directions. The two energy-momentum vectors are then (E, E) and $(E, -E)$, which adds up to $(2E, 0)$, and therefore has a nonzero mass of $2E$.

d / The magnitude of a four-vector can be positive, zero, or negative. A zero magnitude doesn't necessarily imply that the vector is zero.



all of its motion to the ball it hits. (The diagram is actually a little unrealistic, because I've drawn the balls moving at about 20% of c , which would not only destroy them on impact but also kill everyone within a considerable radius.)

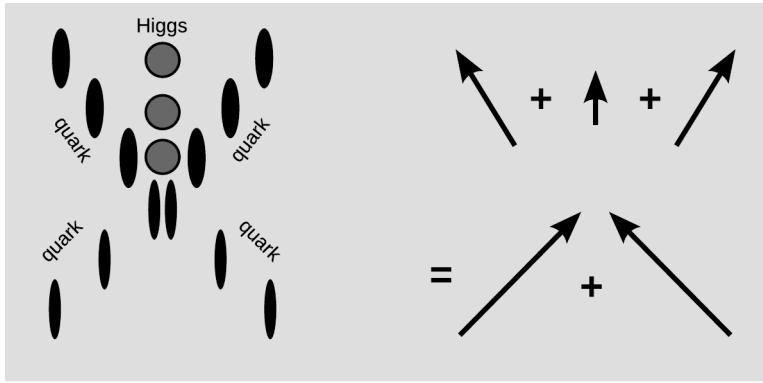
e / The energy-momentum vector is conserved in the collision of two pool balls.



One way to check that energy-momentum is conserved here is to set up each pair of vectors tip-to-tail. An easier way to verify this is that the two sides of the equation are the same except that the order of the things being added up has been reversed; this has no effect on the result because addition of four-vectors is commutative, just like addition of three-vectors.

Most of what we see in this diagram is simply the mass of the balls. The moving ball does however have a little bit of kinetic energy (extra height) and some momentum (leaning to the right, so that its spacelike part isn't zero).

Figure f is a diagram of a somewhat more realistic example, one that really requires relativity. This is a head-on collision of two subatomic particles called quarks, such as occurs at the Large Hadron Collider near Geneva, Switzerland. The quarks are actually parts of larger particles — protons — but on this scale we only see one quark from one proton colliding with one quark from another proton. We don't actually know if quarks are pointlike or have a



f / Production of a Higgs particle in an accelerator experiment, showing the world-lines (left) and the energy-momentum vectors (right). Energy is conserved: the sum of the vectors after the reaction equals the sum before.

finite size, but I've drawn them here as having some bulk to them, and as ovals rather than circles to represent the effect of Lorentz contraction. In this example, some of the energy of the incoming quarks happens to produce something called a Higgs particle. That is, energy has been converted into matter via $E = mc^2$. The Higgs particle was predicted to exist in the 1960s, and the LHC finally confirmed its detection in 2012. When the LHC was being designed, it was expected to be almost certain to find the Higgs. In fact, the fear was that the Higgs was the only thing it would find at all, and that appears to be what has happened. If so, then many particle physicists will be very depressed, but no doubt they will use it as a justification for another project, even larger and more expensive than the LHC.

Finding momentum given m and v

example 5

It can be useful to be able to find a particle's momentum given its mass and velocity. We have

$$m^2 = E^2 - p^2$$

and

$$v = \frac{p}{E}.$$

Eliminating E and solving for p , we find that

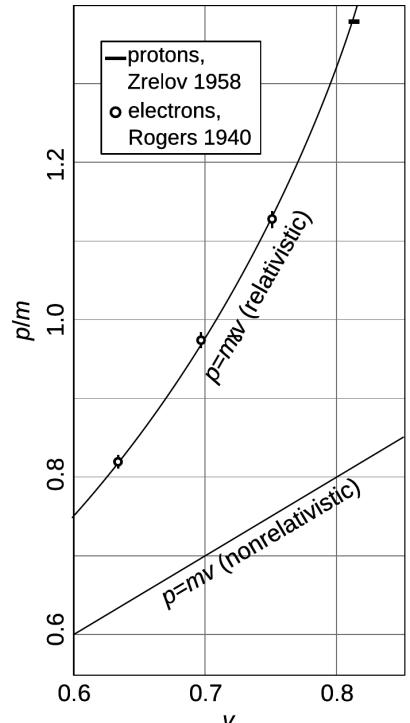
$$\begin{aligned} p &= mv \left(1 - v^2\right)^{-1/2} \\ &= m\gamma v, \end{aligned}$$

where we recall that $\gamma = 1/\sqrt{1 - v^2}$. In other words, the momentum is greater than the Newtonian value mv by a factor of γ . Figure g shows some experimental data testing this result. At low velocities, $\gamma \approx 1$, so we recover the Newtonian approximation. This makes sense because of the correspondence principle.

Electron-positron annihilation

example 6

In example 8, p. 77, we discussed electron-positron annihilation.



g / Two early high-precision tests of the relativistic equation $p = mv\gamma$ for momentum. Graphing p/m rather than p allows the data for electrons and protons to be placed on the same graph. Natural units are used, so that the horizontal axis is the velocity in units of c , and the vertical axis is the unitless quantity p/mc . The very small error bars for the data point from Zrelov are represented by the height of the black rectangle.

Recall that a positron is a form of antimatter, like an electron but with positive charge. If this process is to occur, then it must satisfy both conservation of charge and conservation of energy-momentum. If we only had to satisfy conservation of charge, then we could have a process like

$$e^- + e^+ \rightarrow \gamma,$$

in which an electron and a positron annihilate and produce a single gamma ray. The electric charge on the left adds up to zero, and the gamma ray is also uncharged, since it's a form of electromagnetic radiation.

But now let's consider conservation of energy-momentum. Let m be the mass of an electron or positron, and consider the annihilation process in the center-of-mass frame, i.e., the frame in which the total three-momentum is initially zero, so that the electron and positron are approaching each other symmetrically in a head-on collision. For simplicity, we consider the case where their kinetic energies are small; this tends to be true in most real-world examples, since a positron entering a piece of matter usually has time to slow down before being annihilated.

In the center-of-mass frame, with the electron and positron essentially at rest, we have initial energy-momentum vectors $(E, p) = (m, 0)$ and $(m, 0)$, adding up to $(2m, 0)$. This has a magnitude of $2m$. But the gamma ray must have an energy-momentum vector with a magnitude of zero, i.e., in 1+1 dimensions, something like (E, E) or $(E, -E)$. We can appreciate this problem even without all the fancy four-vectors. If we know that light has plain old momentum, then clearly the gamma ray has to go in some direction and have some momentum in that direction. But this is impossible if the initial momentum is zero.

For this reason, the simplest electron-positron annihilation process is

$$e^- + e^+ \rightarrow 2\gamma.$$

With the emission of *two* gamma rays, we can satisfy conservation of energy-momentum. The initial energy-momentum vector is $(2m, 0)$. We can match this in the final state if we let one gamma ray have energy-momentum (m, m) , and the other one $(m, -m)$, so that the sum is again $(2m, 0)$. Converting back into SI units, we find that the energy of each gamma ray is mc^2 , where m is the mass of the electron. This is the classic signature of matter-antimatter annihilation: back-to-back gamma rays, each with an energy equal to the mass (times c^2) of one of the annihilating particles.

This is the key to medical PET scans (p. 77), in which the patient is injected with a radioactive tracer that emits positrons. When one of these positrons annihilates with an electron, back-to-back

gamma rays are emitted. These are detected, and the two points at which they are detected define a line. In the most recent systems, the difference in arrival times for the two gammas can be determined accurately enough to allow the device to determine at the point on the line from which the emission occurred.

6.6 A tiny bit of linear algebra

With four-vectors, we have generalized the idea of a vector beyond anything that would have been recognized by, for example, my great-great grandfather, who was a steel salesman around 1900. Generalizations can actually be easier than the things they generalize, because when we make a generalization, we're basically throwing away some of the information. If my ancestor had been a general commodities salesman, then he might not have had to know all the specifics of grades of steel, carbon steel, stainless steel, and so on — he could have gotten away with only knowing basics like the fact that steel is heavy and durable. The mathematician's ultimate generalization of the idea of a vector is still more general than what we have considered so far, and it may be helpful at this point to understand what it is, so as to put things in perspective and make life easier. This subject goes by the name of linear algebra, and many students taking this type of physics course are either taking linear algebra concurrently with it, or have already taken it. If this is true for you, then you may just want to skim the following for review.

A *vector space* is a set of objects, which we refer to as vectors, along with operations of addition and scalar multiplication defined on the vectors. The scalars may be the real numbers or the complex numbers.² We require that the addition and scalar multiplication operations have the properties that addition is commutative ($\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$), that we have an additive identity 0 and additive inverses ($\mathbf{v} + (-\mathbf{v}) = 0$), and that both operations are associative and distributive in the ways that we would expect from the notation. The prototypical example of a vector space is vectors in three-dimensional space, with the scalars being the real numbers.

The vector space of polynomials

example 7

Consider the set of all polynomials. If we define addition of polynomials and multiplication of a polynomial by a real number in the obvious ways, then these functions are a vector space. Note that there is no well-defined division operation, since dividing a polynomial by a polynomial typically does not give a polynomial.

A set of vectors is said to be *linearly independent* if it is not

²It will probably not be obvious at this point why anyone would want to use complex numbers as scalars, but we will see when we study quantum physics that when a particle like an electron travels from A to B, it has a wave amplitude and wave phase that are most conveniently expressed as a complex number, while the “vector” is actually a description of the electron’s over-all state.

possible to form the zero vector as a linear combination of them. For vectors in three-dimensional space, a set of three vectors is not linearly independent if they lie in the same plane. The set of polynomials $\{1, x\}$ is linearly independent, but the set $\{P, Q, R\}$, where $P = 1$, $Q = 1 - x$, and $R = 1 + x$, is not, because $-2P + Q + R = 0$.

A *basis* for a vector space is a linearly independent set of vectors, called basis vectors, such that any vector can be formed as a linear combination of basis vectors. The standard basis for vectors in two-dimensional space is $\{\hat{x}, \hat{y}\}$, while a possible basis for the polynomials is the infinite set $\{1, x, x^2, x^3, \dots\}$. A basis exists for any vector space, and in fact there are normally many different bases to choose from, with none being preferred. In the plane, for example, we can choose to rotate the standard $\{\hat{x}, \hat{y}\}$ basis by any angle we like. Every basis for a given vector space has the same number of elements, and this number is called the *dimension* of the vector space. The plane is a two-dimensional vector space. The polynomials are an infinite-dimensional vector space.

A *linear operator* is a function \mathcal{O} that takes a vector as an input and gives a vector as an output, with the properties $\mathcal{O}(\mathbf{u} + \mathbf{v}) = \mathcal{O}(\mathbf{u}) + \mathcal{O}(\mathbf{v})$ and $\mathcal{O}(\alpha\mathbf{u}) = \alpha\mathcal{O}(\mathbf{u})$. A rotation in the plane is a linear operator.

Differentiation as a linear operator

example 8

Consider the set of all differentiable functions, taken as a vector space over either the real numbers or the complex numbers. Then the derivative is a linear operator, as is the second derivative.

For vectors in three-dimensional space, we have a dot product, which is a function that takes two vectors as inputs and gives a scalar as its output. A vector space may or may not come equipped with such an operation. If it does, we call the operation an *inner product*.³

If we have an inner product, then we automatically get a notion of magnitude, whose square we get by taking the dot product of a vector with itself $|\mathbf{v}|^2 = \mathbf{v} \cdot \mathbf{v}$. Conversely, if we have a way of defining a magnitude, then we get an inner product for free, since $|\mathbf{u} + \mathbf{v}|^2 = |\mathbf{u}|^2 + |\mathbf{v}|^2 + 2\mathbf{u} \cdot \mathbf{v}$ can be used to define $\mathbf{u} \cdot \mathbf{v}$.

The inner product can often be described as a measure of how similar two vectors are. For example, in the Euclidean plane, vectors that are perpendicular to each other have a dot product of zero, which tells you that they lie along lines that are completely different. When two vectors have an inner product of zero, we say that they are *orthogonal*.

³We also require that the inner product be linear, $\mathbf{u} \cdot (\alpha\mathbf{v} + \beta\mathbf{w}) = \alpha\mathbf{u} \cdot \mathbf{v} + \beta\mathbf{u} \cdot \mathbf{w}$. Often we want $\mathbf{u} \cdot \mathbf{u} > 0$, but this is not a requirement we want to impose on four-vectors, because their squared magnitudes can be negative.

Inner product on four-vectors

example 9

In this language, we could describe the squared interval \mathcal{I}^2 as the inner product of a displacement four-vector with itself. If a certain frame of reference in relativity has unit vectors $\hat{\mathbf{t}}$, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, then each of these is orthogonal to the other three. The orthogonality $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = 0$ says that the x and y axes have a 90-degree angle between them. The orthogonality $\hat{\mathbf{t}} \cdot \hat{\mathbf{x}} = 0$ says that an observer whose time axis is t considers events along the x axis to be simultaneous (figure c, p. 87).

When a vector space is finite-dimensional and a basis has been chosen, then if we wish we can represent vectors in column vector notation. For example, in the space of first-order polynomials with the basis $\{1, x\}$, the polynomial $3 + 5x$ can be represented by $\begin{pmatrix} 3 \\ 5 \end{pmatrix}$. Linear operators can similarly be represented by matrices, but we will seldom find this possible or useful in this book. For example, we can't represent the derivative as a matrix, because the vector space is infinite-dimensional.

6.7 ★ Tachyons

If we had objects in our universe that could go faster than c , then their energy-momentum vectors would be spacelike. The squared magnitude of a spacelike vector is negative, so such objects would have to have $m^2 < 0$, i.e., their masses would have to be imaginary numbers. Hypothetical subatomic particles with these properties are called tachyons (“TACK-y-ons,” from a Greek word meaning “fast”).

In section 4.1, p. 88, we saw a theoretical argument that no continuous process of acceleration can boost a material object past c , and this is also confirmed by experiments such as the one described in sec. 1.6.2, p. 20. That doesn't, however, address the question of whether one could surpass c through some discontinuous process, such as the “jump to hyperspace” in Star Wars. This loophole now appears to be closed off. We observe that mass is a permanent, fixed property of a material object, and it therefore seems that a material object like the Millennium Falcon could not go faster than c , or else it would have to be transformed into a cloud of tachyons, along with its passengers. However, if tachyons existed, then we could use them to communicate faster than the speed of light (which seems cool), and also to send information back into the past (p. 92).

But do tachyons exist? This is a question that can only be answered by searching experimentally. The most obvious experimental signature of tachyons would be motion at speeds greater than c . Negative results were reported by Murthy and later in 1988 by Clay, who studied showers of particles created in the earth's atmosphere by cosmic rays, looking for precursor particles that arrived before the first gamma rays. One could also look for particles with spacelike

energy-momentum vectors. Alvager and Erman, in a 1965 experiment, studied the radioactive decay of thulium-170, and found that no such particles were emitted at the level of 1 per 10,000 decays.

Some subatomic particles, such as dark matter and neutrinos, don't interact strongly with matter, and are therefore difficult to detect directly. It's possible that tachyons exist but don't interact strongly with matter, in which case they would not have been detectable in the experiments described above. In this scenario, it might still be possible to infer their existence indirectly through missing energy-momentum in nuclear reactions. This is how the neutrino was first discovered. An accelerator experiment by Baltay in 1970 searched for reactions in which the missing energy-momentum was spacelike, and found no such events. They put an upper limit of 1 in 1,000 on the probability of such reactions under their experimental conditions.

For a long time after the discovery of the neutrino, very little was known about its mass, so it was consistent with the experimental evidence to imagine that one or more species of neutrinos were tachyons, and Chodos *et al.* made such speculations in 1985. On p. 21 I described a 2011 experiment in which neutrinos were believed to have been seen moving at a speed slightly greater than c . The experiment turned out to be a mistake, but if it had been correct, then it would have proved that neutrinos were tachyons. An experiment called KATRIN, currently nearing the start of operation at Karlsruhe, will provide the first direct measurement of the mass of the neutrino, by measuring very precisely the missing energy-momentum in the decay of hydrogen-3.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

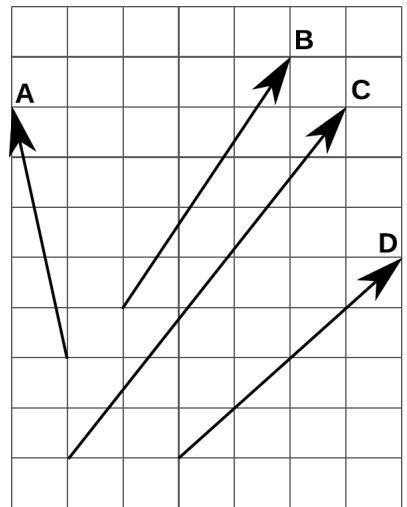
1 The figure shows four different four-vectors. Suppose someone tells you that two of these are actually the energy-momentum vectors of the same particle, measured on two different occasions. Determine which two they are.

2 (a) Find a relativistic equation for the velocity of an object in terms of its mass and momentum (eliminating γ). Start from the result of example 5, p. 153. Use natural units (i.e., discard factors of c) throughout. ✓

(b) Show that your result is approximately the same as the nonrelativistic value, p/m , at low velocities.

(c) Show that very large momenta result in speeds close to the speed of light.

(d) Insert factors of c to make your result from part a usable in SI units. ✓



Problem 1.

3 Example 5 on p. 153 demonstrated how to find the momentum p of an object in terms of m and v . This expression was made more compact by reexpressing part of it in terms of γ . Do a similar calculation to find the energy E in terms of m along with v and/or γ . ▷ Solution, p. 445

4 Example 5 on p. 153 showed that for an object with nonzero mass, $p = m\gamma v$. Expand this in a Taylor series, and find the first two nonvanishing terms. Explain why the vanishing terms are the ones that should vanish physically. Show that the first term is the newtonian expression.

5 For vectors in two dimensions, which of the following are possible choices of a basis?

$$\{\hat{x}\} \quad \{\hat{x}, \hat{y}\} \quad \{-\hat{x}, \hat{x} + \hat{y}\} \quad \{\hat{x}, \hat{y}, \hat{x} + \hat{y}\}$$

▷ Solution, p. 445

6 (a) Consider the set of vectors in two dimensions. This set P is a vector space, and can be visualized as a plane, with each vector being like an arrow that extends from the origin to a particular point. Now consider the line ℓ defined by the equation $y = x$ in Cartesian coordinates, and the ray r defined by $y = x$ with $x \geq 0$. Sketch ℓ and r . If we consider ℓ and r as subsets of the arrows in P , is ℓ a vector space? Is r ?

(b) Consider the set C of angles $0 \leq \theta < 2\pi$. Define addition on C by adding the angles and then, if necessary, bringing the result back into the required range. For example, if $x = \pi$ and $y = 3\pi/2$, then $x+y = \pi/2$. Thus if we visualize C as a circle, every point on the circle has a single number to represent it, not multiple representations such as $\pi/2$ and $5\pi/2$. Suppose we want to make C into a vector space over the real numbers, so that elements of C are the vectors, while a scalar α can be *any* real number, not just a number from 0 to 2π . Then for example if $\alpha = 2$ is a scalar and $v = \pi$ is a vector, then $\alpha v = 0$. Find an example to prove that C is not a vector space, because it violates the distributive property $\alpha(v+w) = \alpha v + \alpha w$.

▷ Solution, p. 445

7 In the SI, we have three base units, the kilogram, the meter, and the second. From these, we form expressions such as m/s to represent units of velocity, and $\text{kg}\cdot\text{m/s}^2$ for force. Show that these expressions form a vector space with the rational numbers as the scalars. What operation on the units should we take as the “addition” operation? What operation should scalar “multiplication” be?

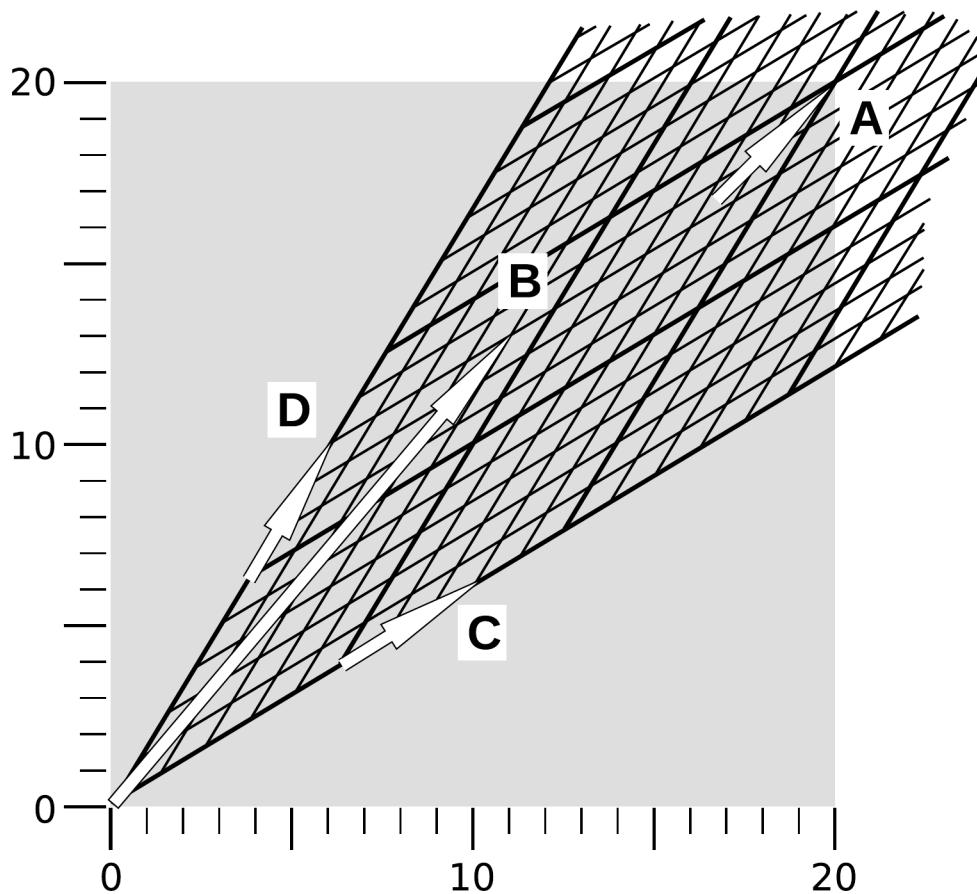
▷ Solution, p. 445

Exercise 6A: Sports in slowlightland

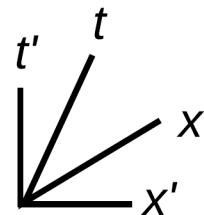
In Slowlightland, the speed of light is $20 \text{ mi/hr} \approx 32 \text{ km/hr} \approx 9 \text{ m/s}$. Think of an example of how relativistic effects would work in sports. Things can get very complex very quickly, so try to think of a simple example that focuses on just one of the following effects:

- relativistic momentum
- relativistic kinetic energy
- relativistic addition of velocities
- time dilation and length contraction
- Doppler shifts of light
- equivalence of mass and energy
- time it takes for light to get to an athlete's eye
- deflection of light rays by gravity

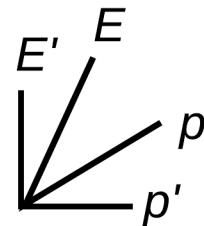
Exercise 6B: Four-vectors and inner products



Interpret the axes as either



or



The figure shows four-vectors **A** through **D**. To avoid cluttering the diagram, only one vector has its tail extended all the way to the origin, but conceptually all four do. As described at the side of the figure, the axes can be interpreted as either spacetime axes or energy-momentum axes. All components are integers in both frames of reference. When reading off components in the primed coordinates, which don't have grid-lines, you will want to use a ruler or a credit card, because otherwise some of the results are optical illusions that look like the wrong value.

1. If the graph is of spacetime, which vector could be a segment of the world-line of a light ray?

If the graph is reinterpreted an energy-momentum graph, what does this tell us about the ray's energy compared to its momentum?

2. If the graph is of spacetime, which vector could be a segment of the world-line of someone who is at rest in the (t, x) frame?

If it's energy-momentum, what can we say about this person's momentum?

How does this person describe the events at the tip and tail of vector **C**?

Turn the page.

3. The inner product of two four-vectors can be evaluated as $(p, q) \cdot (r, s) = pr - qs$. Consider the graph as an energy-momentum graph. In each of the following, calculate the inner product two different ways and give an interpretation:

A · A, evaluated in the (E, p) frame

A · A, evaluated in the (E', p') frame

Why does this make sense?

D · D, evaluated in the (E, p) frame

D · D, evaluated in the (E', p') frame

What does this represent physically?

C · D, evaluated in the (E, p) frame

C · D, evaluated in the (E', p') frame

What does an observer at rest in the unprimed frame say is happening?

B · D, evaluated in the (E, p) frame

B · D, evaluated in the (E', p') frame

Thermodynamics and the microscopic description of matter



A flame is used to fill a balloon with hot air.

Chapter 7

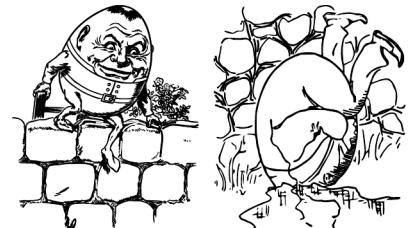
Statistics, equilibrium, and energy sharing

Looking at the two drawings in figure a, you shouldn't have any trouble figuring out which came first in time and which happened later. This process is *irreversible*: eggs don't unscramble.

As we saw in sec. 1.1, p. 13, this kind of seemingly common-sense observation is hard to explain when we consider that the laws of physics are completely symmetric with respect to time-reversal. Why, then, aren't all physical processes *reversible*?

There is something going on that has to do with randomness. Scrambling an egg seems to randomize it, in the same way that shaking up a box full of pennies makes it into some random combination of heads and tails. If someone had carefully arranged the pennies in the box so that they were all heads up, then a thorough shaking would eliminate that orderly pattern, and further shaking would be unlikely to restore it.

Although this seems like a step toward an explanation of time's arrow, it is not yet a complete explanation. For one thing, we could time-reverse the motion of the pennies, and then they would un-



a / Humpty Dumpty.

scramble themselves while again obeying Newton's laws. It does seem like it would be very difficult to set up the time-reversed motion in exactly the right way, but why is it that that sort of thing is so difficult?

In order to pursue this train of thought, we'll need some basic ideas about probability and statistics.

7.1 Basics of probability and statistics

Even if something is random, we can still understand it, and we can still calculate probabilities numerically.

7.1.1 Statistical independence

As an illustration of one general technique for calculating probabilities, suppose you are playing a 25-cent slot machine. Each of the three wheels has one chance in ten of coming up with a cherry. If all three wheels come up cherries, you win \$100. Even though the results of any particular trial are random, you can make certain quantitative predictions. First, you can calculate that your odds of winning on any given trial are $1/10 \times 1/10 \times 1/10 = 1/1000 = 0.001$. Here, I am representing the probabilities as numbers from 0 to 1, which is clearer than statements like "The odds are 999 to 1," and makes the calculations easier. A probability of 0 represents something impossible, and a probability of 1 represents something that will definitely happen.

This calculation was based on the following principle:

the law of independent probabilities

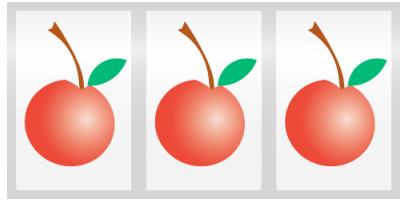
If the probability of one event happening is P_A , and the probability of a second statistically independent event happening is P_B , then the probability that they will both occur is the product of the probabilities, $P_A P_B$.

This can be taken as the definition of statistical independence.

Note that this only applies to independent probabilities. For example, if P_M is the probability that a randomly chosen American owns a motorcycle, and P_K is the probability that a person from the same population gets killed in a motorcycle accident, then the probability of both M and K is *much* greater than the product $P_M P_K$. This is because when K happens, it's usually *because* M was true (the exceptions being cases like people who take a ride on their friend's bike).

7.1.2 Addition of probabilities

The law of independent probabilities tells us to use multiplication to calculate the probability that both A and B will happen, assuming the probabilities are independent. What about the probability of an "or" rather than an "and"? If two events A and B



b / The probability that one wheel will give a cherry is $1/10$. The probability that all three wheels will give cherries is $1/10 \times 1/10 \times 1/10$.

are mutually exclusive, then the probability of one or the other occurring is the sum $P_A + P_B$. For instance, a bowler might have a 30% chance of getting a strike (knocking down all ten pins) and a 20% chance of knocking down nine of them. The bowler's chance of knocking down either nine pins or ten pins is therefore 50%.

It does not make sense to add probabilities of things that are not mutually exclusive, i.e., that could both happen. Say I have a 90% chance of eating lunch on any given day, and a 90% chance of eating dinner. The probability that I will eat either lunch or dinner is not 180%.

7.1.3 Normalization

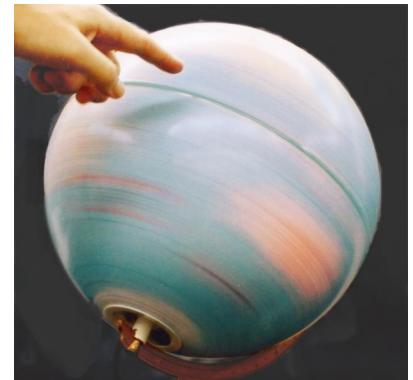
If I spin a globe and randomly pick a point on it, I have about a 70% chance of picking a point that's in an ocean and a 30% chance of picking a point on land. The probability of picking either water or land is $70\% + 30\% = 100\%$. Water and land are mutually exclusive, and there are no other possibilities, so the probabilities had to add up to 100%. It works the same if there are more than two possibilities — if you can classify all possible outcomes into a list of mutually exclusive results, then all the probabilities have to add up to 1, or 100%. This property of probabilities is known as normalization.

7.1.4 Averages

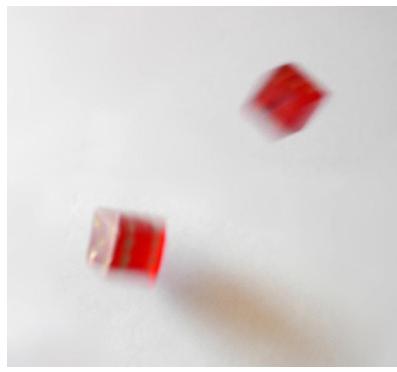
Another way of dealing with randomness is to take averages. The casino knows that in the long run, the number of times you win will approximately equal the number of times you play multiplied by the probability of winning. In the slot-machine game described on page 168, where the probability of winning is 0.001, if you spend a week playing, and pay \$2500 to play 10,000 times, you are likely to win about 10 times ($10,000 \times 0.001 = 10$), and collect \$1000. On the average, the casino will make a profit of \$1500 from you. This is an example of the following rule.

Rule for Calculating Averages

If you conduct N identical, statistically independent trials, and the probability of success in each trial is P , then on the average, the total number of successful trials will be NP . If N is large enough, the relative error in this estimate will become small.



c / Normalization: the probability of picking land plus the probability of picking water adds up to 1.



d / Why are dice random?

self-check A

Which of the following things *must* be independent, which *could* be independent, and which definitely are *not* independent? (1) the probability of successfully making two free-throws in a row in basketball; (2) the probability that it will rain in London tomorrow and the probability that it will rain on the same day in a certain city in a distant galaxy; (3) your probability of dying today and of dying tomorrow. \triangleright Answer, p. 456

Discussion questions

A Why isn't it valid to define randomness by saying that randomness is when all the outcomes are equally likely?

B Newtonian physics is an essentially perfect approximation for describing the motion of a pair of dice. If Newtonian physics is deterministic, why do we consider the result of rolling dice to be random?

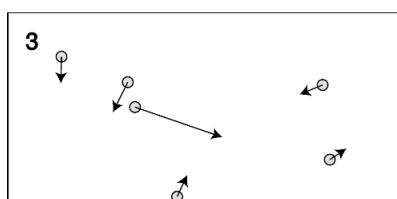
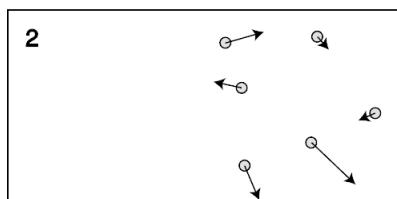
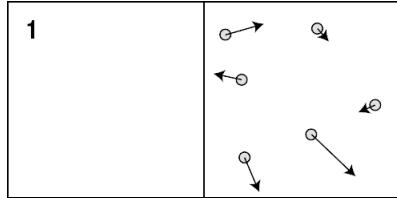
7.2 A statistical argument for irreversible processes

We can now apply our knowledge of probability to begin to flesh out our statistical explanation of why certain processes are irreversible. Figure e/1 shows a box in which all the atoms of the gas are confined on one side. We very quickly remove the barrier between the two sides, e/2. Each snapshot shows both the positions and the momenta of the atoms, which is enough information to allow us in theory to extrapolate the behavior of the system into the future, or the past. After some time, the system has reached a state, e/3, that seems like a typical one that we would observe if the gas hadn't been prepared in any special way.

Let n be the number of atoms in the gas. In figure e, $n = 6$. If we pick a state of the system completely at random, then by symmetry any atom's probability of being on the right is $1/2$. If each atom's probability is independent of the others,¹ then the probability of finding *all* the atoms on the right is 2^{-n} . When n is a relatively small number like 6, this will actually happen once in a while, as in figure f.

Suppose we show figure e/2 to a friend without any further information, and ask her what she can say about the system's behavior in the future. She doesn't know how the system was prepared. Perhaps, she thinks, it was just a strange coincidence that all the atoms happened to be in the right half of the box at this particular moment. In any case, she knows that this unusual situation won't last

¹Independence is an approximation. The atoms have some size, so there will be a tendency for them to crowd each other out. But for this to be a significant effect, we would need to consider gases under considerable compression. In an extreme case, the gas could be very close to the point at it would condense into a liquid or solid. Under such conditions, independence would be a very poor approximation.



e / A gas expands freely, doubling its volume.

for long. She can predict that after the passage of any significant amount of time, a surprise inspection is likely to show roughly half the atoms on each side. The same is true if you ask her to say what happened in the past. She doesn't know about the barrier, so as far as she's concerned, extrapolation into the past is exactly the same kind of problem as extrapolation into the future. We just have to imagine reversing all the momentum vectors, and then all our reasoning works equally well for backwards extrapolation. She would conclude, then, that the gas in the box underwent an unusual fluctuation, f , and she knows that the fluctuation is unlikely to exist very far into the future, or to have existed very far into the past.

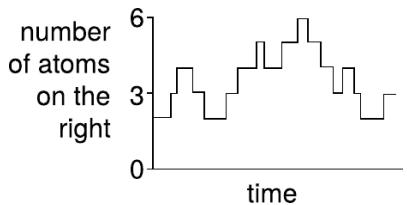
Now $2^{-6} \approx 0.015$ is not really all that small. But suppose we have $n = 100$ atoms. One hundred atoms is still a very small number of atoms by normal standards. A one-liter chamber with only 100 atoms in it would still qualify as an exceptionally good laboratory vacuum. We have $2^{-100} \approx 10^{-30}$, which is very, very small. If we show our friend all hundred atoms on the right, she feels that we've played a trick on her. If she could check once every millisecond, then she would expect to find such a fluctuation only once in about 10^{20} years, which is many orders of magnitude more than her lifetime.

So when n is even somewhat large, and your friend looks at the system and sees all n of the atoms on the right side of the box, she will infer that this is not a fluctuation. She can tell that you've played a trick on her by preparing the system in a special way.

Now when we look around our universe, we see that the distribution of matter is far from uniform. There are vast voids between the galaxies, with only about one molecule on the average in a volume the size of a bus. There are also obviously regions with a lot more matter, such as the region you're occupying right now. If the universe's state soon after the big bang had been chosen at random from among all possible states, then its current state would be absurdly unlikely. We can therefore conclude that our big bang was not a very typical one compared to all possible big bangs. This provides our best current explanation of the “arrow of time” — the past is the direction toward the big bang.

7.3 Sizes of fluctuations

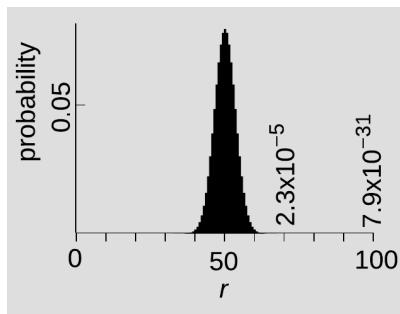
In the thought experiment of the atoms bouncing around between the two sides of the box, we've considered only two cases: the typical case, where about half the atoms are in each side, and the extreme case where all are on one side. The real world isn't like either of these extremes. We expect to see fluctuations, just not huge ones. It's instructive to consider a simple example in which we can easily estimate the the typical sizes of fluctuations. Suppose that we have three atoms in our box, and we label them A, B, and C. Then we can



f / An unusual fluctuation in the distribution of the atoms between the two sides of the box. There has been no external manipulation as in figure e/1.

r	atoms on the left	atoms on the right
0	A, B, C	none
1	B, C	A
1	A, C	B
1	A, B	C
2	A	B, C
2	B	A, C
2	C	A, B
3	none	A, B C

g / A list of the possible ways of putting atoms A, B, and C on the two sides of the box.



h / Probability of finding r out of $n = 100$ balls on the right-hand side of the box. Two values that are too small to see are labeled with numbers.

have A on the left or right, which is two possibilities, and similarly for B and C. This makes a total of $2 \times 2 \times 2 = 8$ possibilities, which is not too many to list by hand. Let r be the number of atoms on the right. Then we have the eight possibilities listed in table g.

Counting up all the possibilities for each value of r , we have the following for the total number of possibilities M at each value of r :

r	M
0	1
1	3
2	3
3	1

Thus the probability of $r = 0$ is $P(0) = 1/8$, and so on.

As a check on our work, we see that the normalization works out, $1/8 + 3/8 + 3/8 + 1/8 = 1$. Furthermore, it makes sense that the probabilities form a symmetric pattern above and below the value 1.5, because our rule for calculating averages predicts that the average value of r should be $3 \times \frac{1}{2}$.

We see that in our example, the extreme values of r are considerably less likely than the moderate values, with probabilities that are one third as big. For larger values of n , this trend becomes more pronounced, and the typical size of the fluctuations in r become very small compared to r .

For example, when $n = 100$, we find (2178) that the probability of getting $r = 70$ is only 2.3×10^{-5} , which is very small. Thus for $n = 100$, even a 20% fluctuation in r is rare. In the graph in figure h, not only is the central peak of the bell-shaped curve fairly narrow, but its “tails” represent extremely small probabilities, as shown by the labels for the invisibly small numbers. And this was only for $n = 100$. For real-world quantities of atoms, we often have $n \sim 10^{23}$, which means that fluctuations will be immeasurably smaller still. We will estimate the sizes of these fluctuations by a different technique in sec. 14.2, p. 328, which will allow us to give a clear proof that they become negligible for large n , in this sense.

7.4 Equipartition

Betsy Salazar of Redwood Cove, California, has 37 pet raccoons, which is theoretically illegal. She admits that she has trouble telling them apart, but she tries to give them all plenty of care and affection (which they reciprocate). There are only so many hours in a day, so there is a fixed total amount of love. The raccoons share this love unequally on any given day, but *on the average* they all get the same amount. This kind of equal-sharing-on-the-average-out-of-some-total-amount is more concisely described using the term

equipartition, meaning equal partitioning, or equal sharing. If Betsy did keep track of how much love she lavished on each animal, using some numerical scale, we would have 37 numbers to keep track of. We say that the love is partitioned among 37 *degrees of freedom*.

By analogy with the raccoons, we've seen that physical systems don't like to put all their eggs in one basket. In our example of the atoms in a box, it is overwhelmingly likely, for real-world quantities of atoms, that the atoms will be shared almost equally between the two sides.

We haven't yet discussed the energy of the atoms, but it's in connection with energy that we normally use the term equipartition. Consider a system consisting of only two atoms, and let the atoms, like the raccoons, differ in size. One has mass M and other m . For simplicity, we take the system to be one-dimensional. Then the total kinetic energy of the two atoms is $\frac{1}{2}MV^2 + \frac{1}{2}mv^2$, where V is the velocity of the big atom and v the velocity of the little one. In thermodynamics we actually usually prefer to discuss momenta rather than velocities (2178), so using the identity $K = p^2/2m$, we have for the energy

$$\frac{P^2}{2M} + \frac{p^2}{2m}.$$

This total energy is conserved, but it can be shared ("partitioned") in various ways between the two atoms. It can be proved that, on the average, each atom has an equal share of the energy, regardless of the unequal masses.

If we consider all three dimensions of space, then equipartition says that on the average, an equal amount of energy goes to each dimension. For example, a particular atom will have some amount of kinetic energy in the form of its x motion, and on the average this will be the *same* as the amount in its y and z motions. In fancy language, each of these three "degrees of freedom" gets an equal share of energy on the average.

Equipartition holds even if the energy isn't kinetic energy or if the system isn't a gas. For example, in a solid, the atoms are stuck near certain positions, and they can't move freely. They can only oscillate about these positions, like masses on little springs. The potential energy of a spring is $(1/2)kx^2$, and because this is a constant multiplied by the square of something, equipartition still holds.

Heat is the random motion of atoms. In a hot substance, the average energy of an atom will be higher. Equipartition gives us a natural way to define temperature, as something like the average energy per degree of freedom. Of course, this number comes out to be inconveniently small in SI units, and in any case there was a preexisting system of temperature units, so we define a scale of temperature T with a fudge factor, notated $k/2$, such that equipartition

comes out like this:

Equipartition

Suppose that a system has an energy that can be expressed in the way described above (as a sum of terms that look like the squares of x 's and/or p 's), and the system is in equilibrium. Then each of these degrees of freedom has, on average, an energy equal to $\frac{1}{2}kT$.

The constant k is called the Boltzmann constant, and in SI units it equals 1.38×10^{-23} J/K. Here K stands for the kelvin, a unit of temperature. One degree kelvin is the same temperature difference as one degree celsius, but the zero of the kelvin scale is absolute zero, the coldest possible temperature, at which all molecular motion ceases.

Equipartition gives us a straightforward way to estimate the *heat capacities* of various substances, meaning how much energy it takes to raise their temperature by a certain amount.

Heating helium

example 1

Helium is a monatomic gas, so it has three degrees of freedom: motion in the x , y , and z directions. (Diatomeric and polyatomic gases have more degrees of freedom, because they can also rotate, and possibly vibrate. See sec. 7.5 and fig. j.) A liter of helium contains $n = 2.5 \times 10^{22}$ atoms. Each atom has an average kinetic energy of $(3/2)kT$, for a total of $(3/2)nkT$. The amount of heat required to raise its temperature by one degree is

$$\frac{3}{2}nk(1\text{ K}) = 0.51\text{ J.}$$

This is the heat capacity at constant volume for this volume of gas. Many other, related quantities can be defined. One can, for example, define the heat capacity per unit mass, the capacity per mole, or the capacity at constant pressure (meaning that the gas must be allowed to expand, and therefore use up some of the input energy by doing work on the walls of the container).

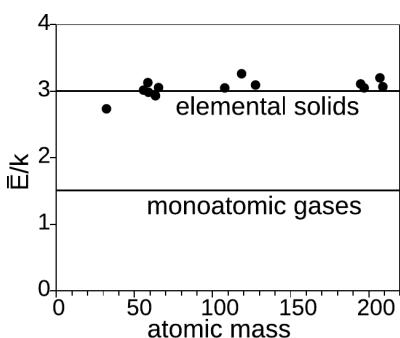
Heating solids

example 2

For a solid, we expect there to be a total of *six* energies per atom: the three kinetic energies, plus potential energies due to its displacement in the x , y , and z directions. Each of these carries an average energy $\frac{1}{2}kT$, for a total of $3kT$. In other words, a solid should have twice the heat capacity found in example 1 for a monatomic gas. This was discovered empirically by Dulong and Petit in 1819, as shown in figure i.

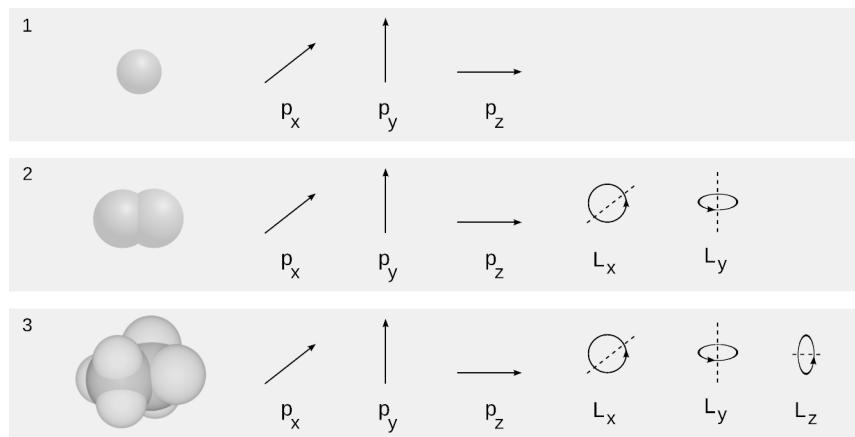
7.5 Heat capacities of gases

Examples 1 and 2 show that by measuring the heat capacity of a substance, we can find out how many microscopic degrees of freedom



i / Heat capacities of solids cluster around $3kT$ per atom. The notation \bar{E} means the average energy. The elemental solids plotted are the ones originally used by Dulong and Petit to infer empirically that the heat capacity of solids per atom was constant. (Modern data.)

it has. As a surprising example of how far this can carry us, consider the three types of molecules shown in figure j.



j / The differing shapes of a helium atom (1), a nitrogen molecule (2), and a difluoroethane molecule (3) have surprising macroscopic effects.

If the gas is monatomic, then we know from the result of example 1 that the relationship between its temperature and its internal energy is $E = (3/2)nkT$, so that to heat it by one degree, we the amount of heat we need to add² is $(3/2)nk$. Here the factor of 3 came ultimately from the fact that the gas was in a three-dimensional space, j/1. Moving in this space, each molecule can have momentum in the x, y, and z directions. It has three degrees of freedom.

What if the gas is not monatomic? Air, for example, is made of diatomic molecules, j/2. There is a subtle difference between the two cases. An individual atom of a monatomic gas is a perfect sphere, so it is exactly the same no matter how it is oriented. Because of this perfect symmetry, there is thus no way to tell whether it is spinning or not, and in fact we find that it can't rotate.³ The diatomic gas, on the other hand, can rotate end over end about the x or y axis, but cannot rotate about the z axis, which is its axis of symmetry. It has a total of five degrees of freedom. A polyatomic molecule with a more complicated, asymmetric shape, j/3, can rotate about all three axis, so it has a total of six degrees of freedom. Summarizing, we have

$$\alpha = \begin{cases} 3, & \text{monatomic gas} \\ 5, & \text{diatomic gas} \\ 6, & \text{polyatomic gas,} \end{cases}$$

for the number of degrees of freedom per molecule α , and the specific heat per molecule, at constant volume, is $(\alpha/2)k$.

²This is called the specific heat capacity, or specific heat, per atom. In practical work it is more common to multiply by Avogadro's number, which gives the specific heat per mole.

³Later in this course, we'll see in more detail that this is a consequence of how quantum physics describes motion

In 1819, Clément and Desormes did a simple experiment that we would today interpret as measuring this heat capacity for air. Its correct interpretation, in comparison with other gases, requires talking about the shapes of molecules. This was in a time when the periodic table hadn't been invented, and the existence of atoms and molecules was considered a mere hypothesis. Who would have dreamed that such simple observations, correctly interpreted, could give us this kind of glimpse of the microcosm?

7.6 The ideal gas law

Suppose that we take a liter of helium gas and get it very hot, increasing its absolute temperature by a factor of 4. This increases the average kinetic energy per atom by a factor of 4, and since kinetic energy depends on the square of the velocity, the typical speed is twice as big as before. This gas exerts a pressure on the sides of its container. The pressure is the force per unit area, measured in SI units of N/m^2 , which can be abbreviated as pascals, $1 \text{ Pa} = 1\text{N}/\text{m}^2$. For two different reasons, the doubling of the speeds of the atoms will now lead to an increase in the pressure. First, the frequency of collisions with the wall has been doubled. Second, the typical momentum of each atom has also been doubled. Since force is the rate of transfer of momentum, $F = \Delta p/\Delta t$, this means that the force is quadrupled. Our conclusion is that when we heat the gas by a factor of 4, while keeping the volume constant, the pressure goes up by a factor of 4. That is, the pressure of an ideal gas is proportional to its temperature. (By an ideal gas, we mean one whose atoms take up negligible space and do not interact with one another, e.g., by condensing to form droplets.) A straightforward extension of these arguments leads to the following law.

Ideal gas law

For an ideal gas,

$$PV = nkT,$$

where P is the pressure, V is the volume, n is the number of molecules, and T is the absolute pressure.

(You may have seen this written elsewhere as $PV = NRT$, where $N = n/N_A$ is the number of moles of atoms, $R = kN_A$, and $N_A = 6.0 \times 10^{23}$, called Avogadro's number, is essentially the number of hydrogen atoms in 1 g of hydrogen.)

Pressure in a car tire

example 3

- ▷ After driving on the freeway for a while, the air in your car's tires heats up from 10°C to 35°C . How much does the pressure increase?
- ▷ The tires may expand a little, but we assume this effect is small, so the volume is nearly constant. From the ideal gas law, the ratio of the pressures is the same as the ratio of the absolute temperatures,

$$\begin{aligned}P_2/P_1 &= T_2/T_1 \\&= (308\text{ K})/(283\text{ K}) \\&= 1.09,\end{aligned}$$

or a 9% increase.

Notes for chapter 7

≥172 Calculating probabilities for the two-sided box when n is large

We demonstrate techniques that make it possible to calculate these probabilities for large n .

In the text, we explicitly listed all the possibilities for distributing $n = 3$ atoms between the two sides of the box. This becomes impractical for large n . A mathematical technique that is more practical involves the use of numbers called binomial coefficients, which you may have encountered before in the context of algebra. In the case of $n = 3$, consider what happens when we calculate $(1 + x)^3 = 1 + 3x + 3x^2 + x^3$. The set of numbers 1, 3, 3, 1 is the same one we computed by listing possibilities for our atoms. The idea is that when we multiply out the cube, we have to go through all the combinations of 1's and x 's in the three factors. This is why these numbers are called binomial coefficients.

The binomial coefficient $\binom{n}{r}$ is defined as the number of ways of choosing r things from among n objects, with the order of the choices not being significant. For example, $\binom{3}{2} = 3$, because we can choose A and B, B and C, or A and C. (But we don't consider "A and C" to be a different choice than "C and A.")

A binomial coefficient can be computed in terms of the factorial function defined by $m! = 1 \times 2 \times \dots \times m$, as $\binom{n}{r} = n! / r!(n-r)!$. This works because we have n choices for the first object, $n-1$ for the second, and so on, down to $n-r+1$ choices for the final one. The product of these factors is the same as $n!/(n-r)!$. But this number counts possibilities more than once if they occur in a different order. Each possibility will be counted $r!$ times. Therefore we divide by $r!$ to get rid of the overcounting. For the example in the text, we have

$$\binom{100}{70} = 29372339821610944823963760,$$

which produces the probability

$$\frac{29372339821610944823963760}{2^{100}} = 2.3 \times 10^{-5}.$$

Calculators and computers will often spit out overflow errors when we ask them to calculate large binomial coefficients like this one. A good way of dealing with this problem is to work with the logarithms of the numbers, and to employ the approximation $\ln n! \approx n \ln n - n$, known as Stirling's formula, which comes from approximating the sum $\sum_{j=1}^n \ln j$ as $\int_1^n \ln x \, dx$.

≥173 Why we use p rather than v for our statistics

To describe the state of a system of particles (in one dimension, for simplicity), we could give either the list of numbers $(x_1, v_1, x_2, v_2, \dots)$ or the list $(x_1, p_1, x_2, p_2, \dots)$. If the particles all have the same mass, then it makes no difference which description we use. But if the particles have different masses, then the latter description is preferable, for the following reason. Suppose we make a graph-paper grid with x_1 on one axis and p_1 on the other. Then a mathematical result called Liouville's theorem shows that a reasonable initial guess about probabilities is to assign equal probability to each square on the grid.

Although the full statement of Liouville's theorem is beyond the scope of this book, it's not hard to see what goes wrong if we try to use velocities instead. When objects of different mass collide, they cannot transfer energy in an arbitrary way while still obeying conservation of momentum. For example, if you throw a golf ball with an energy of 10 joules, and the golf ball hits a bowling ball, it is not possible for the bowling ball to absorb all 10 joules of energy from the golf ball in the form of kinetic energy. The result is that statistically, in such collisions, there is a tendency for the less massive object to undergo bigger accelerations and have larger velocities. This means that it is not reasonable to assign the same probability per unit *velocity* to the golf ball as to the bowling ball.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 Many individuals carry the recessive gene for albinism, but they are not albino unless they receive the gene from both their parents. In the U.S., an individual's probability of receiving the gene from a given parent is about 0.014. What is the probability that a given child will be born albino? ✓

2 Many cities have laws limiting the number of dogs per household. Suppose that in one such city, the legal limit is two dogs. A survey shows that the average number of dogs per household is M , while the average number of dogs, in households that have dogs, is N . Find the fraction of households that have no dog. ✓

3 Devise a method for testing experimentally the hypothesis that a gambler's chance of winning at craps is independent of her previous record of wins and losses. If you don't invoke the mathematical definition of statistical independence, then you haven't proposed a test. This has nothing to do with the details of the rules of craps, or with the fact that it's a game played using dice.

4 (a) Show that under conditions of standard pressure and temperature, the volume of a sample of an ideal gas depends only on the number of molecules in it.
(b) One mole is defined as 6.0×10^{23} atoms. Find the volume of one mole of an ideal gas, in units of liters, at standard temperature and pressure (0°C and 101 kPa). ✓

5 A gas in a cylinder expands its volume by an amount dV , pushing out a piston. Show that the work done by the gas on the piston is given by $dW = P dV$.

6 A sample of gas is enclosed in a sealed chamber. The gas consists of molecules, which are then split in half through some process such as exposure to ultraviolet light, or passing an electric spark through the gas. The gas returns to the same temperature as the surrounding room, but the molecules remain split apart, at least for some amount of time. (To achieve these conditions, we would need an extremely dilute gas. Otherwise the recombination of the molecules would be faster than the cooling down to the same temperature as the room.) How does its pressure now compare with its pressure before the molecules were split?

7 The sun is mainly a mixture of hydrogen and helium, some of which is ionized. As a simplified model, let's pretend that it's made purely out of neutral, monatomic hydrogen, and that the whole mass of the sun is in thermal equilibrium. Given its mass, it would then contain 1.2×10^{57} atoms. It generates energy from nuclear reactions at a rate of 3.8×10^{26} W, and it is in a state of equilibrium in which this amount of energy is radiated off into space as light. Suppose that its ability to radiate light were somehow blocked. Find the rate at which its temperature would increase. ✓

8 Most of the atoms in the universe are in the form of gas that is not part of any star or galaxy: the intergalactic medium (IGM). The IGM consists of about 10^{-5} atoms per cubic centimeter, with a typical temperature of about 10^3 K. These are, in some sense, the density and temperature of the universe (not counting light, or the exotic particles known as "dark matter"). Calculate the pressure of the universe (or, speaking more carefully, the typical pressure due to the IGM). ✓

9 Our sun is powered by nuclear fusion reactions, and as a first step in these reactions, one proton must approach another proton to within a short enough range r . This is difficult to achieve, because the protons have electric charge $+e$ and therefore repel one another electrically. (It's a good thing that it's so difficult, because otherwise the sun would use up all of its fuel very rapidly and explode.) To make fusion possible, the protons must be moving fast enough to come within the required range. Even at the high temperatures present in the core of our sun, almost none of the protons are moving fast enough.

(a) For comparison, the early universe, soon after the Big Bang, had extremely high temperatures. Estimate the temperature T that would have been required so that protons with average energies could fuse. State your result in terms of r , the mass m of the proton, and universal constants.

(b) Show that the units of your answer to part a make sense.

(c) Evaluate your result from part a numerically, using $r = 10^{-15}$ m and $m = 1.7 \times 10^{-27}$ kg. As a check, you should find that this is much hotter than the sun's core temperature of $\sim 10^7$ K.

▷ Solution, p. 445

10 In metals, some electrons, called conduction electrons, are free to move around, rather than being bound to one atom. Classical physics gives an adequate description of many of their properties. Consider a metal at temperature T , and let m be the mass of the electron. Find expressions for (a) the average kinetic energy of a conduction electron, and (b) the average square of its velocity, \bar{v}^2 . (It would not be of much interest to find \bar{v} , which is just zero.) Numerically, $\sqrt{\bar{v}^2}$, called the root-mean-square velocity, comes out to be surprisingly large — about two orders of magnitude greater than the normal thermal velocities we find for atoms in a gas. Why?

Remark: From this analysis, one would think that the conduction electrons would contribute greatly to the heat capacities of metals. In fact they do not contribute very much in most cases; if they did, Dulong and Petit's observations would not have come out as described in the text. The resolution of this contradiction was only eventually worked out by Sommerfeld in 1933, and involves the fact that electrons obey the Pauli exclusion principle. \checkmark

Chapter 8

The macroscopic picture

In ch. 7, we began our study of thermodynamics by taking an abrupt deep dive into the microscopic world. Historically, this was not how the subject was developed, nor do people necessarily need to know anything about atoms and molecules when they check the pressure in their car's tires or take their child's temperature with a fever thermometer. Actually, the best way to operate is to develop some fluency with both the microscopic and macroscopic (i.e., human-scale) descriptions, and be able to switch back and forth between them whenever it's convenient.

You've already seen an example of the advantages of this "bilingual" approach in your study of electromagnetism. On the one hand, your ammeter displays certain behavior and obeys certain rules, such as Kirchoff's junction rule, which apply equally well regardless of whether we believe that it's effectively counting subatomic particles known as electrons. On the other hand, there are times when it is far more convenient to able to appeal to our microscopic knowledge. It would have been awkward to go through an entire semester of electromagnetism without ever using words like "particle" or "electron."

In this chapter, we will investigate some of the purely macroscopic concepts of thermodynamics. We start by taking another look at temperature and pressure. Pressure and temperature were fairly well understood in the age of Newton and Galileo, hundreds of years before there was any firm evidence that atoms and molecules even existed.

Unlike the conserved quantities such as mass-energy, momentum, and angular momentum, neither pressure nor temperature is additive. Two cups of coffee have twice the heat energy of a single cup, but they do not have twice the temperature. Likewise, the painful pressure on your eardrums at the bottom of a pool is not affected if you insert or remove a partition between the two halves of the pool.

We restrict ourselves to a discussion of pressure in fluids at rest and in equilibrium. In physics, the term "fluid" is used to mean either a gas or a liquid. The important feature of a fluid can be demonstrated by comparing with a cube of jello on a plate. The jello is a solid. If you shake the plate from side to side, the jello will

respond by shearing, i.e., by slanting its sides, but it will tend to spring back into its original shape. A solid can sustain shear forces, but a fluid cannot. A fluid does not resist a change in shape unless it involves a change in volume.

8.1 Pressure

If you're at the bottom of a pool, you can't relieve the pain in your ears by turning your head. The water's force on your eardrum is always the same, and is always perpendicular to the surface where the eardrum contacts the water. If your ear is on the east side of your head, the water's force is to the west. If you keep your ear in the same spot while turning around so your ear is on the north, the force will still be the same in magnitude, and it will change its direction so that it is still perpendicular to the eardrum: south. This shows that pressure has no direction in space, i.e., it is a scalar. The direction of the force is determined by the orientation of the surface on which the pressure acts, not by the pressure itself. A fluid flowing over a surface can also exert frictional forces, which are parallel to the surface, but the present discussion is restricted to fluids at rest.

Experiments also show that a fluid's force on a surface is proportional to the surface area. The vast force of the water behind a dam, for example, in proportion to the dam's great surface area. (The bottom of the dam experiences a higher proportion of its force.)

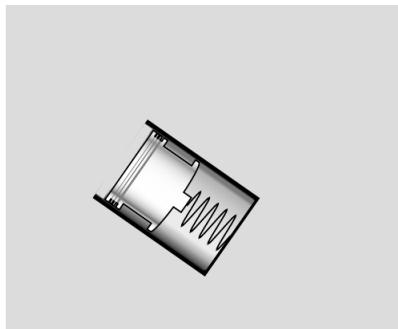
Based on these experimental results, it appears that the useful way to define pressure is as follows. The pressure of a fluid at a given point is defined as F_{\perp}/A , where A is the area of a small surface inserted in the fluid at that point, and F_{\perp} is the component of the fluid's force on the surface which is perpendicular to the surface. (In the case of a moving fluid, fluid friction forces can act parallel to the surface, but we're only dealing with stationary fluids, so there is only an F_{\perp} .)

This is essentially how a pressure gauge works. The reason that the surface must be small is so that there will not be any significant difference in pressure between one part of it and another part. The SI units of pressure are evidently N/m^2 , and this combination can be abbreviated as the pascal, $1 \text{ Pa}=1 \text{ N/m}^2$. The pascal turns out to be an inconveniently small unit, so car tires, for example, normally have pressures imprinted on them in units of kilopascals.

Pressure in U.S. units

example 1

In U.S. units, the unit of force is the pound, and the unit of distance is the inch. The unit of pressure is therefore pounds per square inch, or p.s.i. (Note that the pound is not a unit of mass.)



a / A simple pressure gauge consists of a cylinder open at one end, with a piston and a spring inside. The depth to which the spring is depressed is a measure of the pressure. To determine the absolute pressure, the air needs to be pumped out of the interior of the gauge, so that there is no air pressure acting outward on the piston. In many practical gauges, the back of the piston is open to the atmosphere, so the pressure the gauge registers equals the pressure of the fluid minus the pressure of the atmosphere.

Atmospheric pressure in U.S. and metric units *example 2*

▷ A figure that many people in the U.S. remember is that atmospheric pressure is about 15 pounds per square inch. What is this in metric units?

▷

$$\begin{aligned}(15 \text{ lb})/(1 \text{ in}^2) &= \frac{68 \text{ N}}{(0.0254 \text{ m})^2} \\ &= 1.0 \times 10^5 \text{ N/m}^2 \\ &= 100 \text{ kPa}\end{aligned}$$

8.1.1 Only pressure differences are normally significant.

If you spend enough time on an airplane, the pain in your ears subsides. This is because your body has gradually been able to admit more air into the cavity behind the eardrum. Once the pressure inside is equalized with the pressure outside, the inward and outward forces on your eardrums cancel out, and there is no physical sensation to tell you that anything unusual is going on. For this reason, it is normally only pressure differences that have any physical significance. Thus deep-sea fish are perfectly healthy in their habitat because their bodies have enough internal pressure to cancel the pressure from the water in which they live; if they are caught in a net and brought to the surface rapidly, they explode because their internal pressure is so much greater than the low pressure outside.

Getting killed by a pool pump *example 3*

▷ My house has a pool, which I maintain myself. A pool always needs to have its water circulated through a filter for several hours a day in order to keep it clean. The filter is a large barrel with a strong clamp that holds the top and bottom halves together. My filter has a prominent warning label that warns me not to try to open the clamps while the pump is on, and it shows a cartoon of a person being struck by the top half of the pump. The cross-sectional area of the filter barrel is 0.25 m^2 . Like most pressure gauges, the one on my pool pump actually reads the difference in pressure between the pressure inside the pump and atmospheric pressure. The gauge reads 90 kPa. What is the force that is trying to pop open the filter?

▷ If the gauge told us the absolute pressure of the water inside, we'd have to find the force of the water pushing outward and the force of the air pushing inward, and subtract in order to find the total force. Since air surrounds us all the time, we would have to do such a subtraction every time we wanted to calculate anything useful based on the gauge's reading. The manufacturers of the gauge decided to save us from all this work by making it read the difference in pressure between inside and outside, so all we have

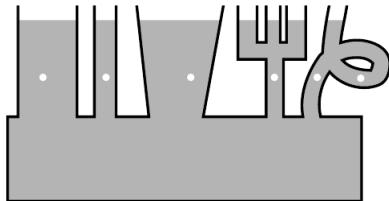
to do is multiply the gauge reading by the cross-sectional area of the filter:

$$\begin{aligned} F &= PA \\ &= (90 \times 10^3 \text{ N/m}^2)(0.25 \text{ m}^2) \\ &= 22000 \text{ N} \end{aligned}$$

That's a lot of force!



b / This doesn't happen. If pressure could vary horizontally in equilibrium, the cube of water would accelerate horizontally. This is a contradiction, since we assumed the fluid was in equilibrium.



c / The pressure is the same at all the points marked with dots.

The word “suction” and other related words contain a hidden misunderstanding related to this point about pressure differences. When you suck water up through a straw, there is nothing in your mouth that is attracting the water upward. The force that lifts the water is from the pressure of the water in the cup. By creating a partial vacuum in your mouth, you decreased the air’s downward force on the water so that it no longer exactly canceled the upward force.

8.1.2 Variation of pressure with depth

The pressure within a fluid in equilibrium can only depend on depth, due to gravity. If the pressure could vary from side to side, then a piece of the fluid in between, b, would be subject to unequal forces from the parts of the fluid on its two sides. Since fluids do not exhibit shear forces, there would be no other force that could keep this piece of fluid from accelerating. This contradicts the assumption that the fluid was in equilibrium.

self-check A

How does this proof fail for solids?

▷ Answer, p. 456

To find the variation with depth, we consider the vertical forces acting on a tiny, imaginary cube of the fluid having infinitesimal height dy and areas dA on the top and bottom. Using positive numbers for upward forces, we have

$$P_{bottom} dA - P_{top} dA - F_g = 0.$$

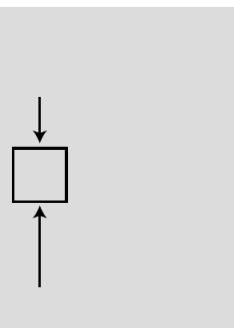
The weight of the fluid is $F_g = mg = \rho Vg = \rho dA dy g$, where ρ is the density of the fluid, so the difference in pressure is

$$dP = -\rho g dy.$$

[variation in pressure with depth for a fluid of density ρ in equilibrium; positive y is up.]

A more elegant way of writing this is in terms of a dot product, $dP = \rho \mathbf{g} \cdot d\mathbf{y}$, which automatically takes care of the plus or minus sign, depending on the relative directions of the \mathbf{g} and $d\mathbf{y}$ vectors, and avoids any requirements about the coordinate system.

The factor of ρ explains why we notice the difference in pressure when diving 3 m down in a pool, but not when going down 3 m of



d / This does happen. The sum of the forces from the surrounding parts of the fluid is upward, canceling the downward force of gravity.

stairs. The equation only tells us the difference in pressure, not the absolute pressure. The pressure at the surface of a swimming pool equals the atmospheric pressure, not zero, even though the depth is zero at the surface. The blood in your body does not even have an upper surface.

In cases where g and ρ are independent of depth, we can integrate both sides of the equation to get everything in terms of finite differences rather than differentials: $\Delta P = -\rho g \Delta y$.

self-check B

In which of the following situations is the equation $\Delta P = -\rho g \Delta y$ valid?

Why? (1) difference in pressure between a tabletop and the feet (i.e., predicting the pressure of the feet on the floor) (2) difference in air pressure between the top and bottom of a tall building (3) difference in air pressure between the top and bottom of Mt. Everest (4) difference in pressure between the top of the earth's mantle and the center of the earth (5) difference in pressure between the top and bottom of an airplane's wing

▷ Answer, p.

456

Pressure of lava underneath a volcano

example 4

▷ A volcano has just finished erupting, and a pool of molten lava is lying at rest in the crater. The lava has come up through an opening inside the volcano that connects to the earth's molten mantle. The density of the lava is 4.1 g/cm^3 . What is the pressure in the lava underneath the base of the volcano, 3000 m below the surface of the pool?

▷

$$\begin{aligned}\Delta P &= \rho g \Delta y \\ &= (4.1 \times 10^3 \text{ kg/m}^3)(9.8 \text{ m/s}^2)(3000 \text{ m}) \\ &= 1.2 \times 10^8 \text{ Pa}\end{aligned}$$

This is the difference between the pressure we want to find and atmospheric pressure at the surface. The latter, however, is tiny compared to the ΔP we just calculated, so what we've found is essentially the pressure, P .

Atmospheric pressure

example 5

Gases, unlike liquids, are quite compressible, and it follows from the ideal gas law that, at a given temperature, the density of a gas is approximately proportional to the pressure. To keep the writing simple, let's just call the constant of proportionality β , $\rho = \beta P$. Using this fact, we can find the variation of atmospheric pressure

with altitude, assuming constant temperature:

$$dP = -\rho g dy = -\beta Pg dy$$

$$\frac{dP}{P} = -\beta g dy$$

$$\ln P = -\beta gy + \text{constant} \quad [\text{integrating both sides}]$$

$$P = (\text{constant})e^{-\beta gy} \quad [\text{exponentiating both sides}]$$

Pressure falls off exponentially with height. There is no sharp cut-off to the atmosphere, but the exponential factor gets extremely small by the time you're ten or a hundred miles up.

8.2 Temperature

8.2.1 Thermal equilibrium

We saw in sec. 7.4, p. 172, that for a system in equilibrium, we can define temperature as a measure of the average energy \bar{E} per degree of freedom, $\bar{E} = \frac{1}{2}kT$. But when we do macroscopic measurements, we never explicitly see microscopic degrees of freedom such as the component p_x of a particular atom's momentum in the x direction. We can, however, say that temperature is a measure of how concentrated the heat energy is in an object. A large, massive object with very little heat energy in it has a low temperature. Of course, we can keep in mind our microscopic insight that the size and mass of the object matter because they relate to the number of microscopic degrees of freedom that it has.



If we're not going to focus on those microscopic degrees of freedom, then we should admit that we haven't really defined anything by saying that temperature measures "how concentrated" the energy is. A good way of handling this is to use an operational definition, i.e., definition of how to measure the thing in question. This is effectively what we did in ch. 1, where we effectively defined time as what a clock measures.

So how do we measure temperature? One common feature of all temperature-measuring devices is that they must be left for a while in contact with the thing whose temperature is being measured. When you take your temperature with a fever thermometer, you are waiting for the mercury inside to come up to the same temperature as your body. The thermometer actually tells you the temperature of its own working fluid (in this case the mercury). In general, the idea of temperature depends on the concept of thermal equilibrium. When you mix cold eggs from the refrigerator with flour that has been at room temperature, they rapidly reach a compromise temperature. What determines this compromise temperature is conservation of energy, and the amount of energy required to heat or cool each substance by one degree. But without even having constructed a temperature scale, we can see that the important point

e / Thermal equilibrium can be prevented. Otters have a coat of fur that traps air bubbles for insulation. If a swimming otter was in thermal equilibrium with cold water, it would be dead. Heat is still conducted from the otter's body to the water, but much more slowly than it would be in a warm-blooded animal that didn't have this special adaptation.

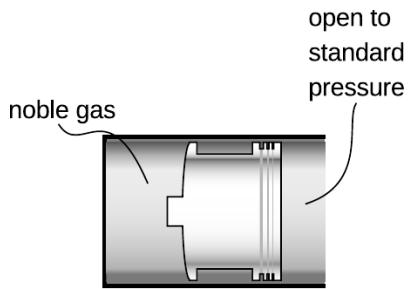
is the phenomenon of thermal equilibrium itself: two objects left in contact will approach the same temperature. We also assume that if object A is at the same temperature as object B, and B is at the same temperature as C, then A is at the same temperature as C. This statement is sometimes known as the zeroth law of thermodynamics, so called because after the first, second, and third laws had been developed, it was realized that there was another law that was even more fundamental.

Although we can understand the zeroth law and the operational definition of temperature without appealing to anything microscopic, they are compatible with the microscopic picture established in ch. 7. Equilibrium occurs when the energy per degree of freedom — or, roughly speaking, the energy per atom — is equalized between the two objects, and a thermometer works by equilibrating its own working fluid with the thing being measured.

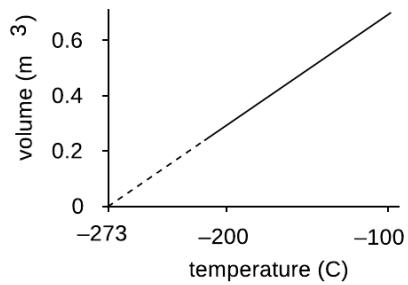
8.2.2 Thermal expansion

The familiar mercury thermometer operates on the principle that the mercury, its working fluid, expands when heated and contracts when cooled. In general, all substances expand and contract with changes in temperature. The zeroth law of thermodynamics guarantees that we can construct a comparative scale of temperatures that is independent of what type of thermometer we use. If a thermometer gives a certain reading when it's in thermal equilibrium with object A, and also gives the same reading for object B, then A and B must be the same temperature, regardless of the details of how the thermometers work.

What about constructing a temperature scale in which every degree represents an equal step in temperature? The Celsius scale has 0 as the freezing point of water and 100 as its boiling point. The hidden assumption behind all this is that since two points define a line, any two thermometers that agree at two points must agree at all other points. In reality if we calibrate a mercury thermometer and an alcohol thermometer in this way, we will find that a graph of one thermometer's reading versus the other is not a perfectly straight $y = x$ line. The subtle inconsistency becomes a drastic one when we try to extend the temperature scale through the points where mercury and alcohol boil or freeze. Gases, however, are much more consistent among themselves in their thermal expansion than solids or liquids, and the noble gases like helium and neon are more consistent with each other than gases in general. Continuing to search for consistency, we find that noble gases are more consistent with each other when their pressure is very low.



f / A simplified version of an ideal gas thermometer. The whole instrument is allowed to come into thermal equilibrium with the substance whose temperature is to be measured, and the mouth of the cylinder is left open to standard pressure. The volume of the noble gas gives an indication of temperature.



g / The volume of 1 kg of neon gas as a function of temperature (at standard pressure). Although neon would actually condense into a liquid at some point, extrapolating the graph gives to zero volume gives the same temperature as for any other gas: absolute zero.

As an idealization, we imagine a gas in which the atoms interact only with the sides of the container, not with each other. Such a gas is perfectly nonreactive (as the noble gases very nearly are), and never condenses to a liquid (as the noble gases do only at extremely low temperatures). Its atoms take up a negligible fraction of the available volume. Any gas can be made to behave very much like this if the pressure is extremely low, so that the atoms hardly ever encounter each other. Such a gas is called an ideal gas, and we define the Celsius scale in terms of the volume of the gas in a thermometer whose working substance is an ideal gas maintained at a fixed (very low) pressure, and which is calibrated at 0 and 100 degrees according to the melting and boiling points of water. The Celsius scale is not just a comparative scale but an additive one as well: every step in temperature is equal, and it makes sense to say that the difference in temperature between 18 and 28°C is the same as the difference between 48 and 58.

8.2.3 Absolute zero and the kelvin scale

We find that if we extrapolate a graph of volume versus temperature, the volume becomes zero at nearly the same temperature for all gases: -273°C . Real gases will all condense into liquids at some temperature above this, but an ideal gas would achieve zero volume at this temperature, known as absolute zero. At the macroscopic level, this is the justification for constructing the kelvin scale, introduced on p. 174.

Scientists use the celsius scale only for comparisons or when a change in temperature is all that is required for a calculation. Only on the kelvin scale does it make sense to discuss ratios of temperatures, e.g., to say that one temperature is twice as hot as another.

Which temperature scale to use

example 6

- ▷ You open an astronomy book and encounter the equation

$$(\text{light emitted}) = (\text{constant}) \times T^4$$

for the light emitted by a star as a function of its surface temperature. What temperature scale is implied?

- ▷ The equation tells us that doubling the temperature results in the emission of 16 times as much light. Such a ratio only makes sense if the Kelvin scale is used.

8.3 Heat

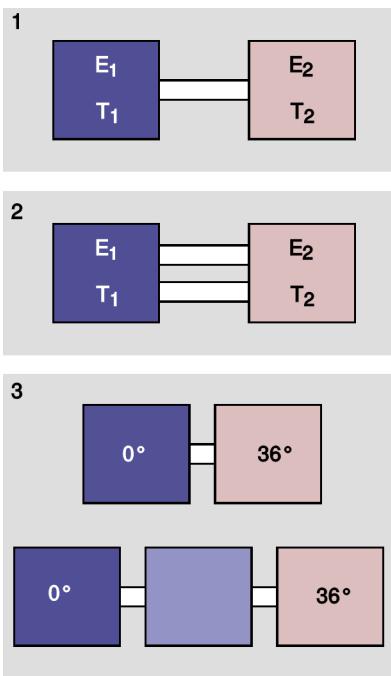
“Heat,” notated Q , is used in thermodynamics as a term for an amount of thermal energy that is transferred. When you put a bite of food in your mouth that is too hot, the pain is caused by the heat transferred from the food to your mouth. People discussing the weather may say “What about this heat today?” or “What about this temperature today?” as if the words were synonyms, but to a physicist they are distinct. Temperature is not additive, but heat is: two sips of hot coffee have the same temperature as one, but two sips will transfer twice the heat to your mouth. Temperature is measured in degrees, heat in joules.

If I give you an object, you can measure its temperature — physicists call temperature a “property of state,” i.e., you can tell what it is from the current state of the object. Heat is a description of a *process* of energy transfer, not a property of state.

It’s relatively easy to detect and measure a *transfer* of thermal energy (the hot bite of food), but to say how much thermal energy an object *has* is much harder — sometimes even impossible in principle.

Heat is distinguished from mechanical work because work is the transfer of energy by a macroscopically measurable force, e.g., the force of a baseball bat on the ball. No such force is needed in order to melt an ice cube; the forces are in microscopic collisions of water molecules with ice molecules.

Heat, like the flow of money or water, is a signed quantity, but the sign is a matter of definition. The bank’s debit is the customer’s withdrawal. It is an arbitrary choice whether to call Q positive when it flows from object A to object B or from B to A, and likewise for the work W . Similar choices arise in the description of flowing fluids or electric currents. We will usually adopt definitions such that as many heats and works as possible are positive. So by our definition, a cute 19th-century steam locomotive takes in positive heat from its boiler, does positive work to pull the cars, and spews out positive heat through its smokestack. When only a single object is being discussed, such as a cylinder of compressed air, we define a heat input as positive and a work output as positive, which is again in accord with the picture of the cute steam engine. No universally consistent convention is possible, since, e.g., if objects A, B, and C all interact, we will always have opposite signs for A’s work on B and B’s work on A, etc.



h / Discussion questions A-C.

Discussion questions

A Figure h/1 shows objects 1 and 2, each with a certain temperature T and a certain amount of thermal energy E . They are connected by a thin rod, so that eventually they will reach thermal equilibrium. We expect that the rate at which heat is transferred into object 1 will be given by some equation $dE_1/dt = k(\dots)$, where k is a positive constant of proportionality and “...” is some expression that depends on the temperatures. Suppose that the following six forms are proposed for the “...” in $dE_1/dt = k(\dots)$.

1. T_1
2. T_2
3. $T_1 - T_2$
4. $T_2 - T_1$
5. T_1/T_2
6. T_2/T_1

Give physical reasons why five of these are not possible.

B How should the rate of heat conduction in h/2 compare with the rate in h/1?

C The example in h/3 is different from the preceding ones because when we add the third object in the middle, we don't necessarily know the intermediate temperature. We could in fact set up this third object with any desired initial temperature. Suppose, however, that the flow of heat is *steady*. For example, the 36° object could be a human body, the 0° object could be the air on a cold day, and the object in between could be a simplified physical model of the insulation provided by clothing or body fat. Under this assumption, what is the intermediate temperature? How does the rate of heat conduction compare in the two cases?

D Based on the conclusions of questions A-C, how should the rate of heat conduction through an object depend on its length and cross-sectional area? If all the linear dimensions of the object are doubled, what happens to the rate of heat conduction through it? How would this apply if we compare an elephant to a shrew?

8.4 Adiabatic expansion of a gas

When you pop the cap off of a bottle of beer, you may see a cloud of mist. This is because the air and carbon dioxide inside the bottle are under pressure, and as they expand, they push on the surrounding air and do work. By conservation of energy, this work can only come at the expense of some loss of the gas's own thermal energy. It cools, and the cooling produces condensation of water vapor.

As an idealization of this type of process, we consider the expansion of a sealed sample of gas inside a cylinder, as it pushes a piston out. We assume that there is no heat conduction through the walls of the cylinder, either because the cylinder is well insulated or because the process is too fast for much heat conduction to occur. Classic examples of this is the expansion stroke of a steam engine,

or, in the opposite direction, the compression stroke of a gasoline engine. When this kind of process occurs, with no possibility of heat conduction, we refer to it as an *adiabatic* process.

As the volume of the cylinder increases by an infinitesimal amount dV , we have from the result of problem 5, p. 179, that it does work $dW = P dV$. Conservation of energy tells us that this will cool the gas by some amount dT , decreasing the gas's thermal energy E . This internal energy equals $(\alpha/2)nkT$, where $\alpha = 3, 5$, or 6 is the number of degrees of freedom, depending on whether the gas is monatomic, diatomic, or polyatomic (sec. 7.5, p. 174). We therefore have $(\alpha/2)nk dT + P dV = 0$. All three variables, T , P , and V , are changing simultaneously, but they are not independent of one another, because the ideal gas law constrains them to $PV = nkT$. We can use this constraint to eliminate any one of the variables. Let's use it to eliminate T . Using the product rule, $nk dT = P dV + V dP$, and separation of variables gives

$$\frac{dP}{P} = -\gamma \frac{dV}{V},$$

where $\gamma = 1 + 2/\alpha$ is referred to as the adiabatic gas constant or adiabatic index. Integrating both sides gives $\ln P = -\gamma \ln V + \text{const}$, or

$$P \propto V^{-\gamma}.$$

Measuring γ using the “spring of air”

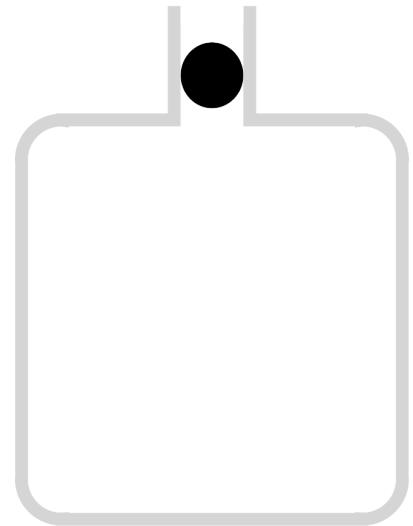
example 7

Figure i shows an experiment that can be used to measure the γ of a gas. When the mass m is inserted into bottle's neck, which has cross-sectional area A , the mass drops until it compresses the air enough so that the pressure is enough to support its weight. The observed frequency ω of oscillations about this equilibrium position y_0 can be used to extract the γ of the gas.

$$\begin{aligned}\omega^2 &= \frac{k}{m} \\ &= -\frac{1}{m} \left. \frac{dF}{dy} \right|_{y_0} \\ &= -\frac{A}{m} \left. \frac{dP}{dy} \right|_{y_0} \\ &= -\frac{A^2}{m} \left. \frac{dP}{dV} \right|_{V_0}\end{aligned}$$

We make the bottle big enough so that its large surface-to-volume ratio prevents the conduction of any significant amount of heat through its walls during one cycle, so $P \propto V^{-\gamma}$, and $dP/dV = -\gamma P/V$. Thus,

$$\omega^2 = \gamma \frac{A^2}{m} \frac{P_0}{V_0}$$



i / Example 7.

The Helmholtz resonator

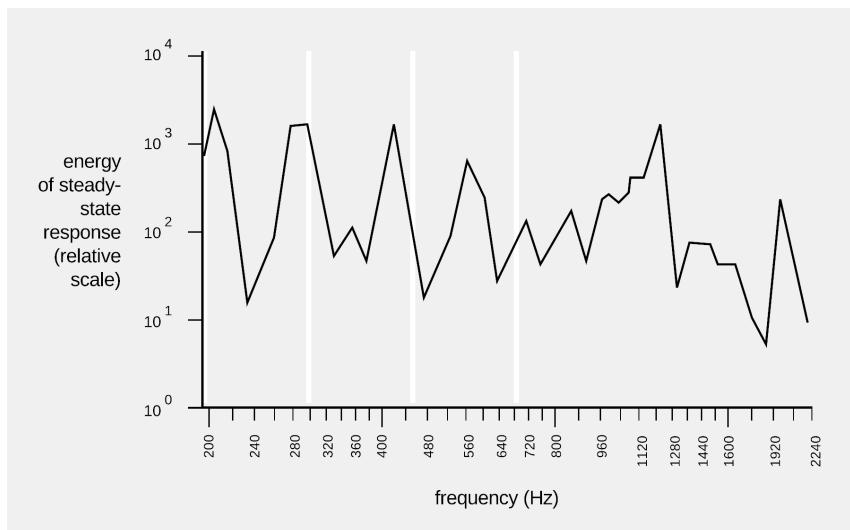
example 8

When you blow over the top of a beer bottle, you produce a pure tone. As you drink more of the beer, the pitch goes down. This is similar to example 7, except that instead of a solid mass m sitting inside the neck of the bottle, the moving mass is the air itself. As air rushes in and out of the bottle, its velocity is highest at the bottleneck, and since kinetic energy is proportional to the square of the velocity, essentially all of the kinetic energy is that of the air that's in the neck. In other words, we can replace m with $AL\rho$, where L is the length of the neck, and ρ is the density of the air. Substituting into the earlier result, we find that the resonant frequency is

$$\omega^2 = \gamma \frac{P_0}{\rho} \frac{A}{LV_0}.$$

This is known as a Helmholtz resonator. As shown in figure j, a violin or an acoustic guitar has a Helmholtz resonance, since air can move in and out through the f-holes.

j / The resonance curve of a 1713 Stradivarius violin, measured by Carleen Hutchins. There are a number of different resonance peaks, some strong and some weak; the ones near 200 and 400 Hz are vibrations of the wood, but the one near 300 Hz is a resonance of the air moving in and out through those holes shaped like the letter F. The white lines show the frequencies of the four strings.

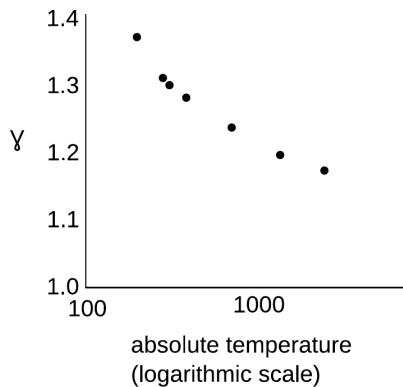


The speed of sound

example 9

We can get a rough and ready derivation of the equation for the speed of sound by analyzing the standing waves in a cylindrical air column as a special type of Helmholtz resonance (example 8), in which the cavity happens to have the same cross-sectional area as the neck. Roughly speaking, the regions of maximum density variation act like the cavity. The regions of minimum density variation, on the other hand, are the places where the velocity of the air is varying the most; these regions throttle back the speed of the vibration, because of the inertia of the moving air. If the cylinder has cross-sectional area A , then the “cavity” and “neck” parts of the wave both have lengths of something

like $\lambda/2$, and the volume of the “cavity” is about $A\lambda/2$. We get $v = f\lambda = (\dots)\sqrt{\gamma P_0/\rho}$, where the factor (...) represents numerical stuff that we can’t possibly hope to have gotten right with such a crude argument. The correct result is in fact $v = \sqrt{\gamma P_0/\rho}$. Isaac Newton attempted the same calculation, but didn’t understand the thermodynamic effects involved, and therefore got a result that didn’t have the correct factor of $\sqrt{\gamma}$.



k / Discussion question A.

Discussion question

A The figure shows a graph of the adiabatic index γ of carbon dioxide gas as a function of temperature. The graph goes down to the lowest temperature at which carbon dioxide is a gas. Solve the following two mysteries:

1. According to the presentation in this section, the value of γ seems like it should be a fixed property of the molecule. How can it vary with temperature? Hint: apply $\gamma = 1 + 2/\alpha$, and infer what is happening to α .
2. At low temperatures, γ is closer to the diatomic value than the polyatomic one. Why should this be? After all, “di-” means “two,” and “poly-” means many. Carbon dioxide has three atoms in it, not two.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 (a) Atmospheric pressure at sea level is 101 kPa. The deepest spot in the world's oceans is a valley called the Challenger Deep, in the Marianas Trench, with a depth of 11.0 km. Find the pressure at this depth, in units of atmospheres. Although water under this amount of pressure does compress by a few percent, assume for the purposes of this problem that it is incompressible.

(b) Suppose that an air bubble is formed at this depth and then rises to the surface. Estimate the change in its volume and radius.

▷ Solution, p. 446

2 Typically the atmosphere gets colder with increasing altitude. However, sometimes there is an *inversion layer*, in which this trend is reversed, e.g., because a less dense mass of warm air moves into a certain area, and rises above the denser colder air that was already present. Suppose that this causes the pressure P as a function of height y to be given by a function of the form $P = P_o e^{-ky}(1 + by)$, where constant temperature would give $b = 0$ and an inversion layer would give $b > 0$. (a) Infer the units of the constants P_o , k , and b . (b) Find the density of the air as a function of y , of the constants, and of the acceleration of gravity g . (c) Check that the units of your answer to part b make sense.

▷ Solution, p. 446

3 (a) The water molecule has a shape like a Mickey Mouse head. Would we expect it to act as a monatomic, diatomic, or polyatomic gas?

(b) In a certain steam engine's expansion stroke, the volume of the cylinder increases by a factor of 5.0. If the final pressure is 1.0 atmospheres (so that the net force on the piston vanishes at that point), find the initial pressure, in units of atmospheres. ✓

4 Show that for a gas undergoing adiabatic expansion, the temperature and pressure are related by a proportionality of the form $T \propto P^b$, and determine b in terms of the number of degrees of freedom α . ✓

5 Estimate the pressure at the center of the Earth, assuming it is of constant density throughout. The gravitational field g is not constant with respect to depth. It equals Gmr/b^3 for r , the distance from the center, less than b , the earth's radius. Here m is the mass of the earth, and G is Newton's universal gravitational constant, which has units of $\text{N}\cdot\text{m}^2/\text{kg}^2$.

- (a) State your result in terms of G , m , and b . ✓
- (b) Show that your answer from part a has the right units for pressure.
- (c) Evaluate the result numerically. ✓
- (d) Given that the earth's atmosphere is on the order of one thousandth the earth's radius, and that the density of the earth is several thousand times greater than the density of the lower atmosphere, check that your result is of a reasonable order of magnitude.

6 (a) Determine the ratio between the escape velocities from the surfaces of the earth and the moon. ✓

(b) The temperature during the lunar daytime gets up to about 130°C . In the extremely thin (almost nonexistent) lunar atmosphere, estimate how the typical velocity of a molecule would compare with that of the same type of molecule in the earth's atmosphere. Assume that the earth's atmosphere has a temperature of 0°C . ✓

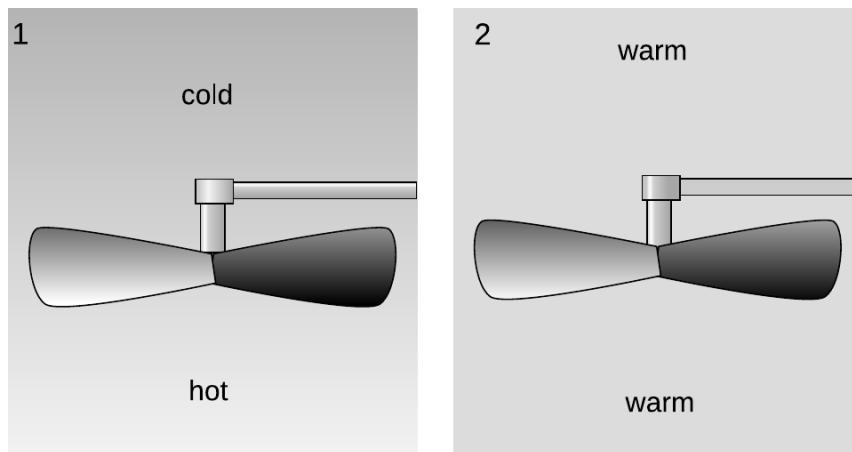
(c) Suppose you were to go to the moon and release some fluorocarbon gas, with molecular formula $\text{C}_n\text{F}_{2n+2}$. Estimate what is the smallest fluorocarbon molecule (lowest n) whose typical velocity would be lower than that of an N_2 molecule on earth in proportion to the moon's lower escape velocity. The moon would be able to retain an atmosphere made of these molecules. ✓

Chapter 9

Entropy

9.1 Heat engines

Heat may be more useful in some forms than in others, i.e., there are different grades of heat energy. In figure a/1, the difference in temperature can be used to extract mechanical work with a fan blade. This principle is used in power plants, where steam is heated by burning oil or by nuclear reactions, and then allowed to expand through a turbine which has cooler steam on the other side. On a smaller scale, there is a Christmas toy, b, that consists of a small propeller spun by the hot air rising from a set of candles, very much like the setup shown in figure a.



a / 1. The temperature difference between the hot and cold parts of the air can be used to extract mechanical energy, for example with a fan blade that spins because of the rising hot air currents. 2. If the temperature of the air is first allowed to become uniform, then no mechanical energy can be extracted. The same amount of heat energy is present, but it is no longer accessible for doing mechanical work.

In figure a/2, however, no mechanical work can be extracted because there is no difference in temperature. Although the air in a/2 has the same total amount of energy as the air in a/1, the heat in a/2 is a lower grade of energy, since none of it is accessible for doing mechanical work.

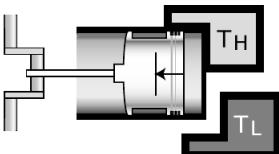
In general, we define a heat engine as any device that takes heat



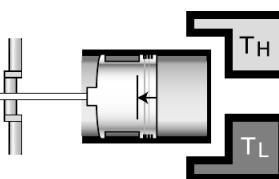
b / A heat engine. Hot air from the candles rises through the fan blades and makes the angels spin.



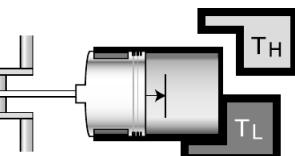
c / Sadi Carnot (1796-1832)



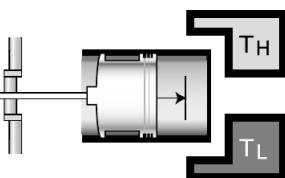
d / The beginning of the first expansion stroke, in which the working gas is kept in thermal equilibrium with the hot reservoir.



e / The beginning of the second expansion stroke, in which the working gas is thermally insulated. The working gas cools because it is doing work on the piston and thus losing energy.



f / The beginning of the first compression stroke. The working gas begins the stroke at the same temperature as the cold reservoir, and remains in thermal contact with it the whole time. The engine does negative work.



g / The beginning of the second compression stroke, in which mechanical work is absorbed, heating the working gas back up to T_H .

from a reservoir of hot matter, extracts some of the heat energy to do mechanical work, and expels a lesser amount of heat into a reservoir of cold matter. The efficiency of a heat engine equals the amount of useful work extracted, W , divided by the amount of energy we had to pay for in order to heat the hot reservoir. This latter amount of heat is the same as the amount of heat the engine extracts from the high-temperature reservoir, Q_H . By conservation of energy, we have $Q_H = W + Q_L$, where Q_L is the amount of heat expelled into the low-temperature reservoir, so the efficiency of a heat engine, W/Q_H , can be rewritten as

$$\text{efficiency} = 1 - \frac{Q_L}{Q_H}. \quad [\text{efficiency of any heat engine}]$$

(As described on p. 191, we take Q_L , Q_H , and W all to be positive.)

It turns out that there is a particular type of heat engine, the Carnot engine, which, although not 100% efficient, is more efficient than any other. The grade of heat energy in a system can thus be unambiguously defined in terms of the amount of heat energy in it that cannot be extracted even by a Carnot engine.

How can we build the most efficient possible engine? Let's start with an unnecessarily inefficient engine like a car engine and see how it could be improved. The radiator and exhaust expel hot gases, which is a waste of heat energy. These gases are cooler than the exploded air-gas mixture inside the cylinder, but hotter than the air that surrounds the car. We could thus improve the engine's efficiency by adding an auxiliary heat engine to it, which would operate with the first engine's exhaust as its hot reservoir and the air as its cold reservoir. In general, any heat engine that expels heat at an intermediate temperature can be made more efficient by changing it so that it expels heat only at the temperature of the cold reservoir.

Similarly, any heat engine that absorbs some energy at an intermediate temperature can be made more efficient by adding an auxiliary heat engine to it which will operate between the hot reservoir and this intermediate temperature.

Based on these arguments, we define a Carnot engine as a heat engine that absorbs heat only from the hot reservoir and expels it only into the cold reservoir. Figures d-g show a realization of a Carnot engine using a piston in a cylinder filled with a monoatomic ideal gas. This gas, known as the working fluid, is separate from, but exchanges energy with, the hot and cold reservoirs. This particular Carnot engine has an efficiency given by ([208](#))

$$\text{efficiency} = 1 - \frac{T_L}{T_H}, \quad [\text{efficiency of a Carnot engine}]$$

where T_L is the temperature of the cold reservoir and T_H is the temperature of the hot reservoir.

Even if you do not wish to dig into the details of the proof, the basic reason for the temperature dependence is not so hard to understand. Useful mechanical work is done on strokes d and e, in which the gas expands. The motion of the piston is in the same direction as the gas's force on the piston, so positive work is done on the piston. In strokes f and g, however, the gas does negative work on the piston. We would like to avoid this negative work, but we must design the engine to perform a complete cycle. Luckily the pressures during the compression strokes are lower than the ones during the expansion strokes, so the engine doesn't undo all its work with every cycle. The ratios of the pressures are in proportion to the ratios of the temperatures, so if T_L is 20% of T_H , the engine is 80% efficient.

We have already proved that any engine that is not a Carnot engine is less than optimally efficient, and it is also true that all Carnot engines operating between a given pair of temperatures T_H and T_L have the same efficiency. (This can be proved by the methods of section 9.3.) Thus a Carnot engine is the most efficient possible heat engine.

9.2 Entropy

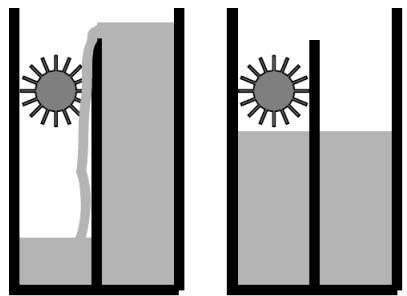
We would like to have some numerical way of measuring the grade of energy in a system. We want this quantity, called entropy, to have the following two properties:

- (1) Entropy is additive. When we combine two systems and consider them as one, the entropy of the combined system equals the sum of the entropies of the two original systems. (Quantities like mass and energy also have this property.)
- (2) The entropy of a system is not changed by operating a Carnot engine within it.

It turns out to be simpler and more useful to define changes in entropy than absolute entropies. Suppose as an example that a system contains some hot matter and some cold matter. It has a relatively high grade of energy because a heat engine could be used to extract mechanical work from it. But if we allow the hot and cold parts to equilibrate at some lukewarm temperature, the grade of energy has gotten worse. Thus putting heat into a hotter area is more useful than putting it into a cold area. Motivated by these considerations, we define a change in entropy as follows:

$$\Delta S = \frac{Q}{T} \quad [\text{change in entropy when adding heat } Q \text{ to matter at temperature } T; \\ \Delta S \text{ is negative if heat is taken out}]$$

A system with a higher grade of energy has a lower entropy.



h / Entropy can be understood using the metaphor of a water wheel. Letting the water levels equalize is like letting the entropy maximize. Taking water from the high side and putting it into the low side increases the entropy. Water levels in this metaphor correspond to temperatures in the actual definition of entropy.

Entropy is additive.

example 1

Since changes in entropy are defined by an additive quantity (heat) divided by a non-additive one (temperature), entropy is additive.

Entropy isn't changed by a Carnot engine.

example 2

The efficiency of a heat engine is defined by

$$\text{efficiency} = 1 - Q_L/Q_H,$$

and the efficiency of a Carnot engine is

$$\text{efficiency} = 1 - T_L/T_H,$$

so for a Carnot engine we have $Q_L/Q_H = T_L/T_H$, which can be rewritten as $Q_L/T_L = Q_H/T_H$. The entropy lost by the hot reservoir is therefore the same as the entropy gained by the cold one.

Entropy increases in heat conduction.

example 3

When a hot object gives up energy to a cold one, conservation of energy tells us that the amount of heat lost by the hot object is the same as the amount of heat gained by the cold one. The change in entropy is $-Q/T_H + Q/T_L$, which is positive because $T_L < T_H$.

Entropy is increased by a non-Carnot engine.

example 4

The efficiency of a non-Carnot engine is less than $1 - T_L/T_H$, so $Q_L/Q_H > T_L/T_H$ and $Q_L/T_L > Q_H/T_H$. This means that the entropy increase in the cold reservoir is greater than the entropy decrease in the hot reservoir.

A book sliding to a stop

example 5

A book slides across a table and comes to a stop. Once it stops, all its kinetic energy has been transformed into heat. As the book and table heat up, their entropies both increase, so the total entropy increases as well.

All of these examples involved closed systems, and in all of them the total entropy either increased or stayed the same. It never decreased. Here are two examples of schemes for decreasing the entropy of a closed system, with explanations of why they don't work.

Using a refrigerator to decrease entropy?

example 6

▷ A refrigerator takes heat from a cold area and dumps it into a hot area. (1) Does this lead to a net decrease in the entropy of a closed system? (2) Could you make a Carnot engine more efficient by running a refrigerator to cool its low-temperature reservoir and eject heat into its high-temperature reservoir?

▷ (1) No. The heat that comes off of the radiator coils is a great deal more than the heat the fridge removes from inside; the difference is what it costs to run your fridge. The heat radiated from the coils is so much more than the heat removed from the inside that the increase in the entropy of the air in the room is greater

than the decrease of the entropy inside the fridge. The most efficient refrigerator is actually a Carnot engine running in reverse, which leads to neither an increase nor a decrease in entropy.

(2) No. The most efficient refrigerator is a reversed Carnot engine. You will not achieve anything by running one Carnot engine in reverse and another forward. They will just cancel each other out.

Maxwell's demon

example 7

▷ Maxwell imagined a pair of rooms, their air being initially in thermal equilibrium, having a partition across the middle with a tiny door. A minuscule demon is posted at the door with a little ping-pong paddle, and his duty is to try to build up faster-moving air molecules in room B and slower moving ones in room A. For instance, when a fast molecule is headed through the door, going from A to B, he lets it by, but when a slower than average molecule tries the same thing, he hits it back into room A. Would this decrease the total entropy of the pair of rooms?

▷ No. The demon needs to eat, and we can think of his body as a little heat engine, and his metabolism is less efficient than a Carnot engine, so he ends up increasing the entropy rather than decreasing it.

Observations such as these lead to the following hypothesis, known as the second law of thermodynamics:

The entropy of a closed system always increases, or at best stays the same: $\Delta S \geq 0$.

Discussion questions

A In this discussion question, you'll think about a car engine in terms of thermodynamics. Note that an internal combustion engine doesn't fit very well into the theoretical straightjacket of a heat engine. For instance, a heat engine has a high-temperature heat reservoir at a single well-defined temperature, T_H . In a typical car engine, however, there are several very different temperatures you could imagine using for T_H : the temperature of the engine block ($\sim 100^\circ\text{C}$), the walls of the cylinder ($\sim 250^\circ\text{C}$), or the temperature of the exploding air-gas mixture ($\sim 1000^\circ\text{C}$, with significant changes over a four-stroke cycle). Let's use $T_H \sim 1000^\circ\text{C}$.

Burning gas supplies heat energy Q_H to your car's engine. The engine does mechanical work W , but also expels heat Q_L into the environment through the radiator and the exhaust. Conservation of energy gives

$$Q_H = Q_L + W,$$

and the relative proportions of Q_L and W are usually about 90% to 10%. (Actually it depends quite a bit on the type of car, the driving conditions, etc.) Here, Q_H , Q_L , and W are all positive according to the sign convention defined on p. 191.

(1) A gallon of gas releases about 140 MJ of heat Q_H when burned. Estimate the change in entropy of the universe due to running a typical car engine and burning one gallon of gas. Note that you'll have to introduce appropriate plus and minus signs, as defined in the relation $\Delta S = Q/T$, in which heat input raises an object's entropy and heat output lowers it. (You'll have to estimate how hot the environment is. For the sake of argument, assume that the work done by the engine, W , remains in the form of mechanical energy, although in reality it probably ends up being changed into heat when you step on the brakes.) Is your result consistent with the second law of thermodynamics?

(2) Q_L is obviously undesirable: you pay for it, but all it does is heat the neighborhood. Suppose that engineers do a really good job of getting rid of the effects that create Q_L , such as friction. Could Q_L ever be reduced to zero, at least theoretically? What would happen if you redid the calculation in #1, but assumed $Q_L = 0$?

B When we run the Carnot engine in figures d-g, there are four parts of the universe that undergo changes in their physical states: the hot reservoir, the cold reservoir, the working gas, and the outside world to which the shaft is connected in order to do physical work. Over one full cycle, discuss which of these parts gain entropy, which ones lose entropy, and which ones keep the same entropy. During which of the four strokes do these changes occur?

9.3 Entropy from a microscopic perspective

The second law of thermodynamics asserts that entropy increases (or stays the same) as we go forward in time. That is, it describes an arrow of time and claims that some processes are not reversible. It doesn't tell us why. It merely asserts a law that seems to hold in experiments. But we have already seen in ch. 7 that there is a more fundamental reason for this, which has to do with statistics and the microscopic scale. People like Carnot didn't realize it, but they were seeing the macroscopic manifestations of these microscopic facts. There we saw that a system tends to move from less probable configurations to more probable ones, i.e., to states in which the number of states is higher. For example, if we see 100 atoms in the right-hand half of the box in figure e, p. 170, and none on the left, we can be sure that when we check on it again at a later time (long enough for equilibrium to happen), we will see very nearly half the atoms on each side. This is because there is only one way of putting all the atoms on the right, $M = 1$, whereas the number of ways of putting them about equally on each side is astronomical, $M \gg 1$ (figure h, p. 172). This leads to the following redefinition of entropy.

Microscopic definition of entropy: The entropy of a system is $S = k \ln M$, where M is the number of available states.

The factor of k is the same Boltzmann constant that we first encountered in ch. 7, and it's there simply for backward compatibility with the previously established macroscopic units of temperature,

energy, and entropy. The point of taking the logarithm of M is that this makes entropy an *additive* quantity. If we defined entropy as the number of possible states M , rather than its log, then entropy would be a multiplicative quantity, not an additive one: if an ice cube in a glass of water has M_1 states available to it, and the number of states available to the water is M_2 , then the number of possible states of the whole system is the product $M_1 M_2$. Logs turn addition into multiplication, so $\ln(M_1 M_2) = \ln M_1 + \ln M_2$, and the entropies of the ice and water simply add.

9.4 Phase space

There is a problem with making this description of entropy into a complete mathematical definition. The problem is that it refers to the number of possible states, but that number may be infinite. To get around the problem, we coarsen our description of the system. For the atoms in our original example of the box with two sides, p. e., we don't really care exactly where each atom is. We only care whether it is in the right side or the left side. If a particular atom's left-right position is described by a coordinate x , then the set of all possible values of x is a line segment along the x axis, containing an infinite number of points. We break this line segment down into two halves, each of width Δx , and we consider two different values of x to be variations on the same state if they both lie in the same half. For our present purposes, we can also ignore completely the y and z coordinates, and all three momentum components, p_x , p_y , and p_z .

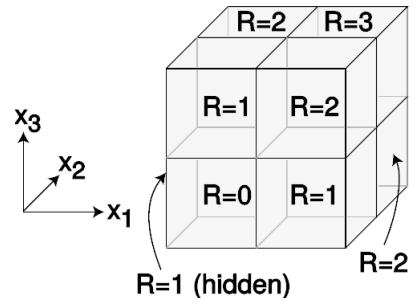
Now let's do a real calculation. Suppose there are only two atoms in the box, with coordinates x_1 and x_2 . We can give all the relevant information about the state of the system by specifying one of the cells in the grid shown in figure i. This grid is known as the *phase space* of the system.¹ The lower right cell, for instance, describes a state in which atom number 1 is in the right side of the box and atom number 2 in the left. Since there are two possible states with $R = 1$ and only one state with $R = 2$, we are twice as likely to observe $R = 1$, and $R = 1$ has higher entropy than $R = 2$.

Figure j shows a corresponding calculation for three atoms, which makes the phase space three-dimensional. Here, the $R = 1$ and 2 states are three times more likely than $R = 0$ and 3. This reproduces the result that we obtained in sec. 7.3, p. 171 simply by making lists of possibilities. Four atoms would require a four-dimensional phase space, which exceeds our ability to visualize. Although our present

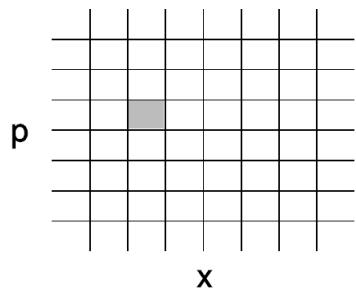
¹The term is a little obscure. Basically the idea is the same as in “my toddler is going through a phase where he always says no.” The “phase” is a stage in the evolution of the system, a snapshot of its state at a moment in time. The usage is also related to the concept of Lissajous figures, in which a particular point on the trajectory is defined by the phases of the oscillations along the x and y axes.

	R=1	R=2
x ₂		
	R=0	R=1
x ₁		

i / The phase space for two atoms in a box.

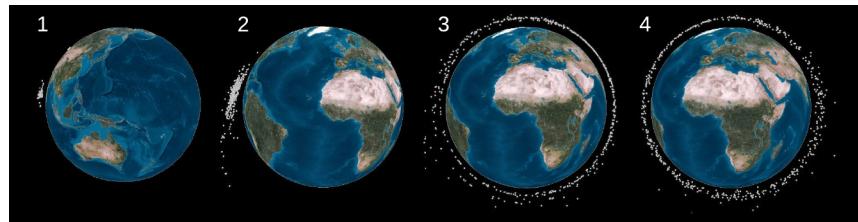


j / The phase space for three atoms in a box.



k / A phase space for a single atom in one dimension, taking momentum into account.

I / Earth orbit is becoming cluttered with space junk, and the pieces can be thought of as the “molecules” comprising an exotic kind of gas. These images show the evolution of a cloud of debris arising from a 2007 Chinese test of an anti-satellite rocket. Panels 1-4 show the cloud five minutes, one hour, one day, and one month after the impact. The entropy seems to have maximized by panel 4.



example doesn’t require it, a phase space can describe momentum as well as position, as shown in figure k. (Re the choice of momentum as the variable, rather than, e.g., velocity, see note [2178](#).) In general, a phase space for a monoatomic gas has six dimensions per atom (one for each coordinate and one for each momentum component). In general, there is one dimension in the phase space per *degree of freedom* (p. 173).

As an example, consider the spreading of the cloud of space debris shown in figure 1. In thermodynamic terms, this is the free expansion of a gas, like opening the valve on an air bottle in the vacuum of outer space. In this analogy, each of the n pieces of debris ($n \sim 2000$) is like one atom. Intuitively, it should be clear that this is an irreversible process, since it seems implausible that the circular spread-out belt of junk in 1/4 would spontaneously gather together into the compressed configuration of figure 1/1, which occupies about 1/10 the volume. Let’s take this circular region of space and break it up into 10 chunks arranged around the circumference of the circle, sort of like the twelve one-hour segments on the dial of an analog clock. Each of these is one cell of phase space.

If we start by considering only a single piece of junk, $n = 1$, then it starts with only one state available to it, $M = 1$, and its entropy is $S = k \ln 1 = 0$. At the end of the process, it has $M = 10$ states available, so $S = k \ln 10$. The entropy is $k \ln(V/V_o)$, where V is the available volume, and V_o is the volume of one cell of our phase space.

For $n = 2$, we can imagine our phase space as a 10×10 grid. The number of states available is $(V/V_o)^2 = 100$. Generalizing to any n , we have an entropy $S = k \ln[(V/V_o)^n]$, which we can rewrite using the rule $\ln(x^y) = y \ln x$ as $S = kn \ln(V/V_o)$. Another way of getting this result would have been simply to use the additivity of entropy: adding up n entropies gives a total entropy that is n times bigger.

Often we don’t really care about an additive constant in our system’s entropy; our macroscopic definition of entropy on p. 201 doesn’t even define what such an additive constant would be. This makes entropy similar to other quantities such as voltage and potential energy. Since $\ln(y/x) = \ln y - \ln x$, we can rewrite the total

entropy of our system as $S = kn \ln V + \text{const}$, and then throw away the constant. Our final result applies not just to this example of the satellite debris but to any gas, if the energy is fixed so that we don't care about the momentum part of the phase space. We have

$$S = kn \ln V$$

for the entropy of a gas consisting of n particles occupying volume V . Note that although it is normally mathematical nonsense to give a transcendental function an input that has units, there is a special dispensation from the pope in the case of logarithms, because the identity $\ln(xy) = \ln x + \ln y$ says that changing units on the input just means adding a constant to the output, which we can ignore if it doesn't affect us. It actually makes sense that we have a usable expression for S that doesn't depend on the volume V_0 , since V_0 is the volume of an entirely fictitious phase cell, which was not even a feature of the macroscopic description of entropy.

9.5 Summary of the laws of thermodynamics

Here is a summary of the laws of thermodynamics:

The zeroth law of thermodynamics (page 189) If object A is at the same temperature as object B, and B is at the same temperature as C, then A is at the same temperature as C.

The first law of thermodynamics Energy is conserved.

The second law of thermodynamics (page 203) The entropy of a closed system always increases, or at best stays the same: $\Delta S \geq 0$.

The third law of thermodynamics The entropy of a system approaches zero as its temperature approaches absolute zero.

From a modern point of view, only the first law deserves to be called a fundamental law of physics. Once Boltzmann discovered the microscopic nature of entropy, the zeroth and second laws could be understood as statements about probability: a system containing a large number of particles is overwhelmingly likely to do a certain thing, simply because the number of possible ways to do it is extremely large compared to the other possibilities. The third law is also now understood to be a consequence of more basic physical principles, but to explain the third law, it's not sufficient simply to know that matter is made of atoms: we also need to understand the quantum-mechanical nature of those atoms, discussed later in this book.

Notes for chapter 9

200 Efficiency of the Carnot engine

The efficiency of a certain Carnot engine is $1 - T_L/T_H$.

This proof is for the Carnot engine described in figures d-g on p. 200, whose working fluid is an ideal gas.

First consider the work done during the constant-temperature strokes. Integrating the equation $dW = P dV$, we have $W = \int P dV$. Since the thermal energy of an ideal gas depends only on its temperature, there is no change in the thermal energy of the gas during this constant-temperature process. Conservation of energy therefore tells us that work done by the gas must be exactly balanced by the amount of heat transferred in from the reservoir.

$$\begin{aligned} Q &= W \\ &= \int P dV \end{aligned}$$

For our proof of the efficiency of the Carnot engine, we need only the ratio of Q_H to Q_L , so we neglect constants of proportionality, and simply substitute $P \propto T/V$, giving

$$Q \propto \int \frac{T}{V} dV \propto T \ln \frac{V_2}{V_1} \propto T \ln \frac{P_1}{P_2}.$$

The efficiency of a heat engine is

$$\text{efficiency} = 1 - \frac{Q_L}{Q_H}.$$

Making use of the result from the previous proof for a Carnot engine with a monoatomic ideal gas as its working gas, we have

$$\text{efficiency} = 1 - \frac{T_L \ln(P_4/P_3)}{T_H \ln(P_1/P_2)},$$

where the subscripts 1, 2, 3, and 4 refer to figures d-g on page 200. We know from problem 4, p. 197, that the temperature is proportional to P^b on the insulated strokes 2-3 and 4-1, the pressures must be related by $P_2/P_3 = P_1/P_4$, which can be rearranged as $P_4/P_3 = P_1/P_2$, and we therefore have

$$\text{efficiency} = 1 - \frac{T_L}{T_H}.$$

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 Even when resting, the human body needs to do a certain amount of mechanical work to keep the heart beating. This quantity is difficult to define and measure with high precision, and also depends on the individual and her level of activity, but it's estimated to be about 1 to 5 watts. Suppose we consider the human body as nothing more than a pump. A person who is just lying in bed all day needs about 1000 kcal/day worth of food to stay alive. (a) Estimate the person's thermodynamic efficiency as a pump, and (b) compare with the maximum possible efficiency imposed by the laws of thermodynamics for a heat engine operating across the difference between a body temperature of 37°C and an ambient temperature of 22°C. (c) Interpret your answer. ▷ Answer, p. 459

2 You use a spoon at room temperature, 22°C, to mix your coffee, which is at 80°C. During this brief period of thermal contact, 1.3 J of heat is transferred from the coffee to the spoon. Find the total change in the entropy of the universe. ✓

3 Object A is a brick. Object B is half of a similar brick. If A is heated, we have $\Delta S = Q/T$. Show that if this equation is valid for A, then it is also valid for B. ▷ Solution, p. 447

4 Refrigerators, air conditioners, and heat pumps are heat engines that work in reverse. You put in mechanical work, and the effect is to take heat out of a cooler reservoir and deposit heat in a warmer one: $Q_L + W = Q_H$. As with the heat engines discussed previously, the efficiency is defined as the energy transfer you want (Q_L for a refrigerator or air conditioner, Q_H for a heat pump) divided by the energy transfer you pay for (W).

Efficiencies are supposed to be unitless, but the efficiency of an air conditioner is normally given in terms of an EER rating (or a more complex version called an SEER). The EER is defined as Q_L/W , but expressed in the barbaric units of Btu/watt-hour. A typical EER rating for a residential air conditioner is about 10 Btu/watt-hour, corresponding to an efficiency of about 3. The standard temperatures used for testing an air conditioner's efficiency are 80°F (27°C) inside and 95°F (35°C) outside.

- (a) What would be the EER rating of a reversed Carnot engine used as an air conditioner? ✓
- (b) If you ran a 3-kW residential air conditioner, with an efficiency of 3, for one hour, what would be the effect on the total entropy of the universe? Is your answer consistent with the second law of thermodynamics? ✓

Optics



Chapter 10

Images, qualitatively

10.1 Vision and the nature of light

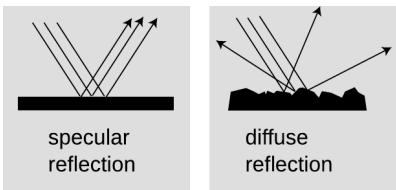
Our eyes are light sensors. When we look at a luminous object such as a candle flame or a cell phone's screen, light comes from the object to us. When we look at a nonluminous object, we see it because it reflects light, which then reaches our eye. Either way, you can't see anything unless light goes in your eye.

Many people might disagree if you told them that light was reflected from the book to the eye, because they think of reflection as something that mirrors do, not something that a book does. They associate reflection with the formation of a reflected image, which does not seem to appear in a piece of paper.

Imagine that you are looking at your reflection in a nice smooth piece of aluminum foil, fresh off the roll. You perceive a face, not a piece of metal. Perhaps you also see the bright reflection of a lamp over your shoulder behind you. Now imagine that the foil is just a little bit less smooth. The different parts of the image are now



a / Two self-portraits of the author, one taken in a mirror and one with a piece of aluminum foil.



b / Specular and diffuse reflection.

a little bit out of alignment with each other. Your brain can still recognize a face and a lamp, but it's a little scrambled, like a Picasso painting. Now suppose you use a piece of aluminum foil that has been crumpled up and then flattened out again. The parts of the image are so scrambled that you cannot recognize an image. Instead, your brain tells you you're looking at a rough, silvery surface.

Mirror-like reflection at a specific angle is known as specular reflection, and random reflection in many directions is called diffuse reflection. Diffuse reflection is how we see nonluminous objects. Specular reflection only allows us to see images of objects other than the one doing the reflecting. In top part of figure a, imagine that the rays of light are coming from the sun. If you are looking down at the reflecting surface, there is no way for your eye-brain system to tell that the rays are not really coming from a sun down below you.

The differences among white, black, and the various shades of gray in between is a matter of what percentage of the light they absorb and what percentage they reflect. That's why light-colored clothing is more comfortable in the summer, and light-colored upholstery in a car stays cooler than dark upholstery.

10.2 The ray model of light

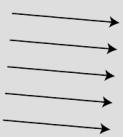
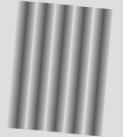
10.2.1 Models of light

Note how I've been casually diagramming the motion of light with pictures showing light rays as lines on the page. More formally, this is known as the ray model of light. The ray model of light seems natural once we convince ourselves that light travels through space, and observe phenomena like sunbeams coming through holes in clouds. Having already been introduced to the concept of light as an electromagnetic wave, you know that the ray model is not the ultimate truth about light, but the ray model is simpler, and in any case science always deals with models of reality, not the ultimate nature of reality. Table c summarizes three models of light.

The ray model is a generic one. By using it we can discuss the path taken by the light, without committing ourselves to any specific description of what it is that is moving along that path. We will use the nice simple ray model for most of our treatment of optics, and with it we can analyze a great many devices and phenomena. Not until ch. 12 will we concern ourselves specifically with wave optics, although in the intervening chapters I will sometimes analyze the same phenomenon using both the ray model and the wave model.

Note that the statements about the applicability of the various models are only rough guides. For instance, wave interference effects

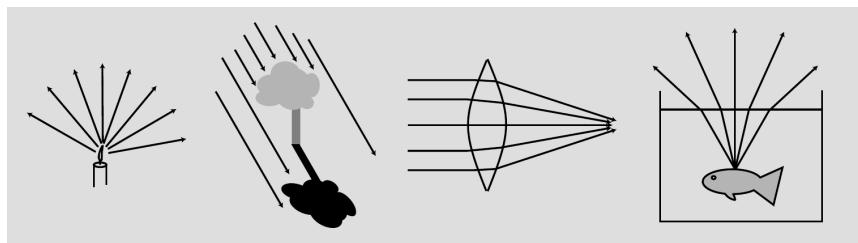
c / Three models of light.

ray model		Advantage: Simplicity.
wave model		Advantage: Color is described naturally in terms of wavelength. Required in order to explain the interaction of light with material objects of sizes comparable to or smaller than a wavelength of light.
particle model		Required in order to explain the interaction of light with individual atoms. At the atomic level, it becomes apparent that a beam of light has a certain graininess to it.

are often detectable, if small, when light passes around an obstacle that is quite a bit bigger than a wavelength. Also, the criterion for when we need the particle model really has more to do with energy scales than distance scales, although the two turn out to be related.

The alert reader may have noticed that the wave model is required at scales smaller than a wavelength of light (on the order of a micrometer for visible light), and the particle model is demanded on the atomic scale or lower (a typical atom being a nanometer or so in size). This implies that at the smallest scales we need *both* the wave model and the particle model. They appear incompatible, so how can we simultaneously use both? The answer is that they are not as incompatible as they seem. Light is both a wave and a particle, but a full understanding of this apparently nonsensical statement is a topic for ch. 15.

d / Examples of ray diagrams.



10.2.2 Ray diagrams

Without even knowing how to use the ray model to calculate anything numerically, we can learn a great deal by drawing ray diagrams. For instance, if you want to understand how eyeglasses help you to see in focus, a ray diagram is the right place to start. Many students under-utilize ray diagrams in optics and instead rely on rote memorization or plugging into formulas. The trouble with memorization and plug-ins is that they can obscure what's really going on, and it is easy to get them wrong. Often the best plan is to do a ray diagram first, then do a numerical calculation, then check that your numerical results are in reasonable agreement with what you expected from the ray diagram.

e / 1. Correct. 2. Incorrect: implies that diffuse reflection only gives one ray from each reflecting point. 3. Correct, but unnecessarily complicated

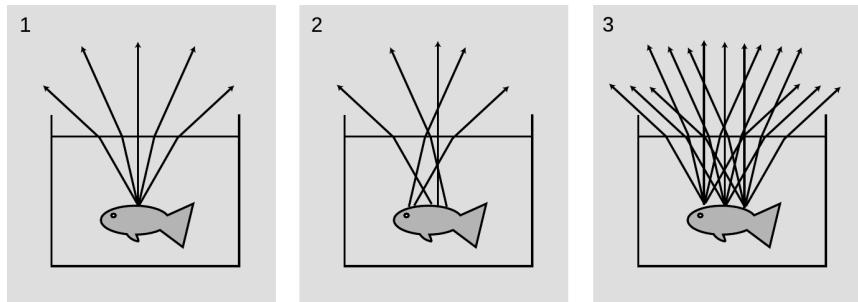


Figure e shows some guidelines for using ray diagrams effectively. The light rays bend when they pass out through the surface of the water (a phenomenon that we'll discuss in more detail later). The rays appear to have come from a point above the goldfish's actual location, an effect that is familiar to people who have tried spear-fishing.

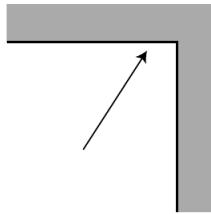
- A stream of light is not really confined to a finite number of narrow lines. We just draw it that way. In e/1, it has been necessary to choose a finite number of rays to draw (five), rather than the theoretically infinite number of rays that will diverge from that point.
- There is a tendency to conceptualize rays incorrectly as objects. In his Optics, Newton goes out of his way to caution

the reader against this, saying that some people “consider ... the refraction of ... rays to be the bending or breaking of them in their passing out of one medium into another.” But a ray is a record of the path traveled by light, not a physical thing that can be bent or broken.

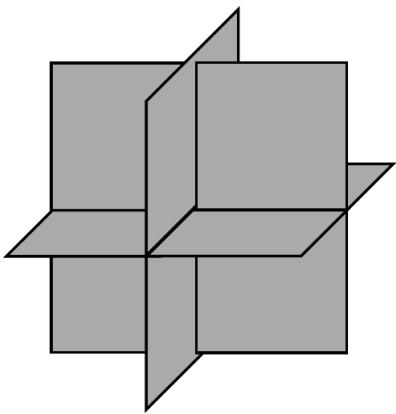
- In theory, rays may continue infinitely far into the past and future, but we need to draw lines of finite length. In e/1, a judicious choice has been made as to where to begin and end the rays. There is no point in continuing the rays any farther than shown, because nothing new and exciting is going to happen to them. There is also no good reason to start them earlier, before being reflected by the fish, because the direction of the diffusely reflected rays is random anyway, and unrelated to the direction of the original, incoming ray.
- When representing diffuse reflection in a ray diagram, many students have a mental block against drawing many rays fanning out from the same point. Often, as in example e/2, the problem is the misconception that light can only be reflected in one direction from one point.
- Another difficulty associated with diffuse reflection, example e/3, is the tendency to think that in addition to drawing many rays coming out of one point, we should also be drawing many rays coming from many points. In e/1, drawing many rays coming out of one point gives useful information, telling us, for instance, that the fish can be seen from any angle. Drawing many sets of rays, as in e/3, does not give us any more useful information, and just clutters up the picture in this example. The only reason to draw sets of rays fanning out from more than one point would be if different things were happening to the different sets.

Discussion question

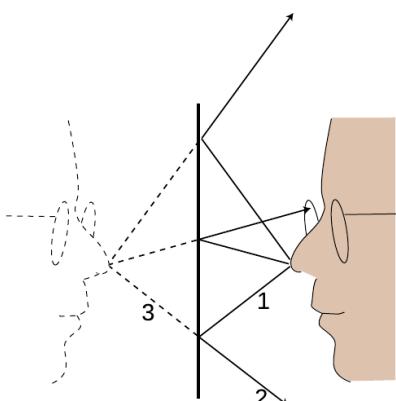
A Suppose an intelligent tool-using fish is spear-hunting for humans. Draw a ray diagram to show how the fish has to correct its aim. Note that although the rays are now passing from the air to the water, the same rules apply (sec. 5.6, p. 120): the rays are closer to being perpendicular to the surface when they are in the water, and rays that hit the air-water interface at a shallow angle are bent the most.



f / Discussion question C.



g / Discussion question D.



h / An image formed by a mirror.

B If a light ray has a velocity vector with components c_x and c_y , what will happen when it is reflected from a surface that lies along the y axis? Make sure your answer does not imply a change in the ray's speed.

C Generalizing your reasoning from discussion question B, what will happen to the velocity components of a light ray that hits a corner, as shown in the figure, and undergoes two reflections?

D Three pieces of sheet metal arranged perpendicularly as shown in the figure form what is known as a radar corner. Let's assume that the radar corner is large compared to the wavelength of the radar waves, so that the ray model makes sense. If the radar corner is bathed in radar rays, at least some of them will undergo three reflections. Making a further generalization of your reasoning from the two preceding discussion questions, what will happen to the three velocity components of such a ray? What would the radar corner be useful for?

10.3 A virtual image

Infants are always fascinated by the antics of the Baby in the Mirror. Now if you want to know something about mirror images that most people don't understand, try this. First bring this page closer and closer to your eyes, until you can no longer focus on it without straining. Then go in the bathroom and see how close you can get your face to the surface of the mirror before you can no longer easily focus on the image of your own eyes. You will find that the shortest comfortable eye-mirror distance is much less than the shortest comfortable eye-paper distance. This demonstrates that the image of your face in the mirror acts as if it had depth and existed in the space *behind* the mirror. If the image was like a flat picture in a book, then you wouldn't be able to focus on it from such a short distance.

We can understand a mirror image using a ray diagram. Figure h shows several light rays, 1, that originated by diffuse reflection at the person's nose. They bounce off the mirror, producing new rays, 2. To anyone whose eye is in the right position to get one of these rays, they appear to have come from a behind the mirror, 3, where they would have originated from a single point. This point is where the tip of the image-person's nose appears to be. A similar analysis applies to every other point on the person's face, so it looks as though there was an entire face behind the mirror. The customary way of describing the situation requires some explanation:

Customary description in physics: There is an image of the face behind the mirror.

Translation: The pattern of rays coming from the mirror is exactly the same as it would be if there were a face behind the mirror. Nothing is really behind the mirror.

This is referred to as a *virtual* image, because the rays do not

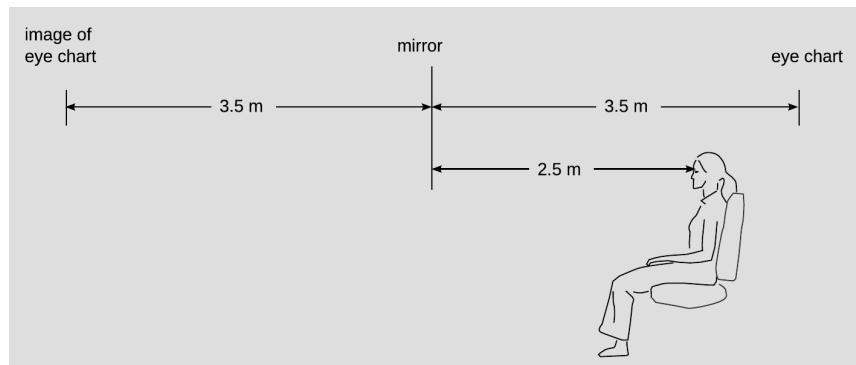
actually cross at the point behind the mirror. They only appear to have originated there.

self-check A

Imagine that the person in figure h moves his face down quite a bit — a couple of feet in real life, or a few inches on this scale drawing. The mirror stays where it is. Draw a new ray diagram. Will there still be an image? If so, where is it visible from?

▷ Answer, p. 456

The geometry of specular reflection tells us that rays 1 and 2 are at equal angles to the normal (the imaginary perpendicular line piercing the mirror at the point of reflection). This means that ray 2's imaginary continuation, 3, forms the same angle with the mirror as ray 1. Since each ray of type 3 forms the same angles with the mirror as its partner of type 1, we see that the distance of the image from the mirror is the same as that of the actual face from the mirror, and it lies directly across from it. The image therefore appears to be the same size as the actual face.



i / Example 1.

An eye exam

example 1

Figure i shows a typical setup in an optometrist's examination room. The patient's vision is supposed to be tested at a distance of 6 meters (20 feet in the U.S.), but this distance is larger than the amount of space available in the room. Therefore a mirror is used to create an image of the eye chart behind the wall.

The Praxinoscope

example 2

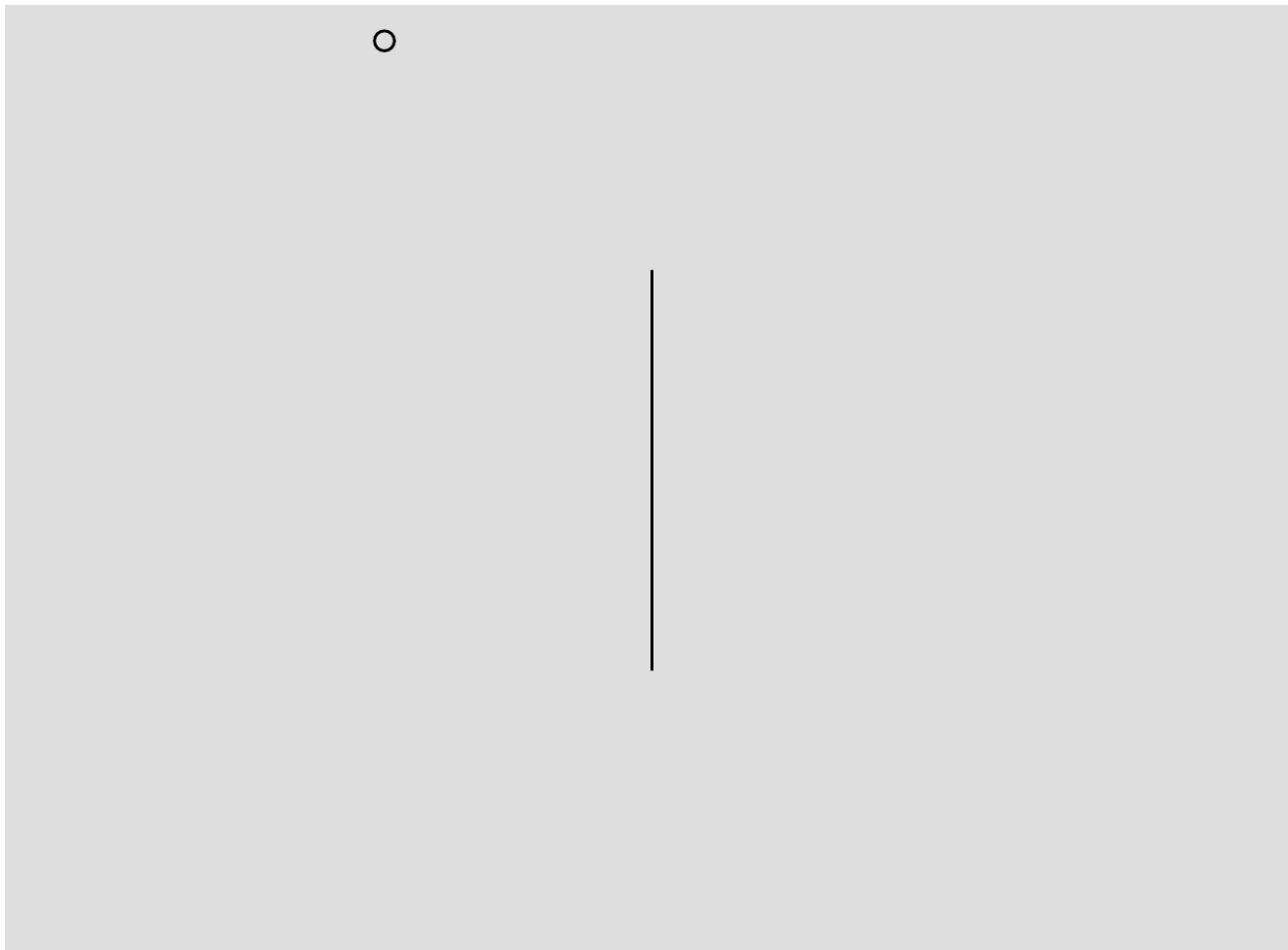
Figure j shows an old-fashioned device called a praxinoscope, which displays an animated picture when spun. The removable strip of paper with the pictures printed on it has twice the radius of the inner circle made of flat mirrors, so each picture's virtual image is at the center. As the wheel spins, each picture's image is replaced by the next.



j / The praxinoscope.

Discussion question

A The figure shows an object that is off to one side of a mirror. Draw a ray diagram. Is an image formed? If so, where is it, and from which directions would it be visible?

A diagram consisting of a light gray rectangular background. On the left side, there is a small black circle labeled 'O'. In the center-right area, there is a single vertical black line.

O

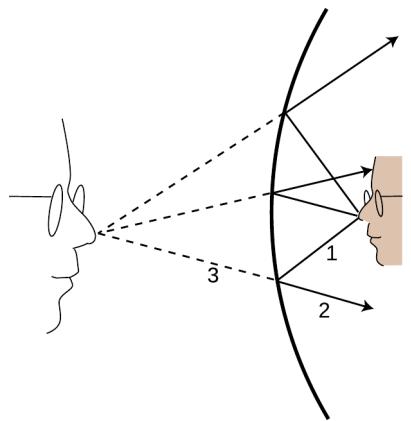
10.4 Curved mirrors

An image in a flat mirror is a pretechnological example: even animals can look at their reflections in a calm pond. We now pass to our first nontrivial example of the manipulation of an image by technology: an image in a curved mirror. Before we dive in, let's consider why this is an important example. If it was just a question of memorizing a bunch of facts about curved mirrors, then you would rightly rebel against an effort to spoil the beauty of your liberally educated brain by force-feeding you technological trivia. The reason this is an important example is not that curved mirrors are so important in and of themselves, but that the results we derive for curved bowl-shaped mirrors turn out to be true for a large class of other optical devices, including mirrors that bulge outward rather than inward, and lenses as well. A microscope or a telescope is simply a combination of lenses or mirrors or both. What you're really learning about here is the basic building block of all optical devices from movie projectors to octopus eyes.

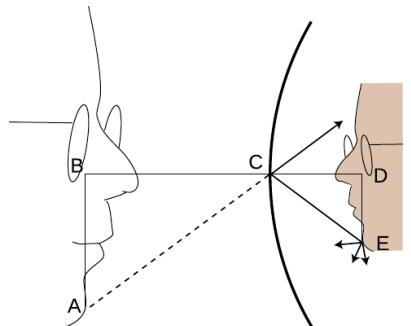
Because the mirror in figure k is curved, it bends the rays back closer together than a flat mirror would: we describe it as *converging*. Note that the term refers to what it does to the light rays, not to the physical shape of the mirror's surface. (The surface itself would be described as *concave*. The term is not all that hard to remember, because the hollowed-out interior of the mirror is like a cave.) It is surprising but true that all the rays like 3 really do converge on a point, forming a good image. We will not prove this fact, but it is true for any mirror whose curvature is gentle enough and that is symmetric with respect to rotation about the perpendicular line passing through its center (not asymmetric like a potato chip). The old-fashioned method of making mirrors and lenses is by grinding them in grit by hand, and this automatically tends to produce an almost perfect spherical surface.

Bending a ray like 2 inward implies bending its imaginary continuation 3 outward, in the same way that raising one end of a seesaw causes the other end to go down. The image therefore forms deeper behind the mirror. This doesn't just show that there is extra distance between the image-nose and the mirror; it also implies that the image itself is bigger from front to back. It has been *magnified* in the front-to-back direction.

It is easy to prove that the same magnification also applies to the image's other dimensions. Consider a point like E in figure l. The trick is that out of all the rays diffusely reflected by E, we pick the one that happens to head for the mirror's center, C. The equal-angle property of specular reflection plus a little straightforward geometry easily leads us to the conclusion that triangles ABC and CDE are the same shape, with ABC being simply a scaled-up version of CDE. The magnification of depth equals the ratio BC/CD, and the up-



k / An image formed by a curved mirror.



l / The image is magnified by the same factor in depth and in its other dimensions.



m / Increased magnification always comes at the expense of decreased field of view.

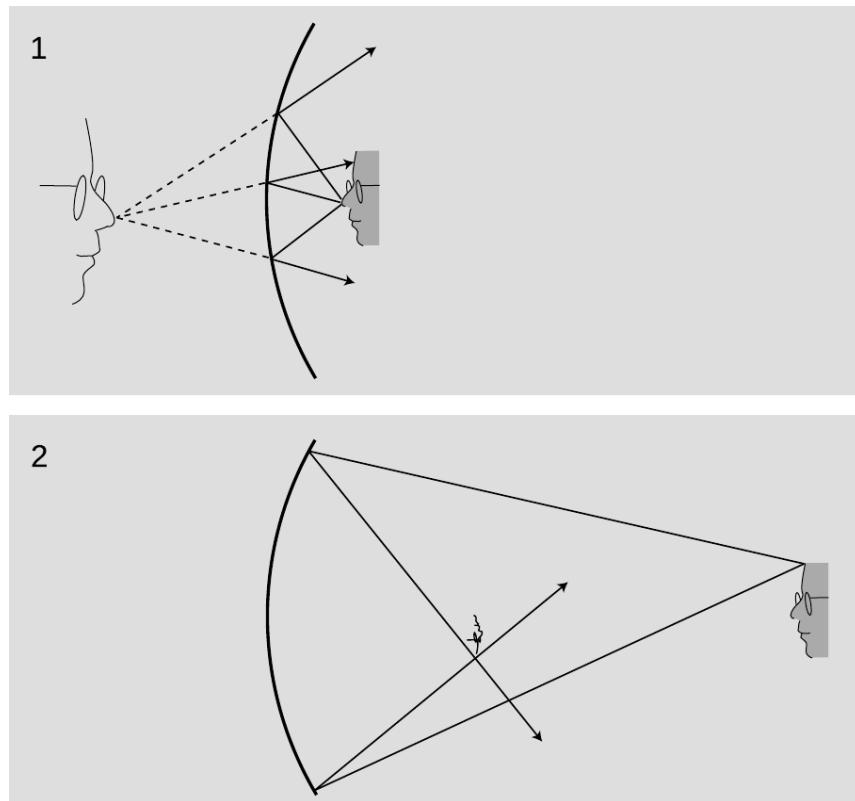
down magnification is AB/DE . A repetition of the same proof shows that the magnification in the third dimension (out of the page) is also the same. This means that the image-head is simply a larger version of the real one, without any distortion. The scaling factor is called the magnification, M . The image in the figure is magnified by a factor $M = 1.9$.

Note that we did not explicitly specify whether the mirror was a sphere, a paraboloid, or some other shape. However, we assumed that a focused image would be formed, which would not necessarily be true, for instance, for a mirror that was asymmetric or very deeply curved.

10.5 A real image

If we start by placing an object very close to the mirror, $n/1$, and then move it farther and farther away, the image at first behaves as we would expect from our everyday experience with flat mirrors, receding deeper and deeper behind the mirror. At a certain point, however, a dramatic change occurs. When the object is more than a certain distance from the mirror, $n/2$, the image appears upside-down and in *front* of the mirror.

$n/1$. A virtual image. 2. A real image. As you'll verify in homework problem 8, the image is upside-down



Here's what's happened. The mirror bends light rays inward, but

when the object is very close to it, as in $n/1$, the rays coming from a given point on the object are too strongly diverging (spreading) for the mirror to bring them back together. On reflection, the rays are still diverging, just not as strongly diverging. But when the object is sufficiently far away, $n/2$, the mirror is only intercepting the rays that came out in a narrow cone, and it is able to bend these enough so that they will reconverge.

Note that the rays shown in the figure, which both originated at the same point on the object, reunite when they cross. The point where they cross is the image of the point on the original object. This type of image is called a *real image*, in contradistinction to the virtual images we've studied before.

Definition: A real image is one where rays actually cross. A virtual image is a point from which rays only appear to have come.

The use of the word "real" is perhaps unfortunate. It sounds as though we are saying the image was an actual material object, which of course it is not.

The distinction between a real image and a virtual image is an important one, because a real image can be projected onto a screen or photographic film. If a piece of paper is inserted in figure $n/2$ at the location of the image, the image will be visible on the paper (provided the object is bright and the room is dark). Your eye uses a lens to make a real image on the retina.

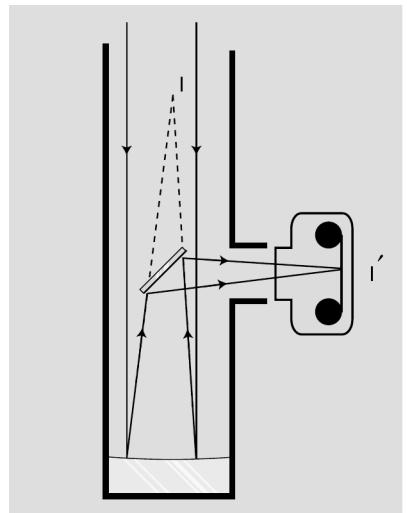
self-check B

Sketch another copy of the face in figure $n/1$, even farther from the mirror, and draw a ray diagram. What has happened to the location of the image? ▷ Answer, p. 456

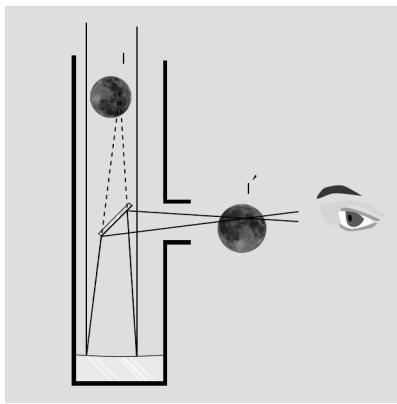
10.6 Images of images

If you are wearing glasses right now, then the light rays from the page are being manipulated first by your glasses and then by the lens of your eye. You might think that it would be extremely difficult to analyze this, but in fact it is quite easy. In any series of optical elements (mirrors or lenses or both), each element works on the rays furnished by the previous element in exactly the same manner as if the image formed by the previous element was an actual object.

Figure o shows an example involving only mirrors. The Newtonian telescope, invented by Isaac Newton, consists of a large curved mirror, plus a second, flat mirror that brings the light out of the tube. (In very large telescopes, there may be enough room to put a camera or even a person inside the tube, in which case the second mirror is not needed.) The tube of the telescope is not vital; it



o / A Newtonian telescope being used with a camera.

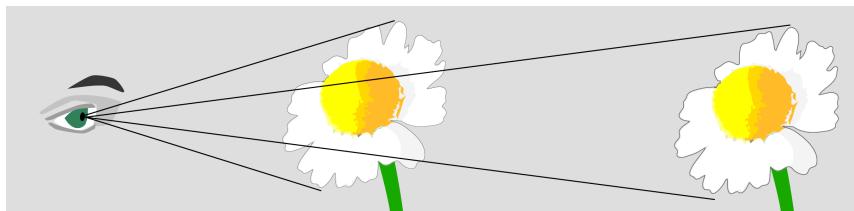


p / A Newtonian telescope being used for visual rather than photographic observing. In real life, an eyepiece lens is normally used for additional magnification, but this simpler setup will also work.

is mainly a structural element, although it can also be helpful for blocking out stray light. The lens has been removed from the front of the camera body, and is not needed for this setup. Note that the two sample rays have been drawn parallel, because an astronomical telescope is used for viewing objects that are extremely far away. These two “parallel” lines actually meet at a certain point, say a crater on the moon, so they can’t actually be perfectly parallel, but they are parallel for all practical purposes since we would have to follow them upward for a quarter of a million miles to get to the point where they intersect.

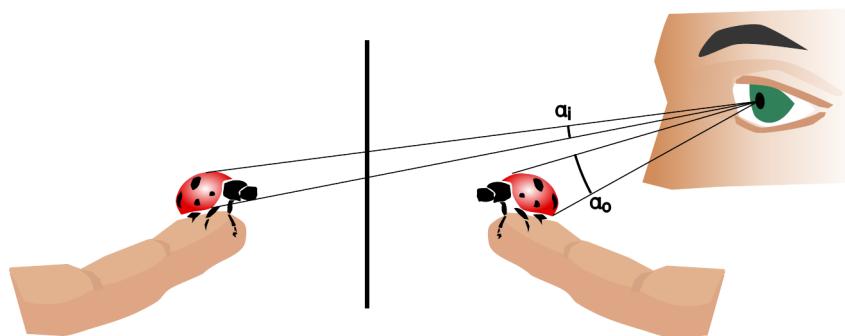
The large curved mirror by itself would form an image I , but the small flat mirror creates an image of the image, I' . The relationship between I and I' is exactly the same as it would be if I was an actual object rather than an image: I and I' are at equal distances from the plane of the mirror, and the line between them is perpendicular to the plane of the mirror.

One surprising wrinkle is that whereas a flat mirror used by itself forms a virtual image of an object that is real, here the mirror is forming a real image of virtual image I . This shows how pointless it would be to try to memorize lists of facts about what kinds of images are formed by various optical elements under various circumstances. You are better off simply drawing a ray diagram.



q / The angular size of the flower depends on its distance from the eye.

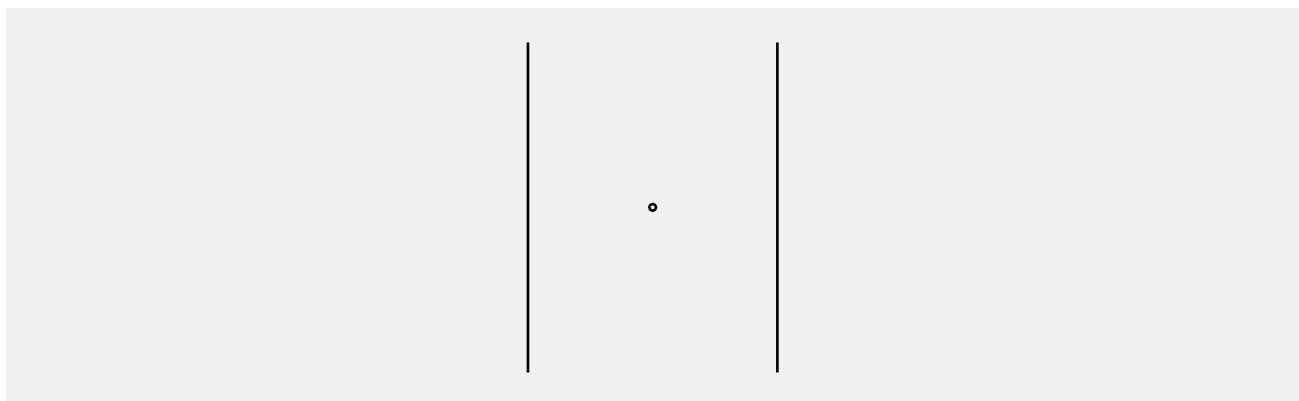
Although the main point here was to give an example of an image of an image, figure p also shows an interesting case where we need to make the distinction between *magnification* and *angular magnification*. If you are looking at the moon through this telescope, then the images I and I' are much *smaller* than the actual moon. Otherwise, for example, image I would not fit inside the telescope! However, these images are very close to your eye compared to the actual moon. The small size of the image has been more than compensated for by the shorter distance. The important thing here is the amount of *angle* within your field of view that the image covers, and it is this angle that has been increased. The factor by which it is increased is called the *angular magnification*, M_a .



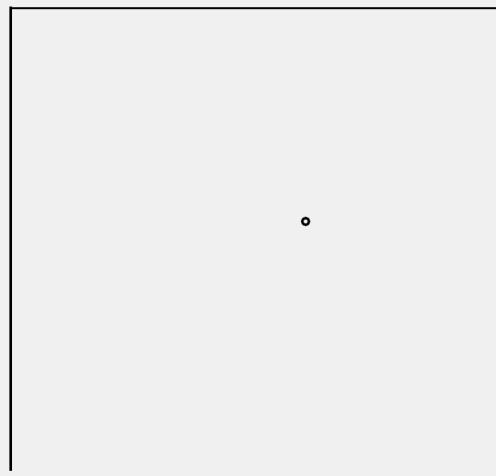
r / The person uses a mirror to get a view of both sides of the ladybug. Although the flat mirror has $M = 1$, it doesn't give an angular magnification of 1. The image is farther from the eye than the object, so the angular magnification $M_a = \alpha_i / \alpha_o$ is less than one.

Discussion questions

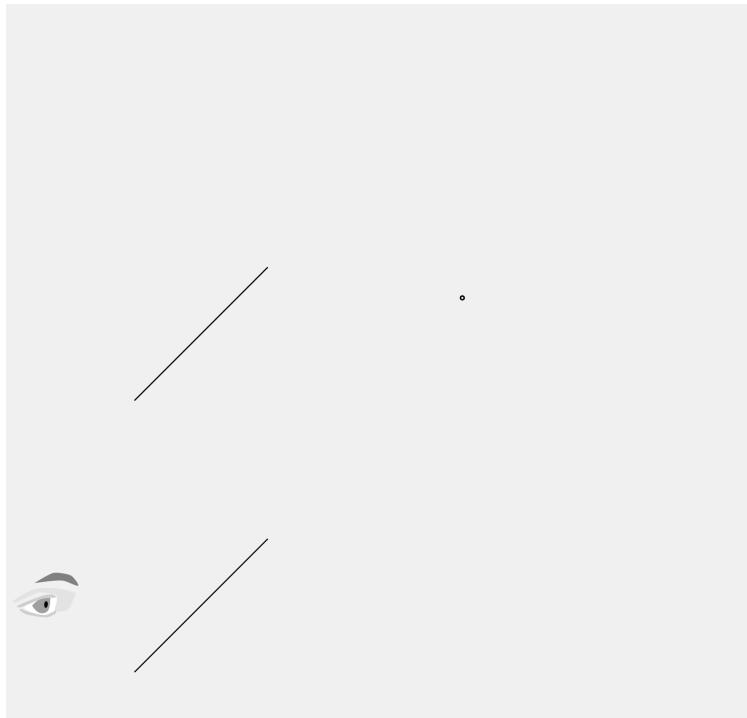
- A** Locate the images of you that will be formed if you stand between two parallel mirrors.



B Locate the images formed by two perpendicular mirrors, as in the figure. What happens if the mirrors are not perfectly perpendicular?



C Locate the images formed by the periscope.



Problems

Key

- ✓ A computerized answer check is available online.
- ★ A difficult problem.

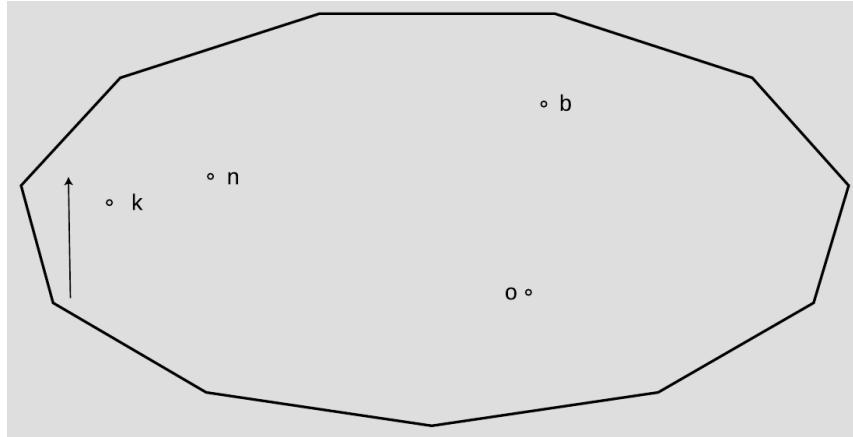
1 Draw a ray diagram showing why a small light source (a candle, say) produces sharper shadows than a large one (e.g., a long fluorescent bulb).

2 The Stealth Bomber is designed with flat, smooth surfaces. Why would this make it difficult to detect using radar?

▷ Solution, p. 447

3 The natives of planet Wumpus play pool using light rays on an eleven-sided table with mirrors for bumpers, shown in the figure. Trace this shot accurately with a ruler to reveal the hidden message. To get good enough accuracy, you'll need to photocopy the page (or download the book and print the page) and construct each reflection using a protractor.

▷ Solution, p. 447



Problem 3.

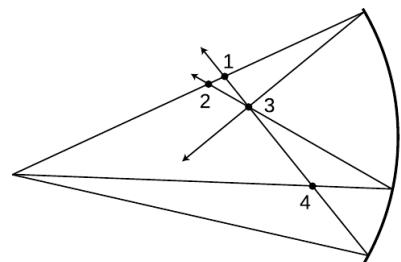
4 A man is walking at 1.0 m/s directly towards a flat mirror. At what speed is his separation from his image decreasing? ✓

5 If a mirror on a wall is only big enough for you to see yourself from your head down to your waist, can you see your entire body by backing up? Test this experimentally and come up with an explanation for your observations, including a ray diagram.

Note that when you do the experiment, it's easy to confuse yourself if the mirror is even a tiny bit off of vertical. One way to check yourself is to artificially lower the top of the mirror by putting a piece of tape or a post-it note where it blocks your view of the top of your head. You can then check whether you are able to see more of yourself both above *and* below by backing up.

6 The figure shows four points where rays cross. Of these, which are image points? Explain.

7 In this chapter we've only done examples of mirrors with hollowed-out shapes (called concave mirrors). Now draw a ray diagram for a curved mirror that has a bulging outward shape (called a convex mirror). (a) How does the image's distance from the mirror compare with the actual object's distance from the mirror? From this comparison, determine whether the magnification is greater than or less than one. (b) Is the image real, or virtual? Could this mirror ever make the other type of image?



Problem 6.

8 In figure n/2 in on page 222, only the image of my forehead was located by drawing rays. Either photocopy the figure or download the book and print out the relevant page. On this copy of the figure, make a new set of rays coming from my chin, and locate its image. To make it easier to judge the angles accurately, draw rays from the chin that happen to hit the mirror at the same points where the two rays from the forehead were shown hitting it. By comparing the locations of the chin's image and the forehead's image, verify that the image is actually upside-down, as shown in the original figure.

Exercise 10: Exploring images with a curved mirror

Equipment:

- concave mirrors with deep curvature
- concave mirrors with gentle curvature
- convex mirrors

1. Obtain a curved mirror from your instructor. If it is silvered on both sides, make sure you're working with the concave side, which bends light rays inward. Look at your own face in the mirror. Now change the distance between your face and the mirror, and see what happens. Explore the full range of possible distances between your face and the mirror.

In these observations you've been changing two variables at once: the distance between the object (your face) and the mirror, and the distance from the mirror to your eye. In general, scientific experiments become easier to interpret if we practice isolation of variables, i.e., only change one variable while keeping all the others constant. In parts 2 and 3 you'll form an image of an object that's not your face, so that you can have independent control of the object distance and the point of view.

2. With the mirror held far away from you, observe the image of something behind you, over your shoulder. Now bring your eye closer and closer to the mirror. Can you see the image with your eye very close to the mirror? See if you can explain your observation by drawing a ray diagram.

—————> turn page

3. Now imagine the following new situation, but *don't actually do it yet*. Suppose you lay the mirror face-up on a piece of tissue paper, put your finger a few cm above the mirror, and look at the image of your finger. As in part 2, you can bring your eye closer and closer to the mirror.

Will you be able to see the image with your eye very close to the mirror? Draw a ray diagram to help you predict what you will observe.

Prediction: _____

Now test your prediction. If your prediction was incorrect, see if you can figure out what went wrong, or ask your instructor for help.

4. For parts 4 and 5, it's more convenient to use concave mirrors that are more gently curved; obtain one from your instructor. Lay the mirror on the tissue paper, and use it to create an image of the overhead lights on a piece of paper above it and a little off to the side. What do you have to do in order to make the image clear? Can you explain this observation using a ray diagram?

—————> *turn page*

5. Now imagine the following experiment, but *don't do it yet*. What will happen to the image on the paper if you cover half of the mirror with your hand?

Prediction:-----

Test your prediction. If your prediction was incorrect, can you explain what happened?

6. Now imagine forming an image with a convex mirror (one that bulges outward), and that therefore bends light rays away from the central axis (i.e., is diverging). Draw a typical ray diagram.

Is the image real, or virtual? Will there be more than one type of image?

Prediction:-----

Test your prediction.

Chapter 11

Images, quantitatively

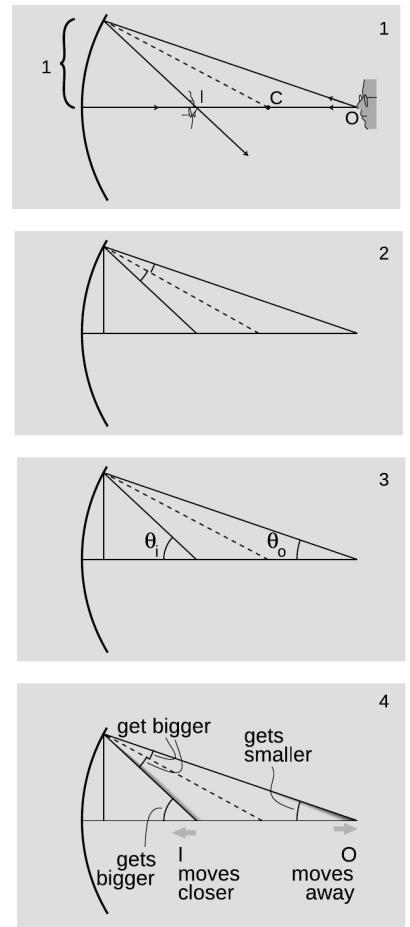
It sounds a bit odd when a scientist refers to a theory as “beautiful,” but to those in the know it makes perfect sense. One mark of a beautiful theory is that it surprises us by being simple. The mathematical theory of lenses and curved mirrors gives us just such a surprise. We expect the subject to be complex because there are so many cases: a converging mirror forming a real image, a diverging lens that makes a virtual image, and so on for a total of six possibilities. If we want to predict the location of the images in all these situations, we might expect to need six different equations, and six more for predicting magnifications. Instead, it turns out that we can use just one equation for the location of the image and one equation for its magnification, and these two equations work in all the different cases with no changes except for plus and minus signs. This is the kind of thing the physicist Eugene Wigner referred to as “the unreasonable effectiveness of mathematics.” Sometimes we can find a deeper reason for this kind of unexpected simplicity, but sometimes it almost seems as if God went out of Her way to make the secrets of universe susceptible to attack by the human thought-tool called math.

11.1 A real image formed by a converging mirror

11.1.1 Location of the image

We will now derive the equation for the location of a real image formed by a converging mirror. We assume for simplicity that the mirror is spherical, but actually this isn’t a restrictive assumption, because any shallow, symmetric curve can be approximated by a sphere. The shape of the mirror can be specified by giving the location of its center, C. A deeply curved mirror is a sphere with a small radius, so C is close to it, while a weakly curved mirror has C farther away. Given the point O where the object is, we wish to find the point I where the image will be formed.

To locate an image, we need to track a minimum of two rays coming from the same point. Since we have proved in the previous chapter that this type of image is not distorted, we can use an on-axis point, O, on the object, as in figure a/1. The results we derive will also hold for off-axis points, since otherwise the image would have



a / The relationship between the object’s position and the image’s can be expressed in terms of the angles θ_o and θ_i .

to be distorted, which we know is not true. We let one of the rays be the one that is emitted along the axis; this ray is especially easy to trace, because it bounces straight back along the axis again. As our second ray, we choose one that strikes the mirror at a distance of 1 from the axis. “One what?” asks the astute reader. The answer is that it doesn’t really matter. When a mirror has shallow curvature, all the reflected rays hit the same point, so 1 could be expressed in any units you like. It could, for instance, be 1 cm, unless your mirror is smaller than 1 cm!

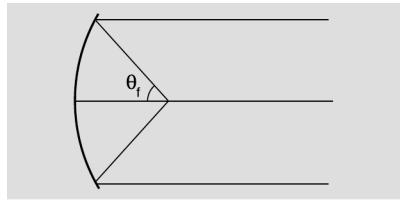
The only way to find out anything mathematical about the rays is to use the sole mathematical fact we possess concerning specular reflection: the incident and reflected rays form equal angles with respect to the normal, which is shown as a dashed line. Therefore the two angles shown in figure a/2 are the same, and skipping some straightforward geometry, this leads to the visually reasonable result that the two angles in figure a/3 are related as follows:

$$\theta_i + \theta_o = \text{constant}$$

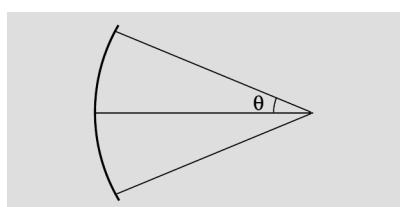
(Note that θ_i and θ_o , which are measured from the image and the object, not from the eye like the angles we referred to in discussing angular magnification on page 224.) For example, move O farther from the mirror. The top angle in figure a/2 is increased, so the bottom angle must increase by the same amount, causing the image point, I, to move closer to the mirror. In terms of the angles shown in figure a/3, the more distant object has resulted in a smaller angle θ_o , while the closer image corresponds to a larger θ_i ; One angle increases by the same amount that the other decreases, so their sum remains constant. These changes are summarized in figure a/4.

The sum $\theta_i + \theta_o$ is a constant. What does this constant represent? Geometrically, we interpret it as double the angle made by the dashed radius line. Optically, it is a measure of the strength of the mirror, i.e., how strongly the mirror focuses light, and so we call it the focal angle, θ_f ,

$$\theta_i + \theta_o = \theta_f.$$



b / The geometrical interpretation of the focal angle.



c / Example 1, an alternative test for finding the focal angle. The mirror is the same as in figure b.

Suppose, for example, that we wish to use a quick and dirty optical test to determine how strong a particular mirror is. We can lay it on the floor as shown in figure c, and use it to make an image of a lamp mounted on the ceiling overhead, which we assume is very far away compared to the radius of curvature of the mirror, so that the mirror intercepts only a very narrow cone of rays from the lamp. This cone is so narrow that its rays are nearly parallel, and θ_o is nearly zero. The real image can be observed on a piece of paper. By moving the paper nearer and farther, we can bring the image into focus, at which point we know the paper is located at the image point. Since $\theta_o \approx 0$, we have $\theta_i \approx \theta_f$, and we can then determine this mirror’s focal angle either by measuring θ_i directly

with a protractor, or indirectly via trigonometry. A strong mirror will bring the rays together to form an image close to the mirror, and these rays will form a blunt-angled cone with a large θ_i and θ_f .

An alternative optical test

example 1

▷ Figure c shows an alternative optical test. Rather than placing the object at infinity as in figure b, we adjust it so that the image is right on top of the object. Points O and I coincide, and the rays are reflected right back on top of themselves. If we measure the angle θ shown in figure c, how can we find the focal angle?

▷ The object and image angles are the same; the angle labeled θ in the figure equals both of them. We therefore have $\theta_i + \theta_o = \theta = \theta_f$. Comparing figures b and c, it is indeed plausible that the angles are related by a factor of two.

At this point, we could consider our work to be done. Typically, we know the strength of the mirror, and we want to find the image location for a given object location. Given the mirror's focal angle and the object location, we can determine θ_o by trigonometry, subtract to find $\theta_i = \theta_f - \theta_o$, and then do more trig to find the image location.

There is, however, a shortcut that can save us from doing so much work. Figure a/3 shows two right triangles whose legs of length 1 coincide and whose acute angles are θ_o and θ_i . These can be related by trigonometry to the object and image distances shown in figure d:

$$\tan \theta_o = 1/d_o \quad \tan \theta_i = 1/d_i$$

Ever since the beginning of this chapter, we've been assuming small angles. For small angles, we can use the small-angle approximation $\tan x \approx x$ (for x in radians), giving simply

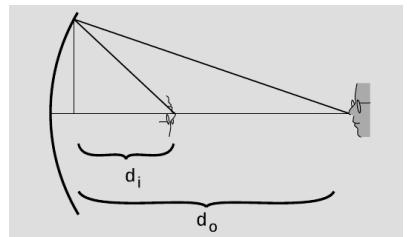
$$\theta_o = 1/d_o \quad \theta_i = 1/d_i.$$

We likewise define a distance called the focal length, f according to $f = 1/\theta_f$. In figure b, f is the distance from the mirror to the place where the rays cross. We can now reexpress the equation relating the object and image positions as

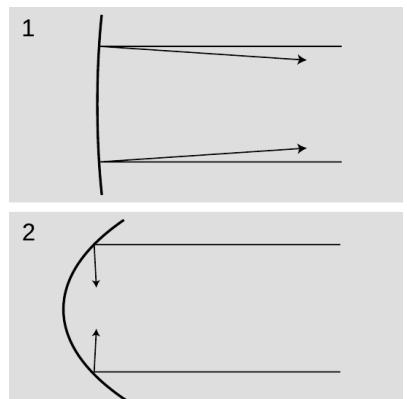
$$\frac{1}{f} = \frac{1}{d_i} + \frac{1}{d_o}.$$

Figure e summarizes the interpretation of the focal length and focal angle.¹

¹There is a standard piece of terminology which is that the “focal point” is the point lying on the optical axis at a distance from the mirror equal to the focal length. This term isn't particularly helpful, because it names a location where nothing normally happens. In particular, it is *not* normally the place where the rays come to a focus! — that would be the *image* point. In other words, we don't normally have $d_i = f$, unless perhaps $d_o = \infty$. A recent online discussion



d / The object and image distances



e / Mirror 1 is weaker than mirror 2. It has a shallower curvature, a longer focal length, and a smaller focal angle. It reflects rays at angles not much different than those that would be produced with a flat mirror.

Which form is better, $\theta_f = \theta_i + \theta_o$ or $1/f = 1/d_i + 1/d_o$? The angular form has in its favor its simplicity and its straightforward visual interpretation, but there are two reasons why we might prefer the second version. First, the numerical values of the angles depend on what we mean by “one unit” for the distance shown as 1 in figure a/1. Second, it is usually easier to measure distances rather than angles, so the distance form is more convenient for number crunching. Neither form is superior overall, and we will often need to use both to solve any given problem.²

A searchlight

example 2

Suppose we need to create a parallel beam of light, as in a searchlight. Where should we place the lightbulb? A parallel beam has zero angle between its rays, so $\theta_i = 0$. To place the lightbulb correctly, however, we need to know a distance, not an angle: the distance d_o between the bulb and the mirror. The problem involves a mixture of distances and angles, so we need to get everything in terms of one or the other in order to solve it. Since the goal is to find a distance, let’s figure out the image distance corresponding to the given angle $\theta_i = 0$. These are related by $d_i = 1/\theta_i$, so we have $d_i = \infty$. (Yes, dividing by zero gives infinity. Don’t be afraid of infinity. Infinity is a useful problem-solving device.) Solving the distance equation for d_o , we have

$$\begin{aligned} d_o &= (1/f - 1/d_i)^{-1} \\ &= (1/f - 0)^{-1} \\ &= f \end{aligned}$$

The bulb has to be placed at a distance from the mirror equal to its focal point.

Diopters

example 3

An equation like $d_i = 1/\theta_i$ really doesn’t make sense in terms of units. Angles are unitless, since radians aren’t really units, so the right-hand side is unitless. We can’t have a left-hand side with units of distance if the right-hand side of the same equation is unitless. This is an artifact of my cavalier statement that the conical bundles of rays spread out to a distance of 1 from the axis where they strike the mirror, without specifying the units used to measure this 1. In real life, optometrists define the thing we’re calling $\theta_i = 1/d_i$ as the “dioptic strength” of a lens or mirror, and measure it in units of inverse meters (m^{-1}), also known as diopters ($1 D=1 m^{-1}$).

among some physics teachers (<https://carnot.physics.buffalo.edu/archives>, Feb. 2006) showed that many disliked the terminology, felt it was misleading, or didn’t know it and would have misinterpreted it if they had come across it. That is, it appears to be what grammarians call a “skunked term” — a word that bothers half the population when it’s used incorrectly, and the other half when it’s used correctly.

²I would like to thank Fouad Ajami for pointing out the pedagogical advantages of using both equations side by side.

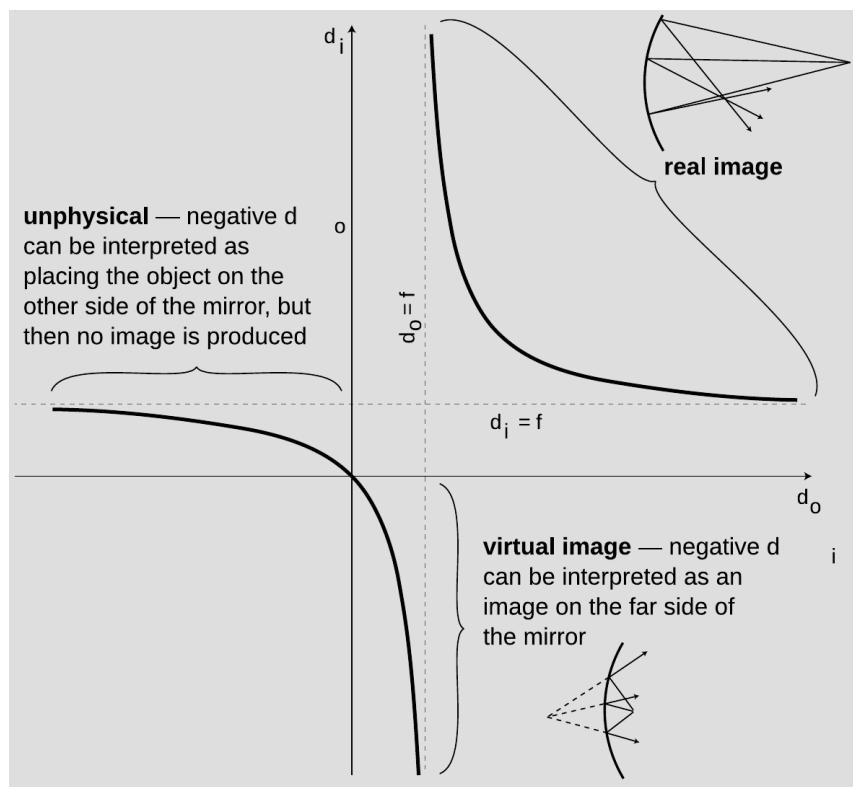
11.1.2 Magnification

We have already discussed in the previous chapter how to find the magnification of a virtual image made by a curved mirror. The result is the same for a real image, and we omit the proof, which is very similar. In our new notation, the result is $M = d_i/d_o$. A numerical example is given in sec. 11.2.

11.2 Other cases with curved mirrors

The equation $d_i = (1/f - 1/d_o)^{-1}$ can easily produce a negative result, but we have been thinking of d_i as a distance, and distances can't be negative. A similar problem occurs with $\theta_i = \theta_f - \theta_o$ for $\theta_o > \theta_f$. What's going on here?

The interpretation of the angular equation is straightforward. As we bring the object closer and closer to the image, θ_o gets bigger and bigger, and eventually we reach a point where $\theta_o = \theta_f$ and $\theta_i = 0$. This large object angle represents a bundle of rays forming a cone that is very broad, so broad that the mirror can no longer bend them back so that they reconverge on the axis. The image angle $\theta_i = 0$ represents an outgoing bundle of rays that are parallel. The outgoing rays never cross, so this is not a real image, unless we want to be charitable and say that the rays cross at infinity. If we go on bringing the object even closer, we get a virtual image.



f / A graph of the image distance d_i as a function of the object distance d_o .

To analyze the distance equation, let's look at a graph of d_i as a function of d_o . The branch on the upper right corresponds to the case of a real image. Strictly speaking, this is the only part of the graph that we've proven corresponds to reality, since we never did any geometry for other cases, such as virtual images. As discussed in the previous section, making d_o bigger causes d_i to become smaller, and vice-versa.

Letting d_o be less than f is equivalent to $\theta_o > \theta_f$: a virtual image is produced on the far side of the mirror. This is the first example of Wigner's “unreasonable effectiveness of mathematics” that we have encountered in optics. Even though our proof depended on the assumption that the image was real, the equation we derived turns out to be applicable to virtual images, provided that we either interpret the positive and negative signs in a certain way, or else modify the equation to have different positive and negative signs.

self-check A

Interpret the three places where, in physically realistic parts of the graph, the graph approaches an asymptote. ▷ Answer, p. 457

A flat mirror

example 4

We can even apply the equation to a flat mirror. As a sphere gets bigger and bigger, its surface is more and more gently curved. The planet Earth is so large, for example, that we cannot even perceive the curvature of its surface. To represent a flat mirror, we let the mirror's radius of curvature, and its focal length, become infinite. Dividing by infinity gives zero, so we have

$$1/d_o = -1/d_i,$$

or

$$d_o = -d_i.$$

If we interpret the minus sign as indicating a virtual image on the far side of the mirror from the object, this makes sense.

It turns out that for any of the six possible combinations of real or virtual images formed by converging or diverging lenses or mirrors, we can apply equations of the form

$$\theta_f = \theta_i + \theta_o$$

and

$$\frac{1}{f} = \frac{1}{d_i} + \frac{1}{d_o},$$

with only a modification of plus or minus signs. There are two possible approaches here. The approach we have been using so far is the more popular approach in American textbooks: leave the equation the same, but attach interpretations to the resulting negative

or positive values of the variables. The trouble with this approach is that one is then forced to memorize tables of sign conventions, e.g., that the value of d_i should be negative when the image is a virtual image formed by a converging mirror. Positive and negative signs also have to be memorized for focal lengths. Ugh! It's highly *unlikely* that any student has ever retained these lengthy tables in his or her mind for more than five minutes after handing in the final exam in a physics course. Of course one can always look such things up when they are needed, but the effect is to turn the whole thing into an exercise in blindly plugging numbers into formulas.

As you have gathered by now, there is another method which I think is better, and which I'll use throughout the rest of this book. In this method, all distances and angles are *positive by definition*, and we put in positive and negative signs in the *equations* depending on the situation. (I thought I was the first to invent this method, but I've been told that this is known as the European sign convention, and that it's fairly common in Europe.) Rather than memorizing these signs, we start with the generic equations

$$\begin{aligned}\theta_f &= \pm\theta_i \pm \theta_o \\ \frac{1}{f} &= \pm\frac{1}{d_i} \pm \frac{1}{d_o},\end{aligned}$$

and then determine the signs by a two-step method that depends on ray diagrams. There are really only two signs to determine, not four; the signs in the two equations match up in the way you'd expect. The following algorithm can be used to find the signs:

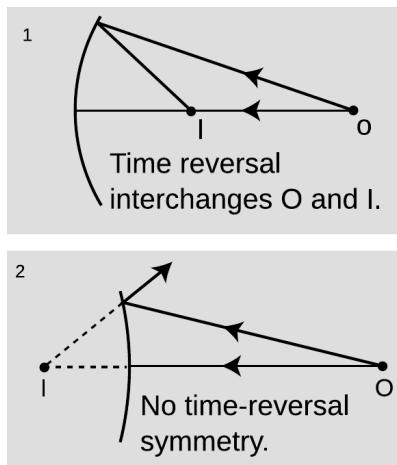
1. Use one of the following methods (your choice — either one works) to determine whether the two signs in the angle equation are the same or opposite.

a. *Long, straightforward method:* Use ray diagrams to decide whether θ_o and θ_i vary in the same way or in opposite ways. (In other words, decide whether making θ_o greater results in a greater value of θ_i or a smaller one.) Based on this, decide whether the two signs in the angle equation are the same or opposite.

b. *Shortcut method using time-reversal:* Check whether the image and object can be interchanged by time-reversing the ray diagram, figure g. (This happens for real images.) If so, then by symmetry the signs have to be the same. In all other cases, it turns out that the signs are opposite.

If the signs are the same, they're both positive, and you're done. If the signs are opposite, go on to step 2 to determine which is positive and which is negative.

2. If the signs are opposite, we need to decide which is the positive one and which is the negative. Since the focal angle is never negative, the smaller angle must be the one with a minus sign.



g / 1. In the case of a real image, time reversal (flipping the arrowheads on the rays) interchanges the roles of object and image. 2. For a virtual image, no such symmetry exists.

In step 1a, many students have trouble drawing the ray diagram correctly. For simplicity, you should always do your diagram for a point on the object that is on the axis of the mirror, and let one of your rays be the one that is emitted along the axis and reflected straight back on itself, as in the figures in sec. 11.1. As shown in figure a/4 in sec. 11.1, there are four angles involved: two at the mirror, one at the object (θ_o), and one at the image (θ_i). Make sure to draw in the normal to the mirror so that you can see the two angles at the mirror. These two angles are equal, so as you change the object position, they fan out or fan in, like opening or closing a book. Once you've drawn this effect, you should easily be able to tell whether θ_o and θ_i change in the same way or in opposite ways.

Although focal lengths are always positive in the method used in this book, you should be aware that diverging mirrors and lenses are assigned negative focal lengths in the other method, so if you see a lens labeled $f = -30$ cm, you'll know what it means.

An anti-shoplifting mirror

example 5

▷ Convenience stores often install a diverging mirror so that the clerk has a view of the whole store and can catch shoplifters. Use a ray diagram to show that the image is reduced, bringing more into the clerk's field of view. If the focal length of the mirror is 3.0

m, and the mirror is 7.0 m from the farthest wall, how deep is the image of the store?

▷ As shown in ray diagram h/1, d_i is less than d_o . The magnification, $M = d_i/d_o$, will be less than one, i.e., the image is actually reduced rather than magnified.

Apply the method above for determining the plus and minus signs. Step 1 (version a): The object is the point on the opposite wall. As an experiment, h/2, move the object closer. I did these drawings using illustration software, but if you were doing them by hand, you'd want to make the scale much larger for greater accuracy. Also, although I split figure h into two separate drawings in order to make them easier to understand, you're less likely to make a mistake if you do them on top of each other.

The two angles at the mirror fan out from the normal. Increasing θ_o has clearly made θ_i larger as well. (All four angles got bigger.) There must be a cancellation of the effects of changing the two terms on the right in the same way, and the only way to get such a cancellation is if the two terms in the angle equation have opposite signs:

$$\theta_f = +\theta_i - \theta_o$$

or

$$\theta_f = -\theta_i + \theta_o.$$

Step 2: Now which is the positive term and which is negative? Since the image angle is bigger than the object angle, the angle equation must be

$$\theta_f = \theta_i - \theta_o,$$

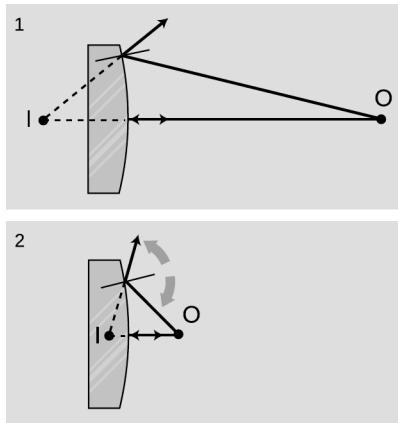
in order to give a positive result for the focal angle. The signs of the distance equation behave the same way:

$$\frac{1}{f} = \frac{1}{d_i} - \frac{1}{d_o}.$$

Solving for d_i , we find

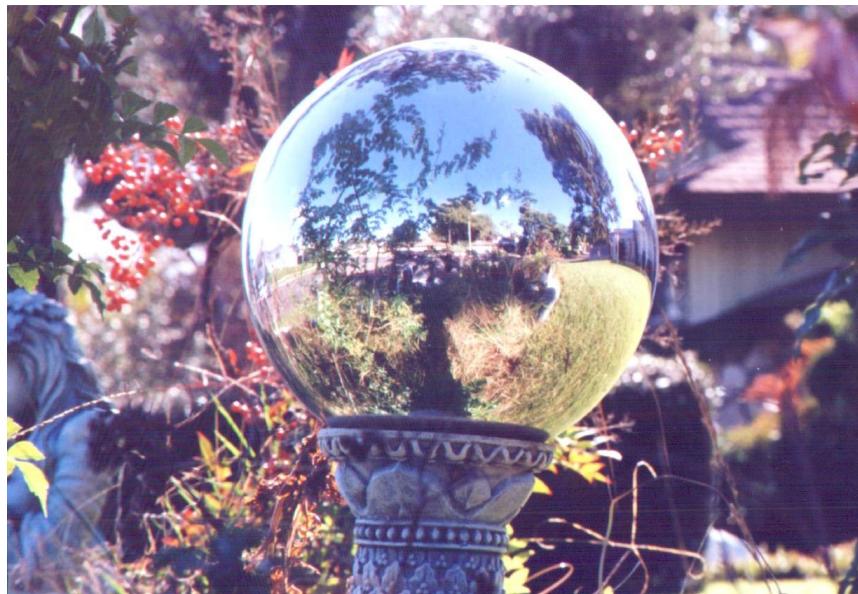
$$\begin{aligned} d_i &= \left(\frac{1}{f} + \frac{1}{d_o} \right)^{-1} \\ &= 2.1 \text{ m.} \end{aligned}$$

The image of the store is reduced by a factor of $2.1/7.0 = 0.3$, i.e., it is smaller by 70%.



h / Example 5.

i / A diverging mirror in the shape of a sphere. The image is reduced ($M < 1$). This is similar to example 5, but here the mirror's curve is not shallow.

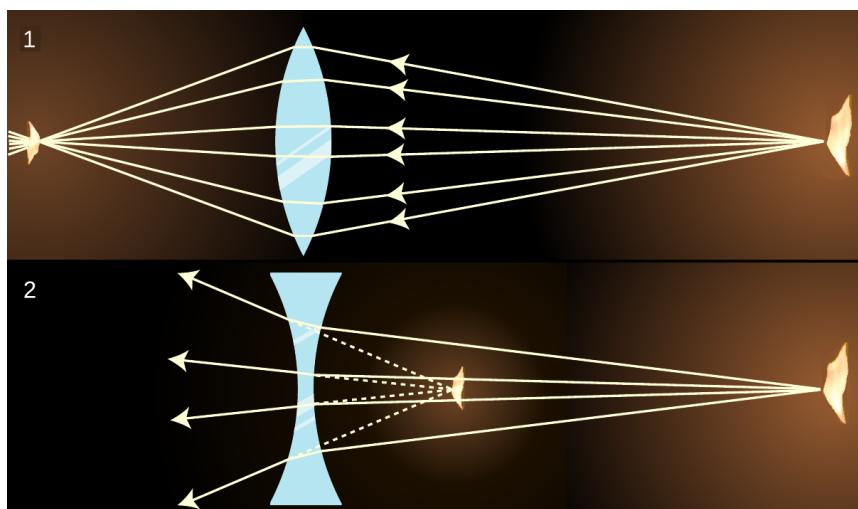


11.3 Images formed by lenses

11.3.1 Lenses

Figures j/1 and j/2 show examples of lenses forming images. There is essentially nothing for you to learn about imaging with lenses that is truly new. You already know how to construct and use ray diagrams, and you know about real and virtual images. The concept of the focal length of a lens is the same as for a curved mirror. The equations for locating images and determining magnifications are of the same form. It's really just a question of flexing your mental muscles on a few examples. The following self-checks and discussion questions will get you started. I've also made a video that demonstrates some applications and how to explain them with ray diagrams: <https://youtu.be/gL8awy6PWLQ>.

j / 1. A converging lens forms an image of a candle flame. 2. A diverging lens.



self-check B

- (1) In figures j/1 and j/2, classify the images as real or virtual.
- (2) Glass has an index of refraction that is greater than that of air. Consider the topmost ray in figure j/1. Explain why the ray makes a slight left turn upon entering the lens, and another left turn when it exits.
- (3) If the flame in figure j/2 was moved closer to the lens, what would happen to the location of the image? ▷ Answer, p. 457

Discussion questions

A In figures j/1 and j/2, the front and back surfaces are parallel to each other at the center of the lens. What will happen to a ray that enters near the center, but not necessarily along the axis of the lens? Draw a BIG ray diagram, and show a ray that comes from off axis.

In discussion questions B-F, don't draw ultra-detailed ray diagrams as in A.

B Suppose you wanted to change the setup in figure j/1 so that the location of the actual flame in the figure would instead be occupied by an image of a flame. Where would you have to move the candle to achieve this? What about in j/2?

C There are three qualitatively different types of image formation that can occur with lenses, of which figures j/1 and j/2 exhaust only two. Figure out what the third possibility is. Which of the three possibilities can result in a magnification greater than one? Cf. problem 8, p. 251.

D Classify the examples shown in figure k according to the types of images delineated in discussion question C.

E In figures j/1 and j/2, the only rays drawn were those that happened to enter the lenses. Discuss this in relation to figure k.

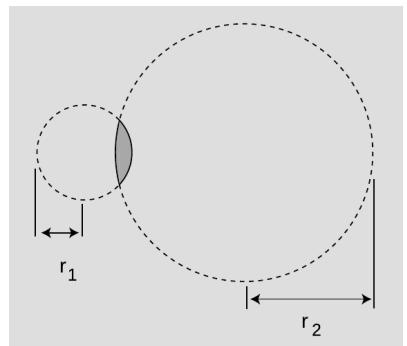
F In the right-hand side of figure k, the image viewed through the lens is in focus, but the side of the rose that sticks out from behind the lens is not. Why?

11.3.2 * The lensmaker's equation

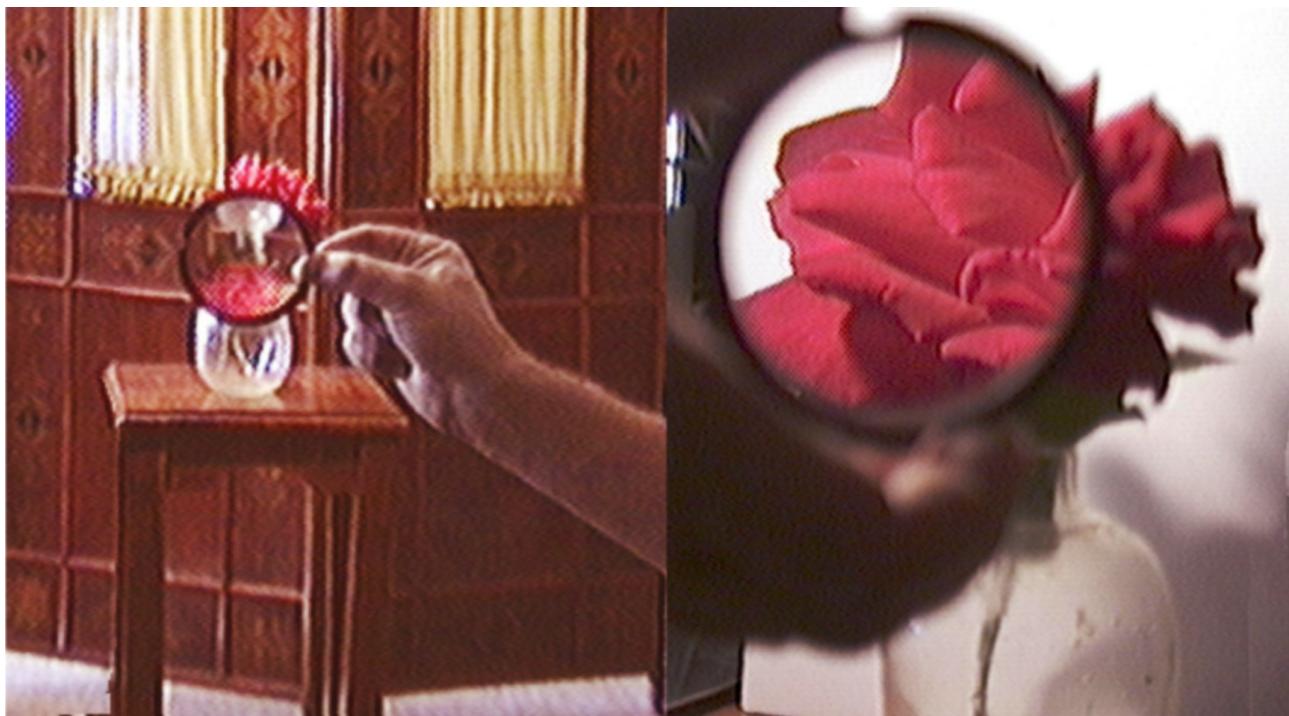
The focal length of a spherical mirror is simply $r/2$, but we cannot expect the focal length of a lens to be given by pure geometry, since it also depends on the index of refraction of the lens. Suppose we have a lens whose front and back surfaces are both spherical. (This is no great loss of generality, since any surface with a sufficiently shallow curvature can be approximated with a sphere.) Then if the lens is immersed in a medium with an index of refraction of 1, its focal length is given approximately by

$$f = \left[(n - 1) \left| \frac{1}{r_1} \pm \frac{1}{r_2} \right| \right]^{-1},$$

where n is the index of refraction and r_1 and r_2 are the radii of curvature of the two surfaces of the lens. This is known as the lensmaker's equation. In my opinion it is not particularly worthy



I / The radii of curvature appearing in the lensmaker's equation.

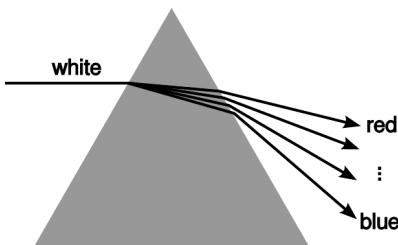


k / Two images of a rose created by the same lens and recorded with the same camera.

of memorization. The positive sign is used when both surfaces are curved outward or both are curved inward; otherwise a negative sign applies. The proof of this equation is left as an exercise to those readers who are sufficiently brave and motivated.

11.3.3 Dispersion

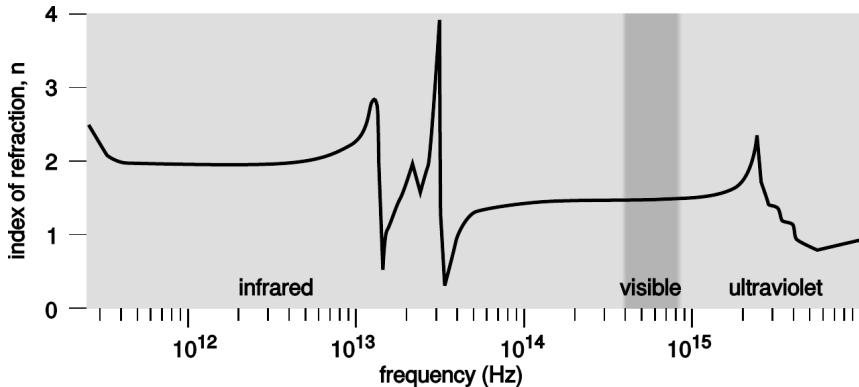
For most materials, we observe that the index of refraction depends slightly on wavelength, being highest at the blue end of the visible spectrum and lowest at the red. For example, white light disperses into a rainbow when it passes through a prism, m. This is the reason for the terminology introduced on p. 70, that any dependence of wave speed on wavelength is referred to as dispersion. Dispersion inside spherical raindrops is responsible for the creation of rainbows in the sky, and in an optical instrument such as the eye or a camera it is responsible for a type of aberration called chromatic aberration (sec. 11.4 and problem 12). Dispersion causes a wave that is not a pure sine wave to have its shape distorted as it travels (p. 70). As we'll see in sec. 16.2, it also causes the speed at which energy and information are transported by the wave to be different from what one might expect from a naive calculation. The microscopic reasons for dispersion of light in matter are discussed in optional section 11.3.4.



m / Dispersion of white light by a prism. White light is a mixture of all the wavelengths of the visible spectrum. Waves of different wavelengths undergo different amounts of refraction.

11.3.4 * Microscopic description of refraction

Given that the speed of light is different in different media, we've seen (sec. 5.6.1, p. 120) that refraction must occur. What we haven't yet explained is why the speed of light does depend on the medium.

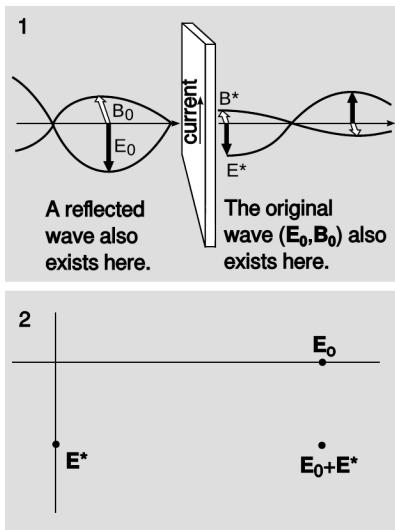


n / Index of refraction of silica glass, redrawn from Kitamura, Pilon, and Jonasz, Applied Optics 46 (2007) 8118, reprinted online at <http://www.seas.ucla.edu/~pilon/Publications/AO2007-1.pdf>.

A good clue as to what's going on comes from the figure n. The relatively minor variation of the index of refraction within the visible spectrum was misleading. At certain specific frequencies, n exhibits wild swings in the positive and negative directions. After each such swing, we reach a new, lower plateau on the graph. These frequencies are resonances. For example, the visible part of the spectrum lies on the left-hand tail of a resonance at about 2×10^{15} Hz, corresponding to the ultraviolet part of the spectrum. This resonance arises from the vibration of the electrons, which are bound to the nuclei as if by little springs. Because this resonance is narrow, the effect on visible-light frequencies is relatively small, but it is stronger at the blue end of the spectrum than at the red end. Near each resonance, not only does the index of refraction fluctuate wildly, but the glass becomes nearly opaque; this is because the vibration becomes very strong, causing energy to be dissipated as heat. The "staircase" effect is the same one visible in any resonance: oscillators have a finite response for $f \ll f_0$, but the response approaches zero for $f \gg f_0$.

So far, we have a qualitative explanation of the frequency-variation of the loosely defined "strength" of the glass's effect on a light wave, but we haven't explained why the effect is observed as a change in speed, or why each resonance is an up-down swing rather than a single positive peak. To understand these effects in more detail, we need to consider the phase response of the oscillator. The phase response reverses itself as we pass through a resonance.

Suppose that a plane wave is normally incident on the left side of a thin sheet of glass, o/1, at $f \ll f_0$. The light wave observed on the right side consists of a superposition of the incident wave consisting of \mathbf{E}_0 and \mathbf{B}_0 with a secondary wave \mathbf{E}^* and \mathbf{B}^* generated by the



o / 1. A wave incident on a sheet of glass excites current in the glass, which produce a secondary wave. 2. The secondary wave superposes with the original wave, as represented in the complex-number representation.

oscillating charges in the glass. Since the frequency is far below resonance, the response $q\mathbf{x}$ of a vibrating charge q is in phase with the driving force \mathbf{E}_0 . The current is the derivative of this quantity, and therefore 90 degrees ahead of it in phase. The magnetic field generated by a sheet of current is just what we would expect from the right-hand rule. We find, o/1, that the secondary wave is 90 degrees ahead of the incident one in phase. The incident wave still exists on the right side of the sheet, but it is superposed with the secondary one. Their addition is shown in o/2 using the complex number representation. The superposition of the two fields lags behind the incident wave, which is the effect we would expect if the wave had traveled more slowly through the glass.

In the case $f \gg f_0$, the same analysis applies except that the phase of the secondary wave is reversed. The transmitted wave is advanced rather than retarded in phase. This explains the dip observed in figure n after each spike.

All of this is in accord with our understanding of relativity, in which we saw that the universal speed c was to be understood fundamentally as a conversion factor between the units used to measure time and space — not as the speed of light. Since c isn't defined as the speed of light, it's of no fundamental importance whether light has a different speed in matter than it does in vacuum. In fact, the picture we've built up here is one in which all of our electromagnetic waves travel at c ; propagation at some other speed is only what appears to happen because of the superposition of the $(\mathbf{E}_0, \mathbf{B}_0)$ and $(\mathbf{E}^*, \mathbf{B}^*)$ waves, both of which move at c .

But it is worrisome that at the frequencies where $n < 1$, the speed of the wave is greater than c . According to special relativity, information is never supposed to be transmitted at speeds greater than c , since this would produce situations in which a signal could be received before it was transmitted! This difficulty is resolved in section 16.2, where we show that there are two different velocities that can be defined for a wave in a dispersive medium, the phase velocity and the group velocity. The group velocity is the velocity at which information is transmitted, and it is always less than c .

11.4 ★ Aberrations

An imperfection or distortion in an image is called an aberration. An aberration can be produced by a flaw in a lens or mirror, but even with a perfect optical surface some degree of aberration is unavoidable. To see why, consider the mathematical approximation we've been making, which is that the depth of the mirror's curve is small compared to d_o and d_i . Since only a flat mirror can satisfy this shallow-mirror condition perfectly, any curved mirror will deviate somewhat from the mathematical behavior we derived by assuming that condition. There are two main types of aberration in

curved mirrors, and these also occur with lenses.

(1) An object on the axis of the lens or mirror may be imaged correctly, but off-axis objects may be out of focus or distorted. In a camera, this type of aberration would show up as a fuzziness or warping near the sides of the picture when the center was perfectly focused. An example of this is shown in figure p, and in that particular example, the aberration is not a sign that the equipment was of low quality or wasn't right for the job but rather an inevitable result of trying to flatten a panoramic view; in the limit of a 360-degree panorama, the problem would be similar to the problem of representing the Earth's surface on a flat map, which can't be accomplished without distortion.

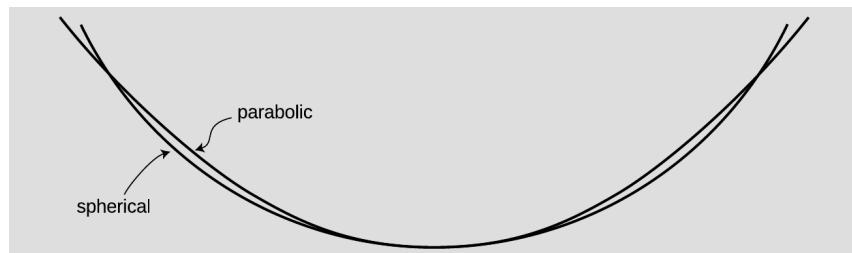


p / This photo was taken using a “fish-eye lens,” which gives an extremely large field of view.

(2) The image may be sharp when the object is at certain distances and blurry when it is at other distances. The blurriness occurs because the rays do not all cross at exactly the same point. If we know in advance the distance of the objects with which the mirror or lens will be used, then we can optimize the shape of the optical surface to make in-focus images in that situation. For instance, a spherical mirror will produce a perfect image of an object that is at the center of the sphere, because each ray is reflected di-

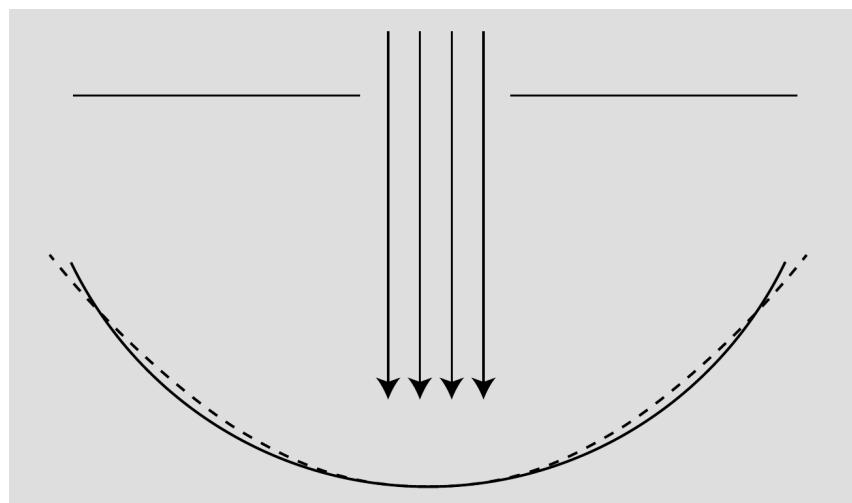
rectly onto the radius along which it was emitted. For objects at greater distances, however, the focus will be somewhat blurry. In astronomy the objects being used are always at infinity, so a spherical mirror is a poor choice for a telescope. A different shape (a parabola) is better specialized for astronomy.

q / Spherical mirrors are the cheapest to make, but parabolic mirrors are better for making images of objects at infinity. A sphere has equal curvature everywhere, but a parabola has tighter curvature at its center and gentler curvature at the sides.



One way of decreasing aberration is to use a small-diameter mirror or lens, or block most of the light with an opaque screen with a hole in it, so that only light that comes in close to the axis can get through. Either way, we are using a smaller portion of the lens or mirror whose curvature will be more shallow, thereby making the shallow-mirror (or thin-lens) approximation more accurate. Your eye does this by narrowing down the pupil to a smaller hole. In a camera, there is either an automatic or manual adjustment, and narrowing the opening is called “stopping down.” The disadvantage of stopping down is that light is wasted, so the image will be dimmer or a longer exposure must be used.

r / Even though the spherical mirror (solid line) is not well adapted for viewing an object at infinity, we can improve its performance greatly by stopping it down. Now the only part of the mirror being used is the central portion, where its shape is virtually indistinguishable from a parabola (dashed line).



What I would suggest you take away from this discussion for the sake of your general scientific education is simply an understanding of what an aberration is, why it occurs, and how it can be reduced, not detailed facts about specific types of aberrations.



The Hubble Space Telescope was placed into orbit with faulty optics in 1990. Its main mirror was supposed to have been nearly parabolic, since it is an astronomical telescope, meant for producing images of objects at infinity. However, contractor Perkin Elmer had delivered a faulty mirror, which produced aberrations. The large photo shows astronauts putting correcting mirrors in place in 1993. The two small photos show images produced by the telescope before and after the fix.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 Apply the equation $M = d_i/d_o$ to the case of a flat mirror.
▷ Solution, p. 447

2 A concave surface that reflects sound waves can act just like a converging mirror. Suppose that, standing near such a surface, you are able to find a point where you can place your head so that your own whispers are focused back on your head, so that they sound loud to you. Given your distance to the surface, what is the surface's focal length? ✓



3 Use the method described in the text to derive the equation relating object distance to image distance for the case of a virtual image produced by a converging mirror. ▷ Solution, p. 448

4 Rank the focal lengths of the mirrors in the figure, from shortest to longest. Explain.

5 (a) A converging mirror with a focal length of 20 cm is used to create an image, using an object at a distance of 10 cm. Is the image real, or is it virtual? (b) How about $f = 20$ cm and $d_o = 30$ cm? (c) What if it was a *diverging* mirror with $f = 20$ cm and $d_o = 10$ cm? (d) A diverging mirror with $f = 20$ cm and $d_o = 30$ cm? ▷ Solution, p. 448

Problem 4.

6 (a) Make up a numerical example of a virtual image formed by a converging mirror with a certain focal length, and determine the magnification. (You will need the result of problem 3.) Make sure to choose values of d_o and f that would actually produce a virtual image, not a real one. Now change the location of the object *a little bit* and redetermine the magnification, showing that it changes. At my local department store, the cosmetics department sells hand mirrors advertised as giving a magnification of 5 times. How would you interpret this?

(b) Suppose a Newtonian telescope is being used for astronomical observing. Assume for simplicity that no eyepiece is used, and assume a value for the focal length of the mirror that would be reasonable for an amateur instrument that is to fit in a closet. Is the angular magnification different for objects at different distances? For example, you could consider two planets, one of which is twice as far as the other. ▷ Solution, p. 448

7 (a) Find a case where the magnification of a curved mirror is infinite. Is the *angular* magnification infinite from any realistic viewing position? (b) Explain why an arbitrarily large magnification can't be achieved by having a sufficiently small value of d_o .

▷ Solution, p. 448

8 As discussed in question 7, there are two types of curved mirrors, concave and convex. Make a list of all the possible combinations of types of images (virtual or real) with types of mirrors (concave and convex). (Not all of the four combinations are physically possible.) Now for each one, use ray diagrams to determine whether increasing the distance of the object from the mirror leads to an increase or a decrease in the distance of the image from the mirror.

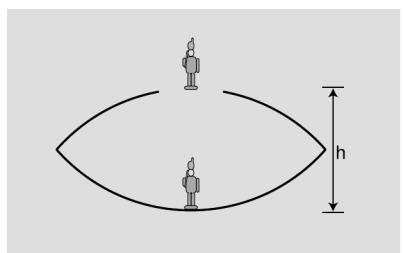
Draw BIG ray diagrams! Each diagram should use up about half a page of paper.

Some tips: To draw a ray diagram, you need two rays. For one of these, pick the ray that comes straight along the mirror's axis, since its reflection is easy to draw. After you draw the two rays and locate the image for the original object position, pick a new object position that results in the same type of image, and start a new ray diagram, in a different color of pen, right on top of the first one. For the two new rays, pick the ones that just happen to hit the mirror at the same two places; this makes it much easier to get the result right without depending on extreme accuracy in your ability to draw the reflected rays.

9 If the user of an astronomical telescope moves her head closer to or farther away from the image she is looking at, does the magnification change? Does the angular magnification change? Explain. (For simplicity, assume that no eyepiece is being used.)

▷ Solution, p. 449

10 (a) A converging mirror is being used to create a virtual image. What is the range of possible magnifications? (b) Do the same for the other types of images that can be formed by curved mirrors (both converging and diverging).



Problem 29.

11 Suppose a converging lens is constructed of a type of plastic whose index of refraction is less than that of water. How will the lens's behavior be different if it is placed underwater?

▷ Solution, p. 449

12 There are two main types of telescopes, refracting (using a lens) and reflecting (using a mirror as in figure p on p. 224). (Some telescopes use a mixture of the two types of elements: the light first encounters a large curved mirror, and then goes through an eyepiece that is a lens. To keep things simple, assume no eyepiece is used.) What implications would the color-dependence of focal length have for the relative merits of the two types of telescopes? Describe the case where an image is formed of a white star. You may find it helpful to draw a ray diagram.

13 Based on Snell's law, explain why rays of light passing through the edges of a converging lens are bent more than rays passing through parts closer to the center. It might seem like it should be the other way around, since the rays at the edge pass through less glass — shouldn't they be affected less? In your answer:

- Include a ray diagram showing a huge, full-page, close-up view of the relevant part of the lens.
- Make use of the fact that the front and back surfaces aren't always parallel; a lens in which the front and back surfaces *are* always parallel doesn't focus light at all, so if your explanation doesn't make use of this fact, your argument must be incorrect.
- Make sure your argument still works even if the rays don't come in parallel to the axis or from a point on the axis.

▷ Solution, p. 449

14 When you take pictures with a camera, the distance between the lens and the film or chip has to be adjusted, depending on the distance at which you want to focus. This is done by moving the lens. If you want to change your focus so that you can take a picture of something farther away, which way do you have to move the lens? Explain using ray diagrams. [Based on a problem by Eric Mazur.]

15 When swimming underwater, why is your vision made much clearer by wearing goggles with flat pieces of glass that trap air behind them? [Hint: You can simplify your reasoning by considering the special case where you are looking at an object far away, and along the optic axis of the eye.]

▷ Solution, p. 450

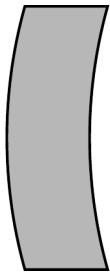
- 16** An object is more than one focal length from a converging lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in chapter 11, determine the positive and negative signs in the equation $1/f = \pm 1/d_i \pm 1/d_o$. (c) The images of the rose in section 4.2 were made using a lens with a focal length of 23 cm. If the lens is placed 80 cm from the rose, locate the image. \checkmark

- 17** Zahra likes to play practical jokes on the friends she goes hiking with. One night, by a blazing camp fire, she stealthily uses a lens of focal length f to gather light from the fire and make a hot spot on Becky's neck. (a) Using the method of section 11.2, p. 237, draw a ray diagram and set up the equation for the image location, inferring the correct plus and minus signs from the diagram. (b) Let A be the distance from the lens to the campfire, and B the distance from the lens to Becky's neck. Consider the following nine possibilities:

B		
$< f$ $= f$ $> f$		
$A < f$	[]	[]
$A = f$	[]	[]
$A > f$	[]	[]

By reasoning about your equation from part a, determine which of these are possible and which are not. \triangleright Solution, p. 451

- 18** The figure shows a lens with surfaces that are curved, but whose thickness is constant along any horizontal line. Use the lens-maker's equation to prove that this "lens" is not really a lens at all. \triangleright Solution, p. 451

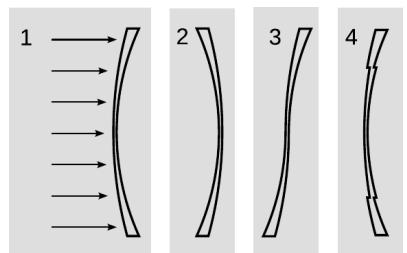


Problem 18.

- 19** The figure shows four lenses. Lens 1 has two spherical surfaces. Lens 2 is the same as lens 1 but turned around. Lens 3 is made by cutting through lens 1 and turning the bottom around. Lens 4 is made by cutting a central circle out of lens 1 and recessing it.

- (a) A parallel beam of light enters lens 1 from the left, parallel to its axis. Reasoning based on Snell's law, will the beam emerging from the lens be bent inward, or outward, or will it remain parallel to the axis? Explain your reasoning. As part of your answer, make a huge drawing of one small part of the lens, and apply Snell's law at both interfaces. Recall that rays are bent more if they come to the interface at a larger angle with respect to the normal.

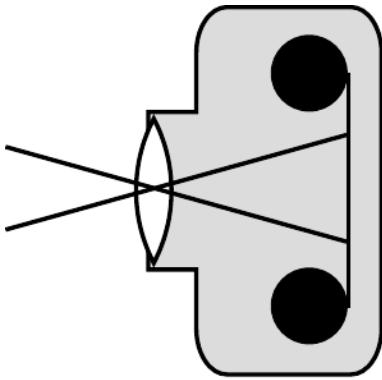
- (b) What will happen with lenses 2, 3, and 4? Explain. Drawings are not necessary. \triangleright Solution, p. 452



Problem 19.

20 An object is less than one focal length from a converging lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in chapter 11, determine the positive and negative signs in the equation $1/f = \pm 1/d_i \pm 1/d_o$. (c) The images of the rose in section 4.2 were made using a lens with a focal length of 23 cm. If the lens is placed 10 cm from the rose, locate the image. ✓

21 Nearsighted people wear glasses whose lenses are diverging. (a) Draw a ray diagram. For simplicity pretend that there is no eye behind the glasses. (b) Using reasoning like that developed in chapter 11, determine the positive and negative signs in the equation $1/f = \pm 1/d_i \pm 1/d_o$. (c) If the focal length of the lens is 50.0 cm, and the person is looking at an object at a distance of 80.0 cm, locate the image. ✓



Problem 22.

22 Two standard focal lengths for camera lenses are 50 mm (standard) and 28 mm (wide-angle). To see how the focal lengths relate to the angular size of the field of view, it is helpful to visualize things as represented in the figure. Instead of showing many rays coming from the same point on the same object, as we normally do, the figure shows two rays from two different objects. Although the lens will intercept infinitely many rays from each of these points, we have shown only the ones that pass through the center of the lens, so that they suffer no angular deflection. (Any angular deflection at the front surface of the lens is canceled by an opposite deflection at the back, since the front and back surfaces are parallel at the lens's center.) What is special about these two rays is that they are aimed at the edges of one 35-mm-wide frame of film; that is, they show the limits of the field of view. Throughout this problem, we assume that d_o is much greater than d_i . (a) Compute the angular width of the camera's field of view when these two lenses are used. (b) Use small-angle approximations to find a simplified equation for the angular width of the field of view, θ , in terms of the focal length, f , and the width of the film, w . Your equation should not have any trig functions in it. Compare the results of this approximation with your answers from part a. (c) Suppose that we are holding constant the aperture (amount of surface area of the lens being used to collect light). When switching from a 50-mm lens to a 28-mm lens, how many times longer or shorter must the exposure be in order to make a properly developed picture, i.e., one that is not under- or overexposed? [Based on a problem by Arnold Arons.]

▷ Solution, p. 452

23 A nearsighted person is one whose eyes focus light too strongly, and who is therefore unable to relax the lens inside her eye sufficiently to form an image on her retina of an object that is too far away.

- (a) Draw a ray diagram showing what happens when the person tries, with uncorrected vision, to focus at infinity.
- (b) What type of lenses do her glasses have? Explain.
- (c) Draw a ray diagram showing what happens when she wears glasses. Locate both the image formed by the glasses and the final image.
- (d) Suppose she sometimes uses contact lenses instead of her glasses. Does the focal length of her contacts have to be less than, equal to, or greater than that of her glasses? Explain.

24 Fred's eyes are able to focus on things as close as 5.0 cm. Fred holds a magnifying glass with a focal length of 3.0 cm at a height of 2.0 cm above a flatworm. (a) Locate the image, and find the magnification. (b) Without the magnifying glass, from what distance would Fred want to view the flatworm to see its details as well as possible? With the magnifying glass? (c) Compute the angular magnification.

25 It would be annoying if your eyeglasses produced a magnified or reduced image. Prove that when the eye is very close to a lens, and the lens produces a virtual image, the angular magnification is always approximately equal to 1 (regardless of whether the lens is diverging or converging).

26 Under ordinary conditions, gases have indices of refraction only a little greater than that of vacuum, i.e., $n = 1 + \epsilon$, where ϵ is some small number. Suppose that a ray crosses a boundary between a region of vacuum and a region in which the index of refraction is $1 + \epsilon$. Find the maximum angle by which such a ray can ever be deflected, in the limit of small ϵ . ▷ Hint, p. 443 ✓

27 A converging mirror has focal length f . An object is located at a distance $(1 + \epsilon)f$ from the mirror, where ϵ is small. Find the distance of the image from the mirror, simplifying your result as much as possible by using the assumption that ϵ is small.

▷ Answer, p. 459

28 A diverging mirror of focal length f is fixed, and faces down. An object is dropped from the surface of the mirror, and falls away from it with acceleration g . The goal of the problem is to find the maximum velocity of the image.

- (a) Describe the motion of the image verbally, and explain why we should expect there to be a maximum velocity.
- (b) Use arguments based on units to determine the form of the solution, up to an unknown unitless multiplicative constant.
- (c) Complete the solution by determining the unitless constant.

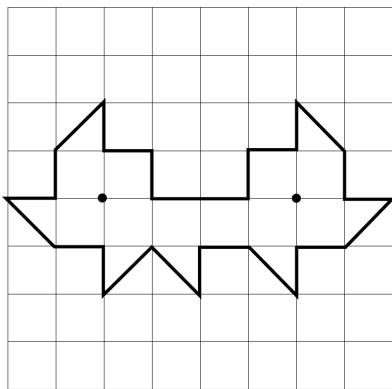
✓

29 The figure shows a device for constructing a realistic optical illusion. Two mirrors of equal focal length are put against each other with their silvered surfaces facing inward. A small object placed in the bottom of the cavity will have its image projected in the air above. The way it works is that the top mirror produces a virtual image, and the bottom mirror then creates a real image of the virtual image. (a) Show that if the image is to be positioned as shown, at the mouth of the cavity, then the focal length of the mirrors is related to the dimension h via the equation

$$\frac{1}{f} = \frac{1}{h} + \frac{1}{h + \left(\frac{1}{h} - \frac{1}{f}\right)^{-1}}.$$

- (b) Restate the equation in terms of a single variable $x = h/f$, and show that there are two solutions for x . Which solution is physically consistent with the assumptions of the calculation? *

30 Suppose we have a polygonal room whose walls are mirrors, and there a pointlike light source in the room. In most such examples, every point in the room ends up being illuminated by the light source after some finite number of reflections. A difficult mathematical question, first posed in the middle of the last century, is whether it is ever possible to have an example in which the whole room is *not* illuminated. (Rays are assumed to be absorbed if they strike exactly at a vertex of the polygon, or if they pass exactly through the plane of a mirror.)



Problem 30.

The problem was finally solved in 1995 by G.W. Tokarsky, who found an example of a room that was not illuminable from a certain point. Figure 30 shows a slightly simpler example found two years later by D. Castro. If a light source is placed at either of the locations shown with dots, the other dot remains unilluminated, although every other point is lit up. It is not straightforward to prove rigorously that Castro's solution has this property. However, the plausibility of the solution can be demonstrated as follows.

Suppose the light source is placed at the right-hand dot. Locate all the images formed by single reflections. Note that they form a

regular pattern. Convince yourself that none of these images illuminates the left-hand dot. Because of the regular pattern, it becomes plausible that even if we form images of images, images of images of images, etc., none of them will ever illuminate the other dot.

There are various other versions of the problem, some of which remain unsolved. The book by Klee and Wagon gives a good introduction to the topic, although it predates Tokarsky and Castro's work.

References:

- G.W. Tokarsky, "Polygonal Rooms Not Illuminable from Every Point." Amer. Math. Monthly 102, 867-879, 1995.
D. Castro, "Corrections." Quantum 7, 42, Jan. 1997.
V. Klee and S. Wagon, *Old and New Unsolved Problems in Plane Geometry and Number Theory*. Mathematical Association of America, 1991. *

31 The intensity of a beam of light is defined as the power per unit area incident on a perpendicular surface. Suppose that a beam of light in a medium with index of refraction n reaches the surface of the medium, with air on the outside. Its incident angle with respect to the normal is θ . (All angles are in radians.) Only a fraction f of the energy is transmitted, the rest being reflected. Because of this, we might expect that the transmitted ray would always be less intense than the incident one. But because the transmitted ray is refracted, it becomes narrower, causing an additional change in intensity by a factor $g > 1$. The product of these factors $I = fg$ can be greater than one. The purpose of this problem is to estimate the maximum amount of intensification.

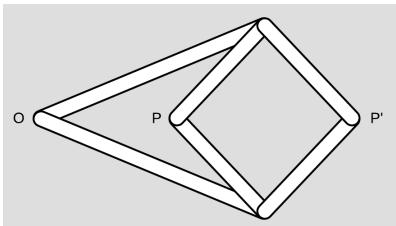
We will use the small-angle approximation $\theta \ll 1$ freely, in order to make the math tractable. In our previous studies of waves, we have only studied the factor f in the one-dimensional case where $\theta = 0$. The generalization to $\theta \neq 0$ is rather complicated and depends on the polarization, but for unpolarized light, we can use Schlick's approximation,

$$f(\theta) = f(0)(1 - \cos \theta)^5,$$

where the value of f at $\theta = 0$ is found as in problem 10 on p. 143.

- (a) Using small-angle approximations, obtain an expression for g of the form $g \approx 1 + P\theta^2$, and find the constant P . \triangleright Answer, p. 459
(b) Find an expression for I that includes the two leading-order terms in θ . We will call this expression I_2 . Obtain a simple expression for the angle at which I_2 is maximized. As a check on your work, you should find that for $n = 1.3$, $\theta = 63^\circ$. (Trial-and-error maximization of I gives 60° .)
(c) Find an expression for the maximum value of I_2 . You should find that for $n = 1.3$, the maximum intensification is 31%.

*



Problem 32.

32 A mechanical linkage is a device that changes one type of motion into another. The most familiar example occurs in a gasoline car's engine, where a connecting rod changes the linear motion of the piston into circular motion of the crankshaft. The top panel of the figure shows a mechanical linkage invented by Peaucellier in 1864, and independently by Lipkin around the same time. It consists of six rods joined by hinges, the four short ones forming a rhombus. Point O is fixed in space, but the apparatus is free to rotate about O. Motion at P is transformed into a different motion at P' (or vice versa).

Geometrically, the linkage is a mechanical implementation of the ancient problem of inversion in a circle. Considering the case in which the rhombus is folded flat, let the k be the distance from O to the point where P and P' coincide. Form the circle of radius k with its center at O. As P and P' move in and out, points on the inside of the circle are always mapped to points on its outside, such that $rr' = k^2$. That is, the linkage is a type of analog computer that exactly solves the problem of finding the inverse of a number r . Inversion in a circle has many remarkable geometrical properties, discussed in H.S.M. Coxeter, *Introduction to Geometry*, Wiley, 1961. If a pen is inserted through a hole at P, and P' is traced over a geometrical figure, the Peaucellier linkage can be used to draw a kind of image of the figure.

A related problem is the construction of pictures, like the one in the bottom panel of the figure, called anamorphs. The drawing of the column on the paper is highly distorted, but when the reflecting cylinder is placed in the correct spot on top of the page, an undistorted image is produced inside the cylinder. (Wide-format movie technologies such as Cinemascope are based on similar principles.)

Show that the Peaucellier linkage does *not* convert correctly between an image and its anamorph, and design a modified version of the linkage that does. Some knowledge of analytic geometry will be helpful. *

Exercise 11A: Object and image distances

Equipment:

- optical benches
- converging mirrors
- illuminated objects

1. Set up the optical bench with the mirror at zero on the centimeter scale. Set up the illuminated object on the bench as well.
2. Each group will locate the image for their own value of the object distance, by finding where a piece of paper has to be placed in order to see the image on it. (The instructor will do one point as well.) Note that you will have to tilt the mirror a little so that the paper on which you project the image doesn't block the light from the illuminated object.

Is the image real or virtual? How do you know? Is it inverted, or uninverted?

Draw a ray diagram.

3. Measure the image distance and write your result in the table on the board. Do the same for the magnification.
4. What do you notice about the trend of the data on the board? Draw a second ray diagram with a different object distance, and show why this makes sense. Some tips for doing this correctly: (1) For simplicity, use the point on the object that is on the mirror's axis. (2) You need to trace two rays to locate the image. To save work, don't just do two rays at random angles. You can either use the on-axis ray as one ray, or do two rays that come off at the same angle, one above and one below the axis. (3) Where each ray hits the mirror, draw the normal line, and make sure the ray is at equal angles on both sides of the normal.
5. We will find the mirror's focal length from the instructor's data-point. Then, using this focal length, calculate a theoretical prediction of the image distance, and write it on the board next to the experimentally determined image distance.

Exercise 11B: How strong are your glasses?

This exercise was created by Dan MacIsaac.

Equipment:

eyeglasses

diverging lenses for students who don't wear glasses, or who use glasses with converging lenses

rulers and metersticks

scratch paper

marking pens

Most people who wear glasses have glasses whose lenses are outbending, which allows them to focus on objects far away. Such a lens cannot form a real image, so its focal length cannot be measured as easily as that of a converging lens. In this exercise you will determine the focal length of your own glasses by taking them off, holding them at a distance from your face, and looking through them at a set of parallel lines on a piece of paper. The lines will be reduced (the lens's magnification is less than one), and by adjusting the distance between the lens and the paper, you can make the magnification equal $1/2$ exactly, so that two spaces between lines as seen through the lens fit into one space as seen simultaneously to the side of the lens. This object distance can be used in order to find the focal length of the lens.

1. Use a marker to draw three evenly spaced parallel lines on the paper. (A spacing of a few cm works well.)
2. Does this technique really measure magnification or does it measure angular magnification? What can you do in your experiment in order to make these two quantities nearly the same, so the math is simpler?
3. Before taking any numerical data, use algebra to find the focal length of the lens in terms of d_o , the object distance that results in a magnification of $1/2$.
4. Measure the object distance that results in a magnification of $1/2$, and determine the focal length of your lens.

Chapter 12

Wave optics

Electron microscopes can make images of individual atoms, but why will a visible-light microscope never be able to? Stereo speakers create the illusion of music that comes from a band arranged in your living room, but why doesn't the stereo illusion work with bass notes? Why are computer chip manufacturers investing billions of dollars in equipment to etch chips with x-rays instead of visible light?

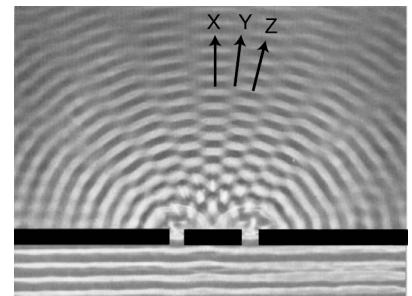
The answers to all of these questions have to do with the subject of wave optics. So far this book has discussed the interaction of light waves with matter, and its practical applications to optical devices like mirrors, but we have used the ray model of light almost exclusively. Hardly ever have we explicitly made use of the fact that light is an electromagnetic wave. We were able to get away with the simple ray model because the chunks of matter we were discussing, such as lenses and mirrors, were thousands of times larger than a wavelength of light. We now turn to phenomena and devices that can only be understood using the wave model of light.

12.1 Diffraction

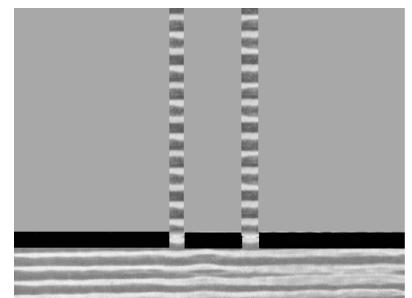
Figure a shows a typical problem in wave optics, enacted with water waves. It may seem surprising that we don't get a simple pattern like figure b, but the pattern would only be that simple if the wavelength was hundreds of times shorter than the distance between the gaps in the barrier and the widths of the gaps.

Wave optics is a broad subject, but this example will help us to pick out a reasonable set of restrictions to make things more manageable:

- (1) We restrict ourselves to cases in which a wave travels through a uniform medium, encounters a certain area in which the medium has different properties, and then emerges on the other side into a second uniform region.
- (2) We assume that the incoming wave is a nice tidy sine-wave pattern with wavefronts that are lines (or, in three dimensions, planes).
- (3) In figure a we can see that the wave pattern immediately beyond the barrier is rather complex, but farther on it sorts itself



a / In this view from overhead, a straight, sinusoidal water wave encounters a barrier with two gaps in it. Strong wave vibration occurs at angles X and Z, but there is none at all at angle Y. (The figure has been retouched from a real photo of water waves. In reality, the waves beyond the barrier would be much weaker than the ones before it, and they would therefore be difficult to see.)



b / This doesn't happen.

out into a set of wedges separated by gaps in which the water is still. We will restrict ourselves to studying the simpler wave patterns that occur farther away, so that the main question of interest is how intense the outgoing wave is at a given angle.

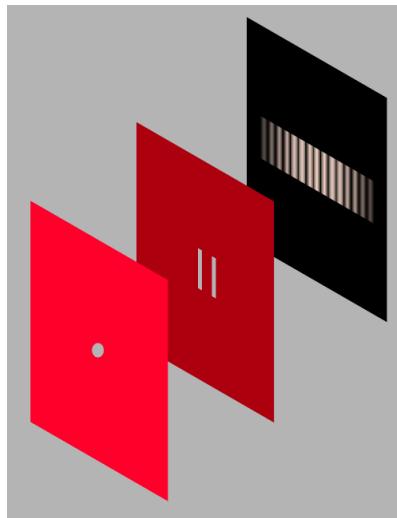
The kind of phenomenon described by restriction (1) is called *diffraction*. Diffraction can be defined as the behavior of a wave when it encounters an obstacle or a nonuniformity in its medium. In general, diffraction causes a wave to bend around obstacles and make patterns of strong and weak waves radiating out beyond the obstacle. Understanding diffraction is the central problem of wave optics. If you understand diffraction, even the subset of diffraction problems that fall within restrictions (2) and (3), the rest of wave optics is icing on the cake.

Diffraction can be used to find the structure of an unknown diffracting object: even if the object is too small to study with ordinary imaging, it may be possible to work backward from the diffraction pattern to learn about the object. The structure of a crystal, for example, can be determined from its x-ray diffraction pattern.

Diffraction can also be a bad thing. In a telescope, for example, light waves are diffracted by all the parts of the instrument. This will cause the image of a star to appear fuzzy even when the focus has been adjusted correctly. By understanding diffraction, one can learn how a telescope must be designed in order to reduce this problem — essentially, it should have the biggest possible diameter.

There are two ways in which restriction (2) might commonly be violated. First, the light might be a mixture of wavelengths. If we simply want to observe a diffraction pattern or to use diffraction as a technique for studying the object doing the diffracting (e.g., if the object is too small to see with a microscope), then we can pass the light through a colored filter before diffracting it.

A second issue is that light from sources such as the sun or a lightbulb does not consist of a nice neat plane wave, except over very small regions of space. Different parts of the wave are out of step with each other, and the wave is referred to as *incoherent*. One way of dealing with this is shown in figure c. After filtering to select a certain wavelength of red light, we pass the light through a small pinhole. The region of the light that is intercepted by the pinhole is so small that one part of it is not out of step with another. Beyond the pinhole, light spreads out in a spherical wave; this is analogous to what happens when you speak into one end of a paper towel roll and the sound waves spread out in all directions from the other end. By the time the spherical wave gets to the double slit it has spread out and reduced its curvature, so that we can now think of it as a simple plane wave.

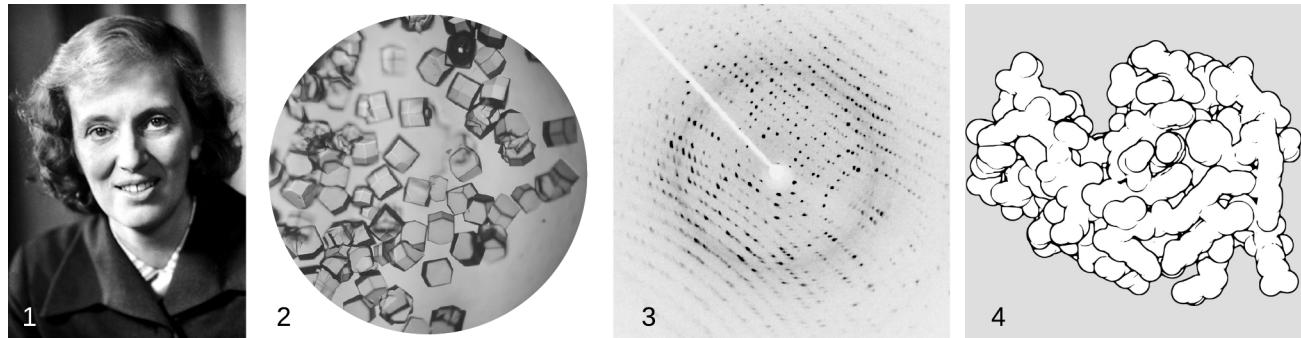


c / A practical, low-tech setup for observing diffraction of light.

If this seems laborious, you may be relieved to know that modern technology gives us an easier way to produce a single-wavelength, coherent beam of light: the laser.

The parts of the final image on the screen in c are called diffraction fringes. The center of each fringe is a point of maximum brightness, and halfway between two fringes is a minimum.

Because the diffraction spreads out in all directions, the fringes do not have a spacing or locations that we can define in units of meters — such dimensions would get bigger if we moved the screen farther away. But a diffraction does have definable *angular* dimensions. An example would be the angle between lines X and Z in figure a. A common problem in diffraction is to find some angular dimension or position θ , given one or more linear dimensions, such as the distance d between a pair of double slits. Conversely, one can observe a θ and determine an unknown d . In an example such as figure d, there could actually be many θ 's and many d 's.



d / 1. Dorothy Hodgkin was awarded the Nobel Prize in chemistry in 1964 for her work in determining the structures of organic molecules using x-ray diffraction. The following figures sketch the technique. 2. The unknown molecule, in this case a natural antibacterial enzyme called lysozyme, is crystallized. Without crystallization, the random orientations of all the molecules would make the diffraction pattern a blurred average over orientations. 3. The x-ray diffraction pattern is observed. 4. The three-dimensional structure of the molecule is determined. As a much simpler example of the determination of an unknown structure from its diffraction pattern, see problem 9, p. 278. The more general technique is called Fourier analysis.

12.2 Scaling of diffraction

This chapter has “optics” in its title, so it is nominally about light, but we started out with an example involving water waves. Water waves are certainly easier to visualize, but is this a legitimate comparison? In fact the analogy works quite well, despite the fact that a light wave has a wavelength about a million times shorter. This is because diffraction effects scale uniformly. That is, if we enlarge or reduce the whole diffraction situation by the same factor, including both the wavelengths and the sizes of the obstacles the wave encounters, the result is still a valid solution.

This is unusually simple behavior! Most physical phenomena do not scale in any simple way. For example, we can’t have spiders the size of horses, because weight scales like the cube of the dimensions, while the strength of the animal’s legs scale like the square.

Of course water waves and light waves differ in many ways, not just in scale, but the general facts you will learn about diffraction are applicable to all waves.

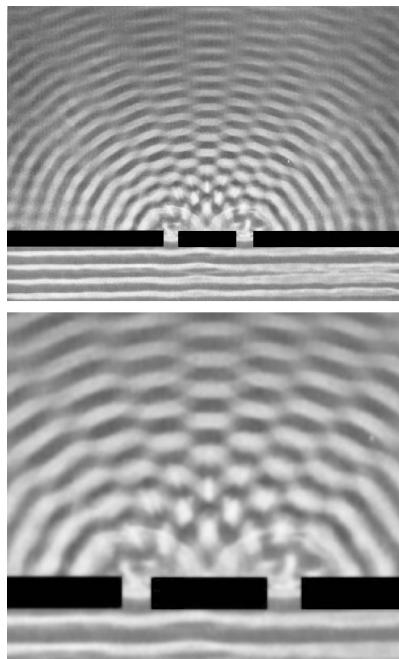
Another way of stating the simple scaling behavior of diffraction is that the diffraction angles we get depend only on the unitless ratio λ/d , where λ is the wavelength of the wave and d is some dimension of the diffracting objects, e.g., the center-to-center spacing between the slits in figure a. If, for instance, we scale up both λ and d by a factor of 37, the ratio λ/d will be unchanged.

Discussion question

- A Why would x-rays rather than visible light be used to find the structure of a crystal, as in figure d? Sound waves are used to make images of fetuses in the womb. What would influence the choice of wavelength?

12.3 The correspondence principle

The only reason we don’t usually notice diffraction of light in everyday life is that we don’t normally deal with objects that are comparable in size to a wavelength of visible light, which is about a millionth of a meter. Does this mean that wave optics contradicts ray optics, or that wave optics sometimes gives wrong results? No. If you hold three fingers out in the sunlight and cast a shadow with them, either wave optics or ray optics can be used to predict the straightforward result: a shadow pattern with two bright lines where the light has gone through the gaps between your fingers. Wave optics is a more general theory than ray optics, so in any case where ray optics is valid, the two theories will agree. This is an example of a general idea enunciated by the physicist Niels Bohr, called the *correspondence principle*: when flaws in a physical theory lead to the creation of a new and more general theory, the new theory must still agree with the old theory within its more restricted area of applicability. After all, a theory is only created as a way of describing



e / The bottom figure is simply a copy of the middle portion of the top one, scaled up by a factor of two. All the angles are the same. Physically, the angular pattern of the diffraction fringes can't be any different if we scale both λ and d by the same factor, leaving λ/d unchanged.

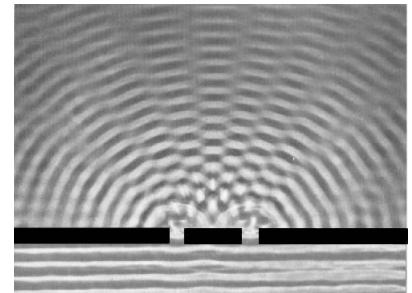
experimental observations. If the original theory had not worked in any cases at all, it would never have become accepted.

In the case of optics, the correspondence principle tells us that when λ/d is small, both the ray and the wave model of light must give approximately the same result. Suppose you spread your fingers and cast a shadow with them using a coherent light source. The quantity λ/d is about 10^{-4} , so the two models will agree very closely. (To be specific, the shadows of your fingers will be outlined by a series of light and dark fringes, but the angle subtended by a fringe will be on the order of 10^{-4} radians, so they will be too tiny to be visible.)

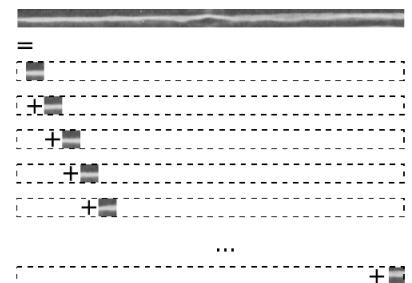
self-check A

What kind of wavelength would an electromagnetic wave have to have in order to diffract dramatically around your body? Does this contradict the correspondence principle?

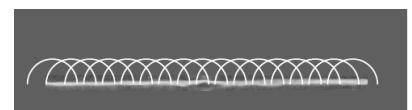
▷ Answer, p. 457



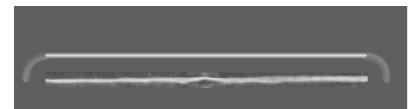
f / Double-slit diffraction.



g / A wavefront can be analyzed by the principle of superposition, breaking it down into many small parts.



h / If it was by itself, each of the parts would spread out as a circular ripple.



i / Adding up the ripples produces a new wavefront.

12.4 Huygens' principle

Returning to the example of double-slit diffraction, f, note the strong visual impression of two overlapping sets of concentric semi-circles. This is an example of *Huygens' principle*, named after a Dutch physicist and astronomer. (The first syllable rhymes with “boy.”) Huygens' principle states that any wavefront can be broken down into many small side-by-side wave peaks, g, which then spread out as circular ripples, h, and by the principle of superposition, the result of adding up these sets of ripples must give the same result as allowing the wave to propagate forward, i. In the case of sound or light waves, which propagate in three dimensions, the “ripples” are actually spherical rather than circular, but we can often imagine things in two dimensions for simplicity.

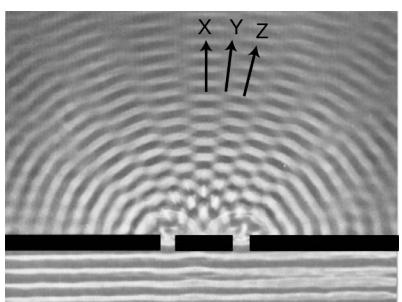
In double-slit diffraction the application of Huygens' principle is visually convincing: it is as though all the sets of ripples have been blocked except for two. It is a rather surprising mathematical fact, however, that Huygens' principle gives the right result in the case of an unobstructed linear wave, h and i. A theoretically infinite number of circular wave patterns somehow conspire to add together and produce the simple linear wave motion with which we are familiar.

Since Huygens' principle is equivalent to the principle of superposition, and superposition is a property of waves, what Huygens had created was essentially the first wave theory of light. However, he imagined light as a series of pulses, like hand claps, rather than as a sinusoidal wave.

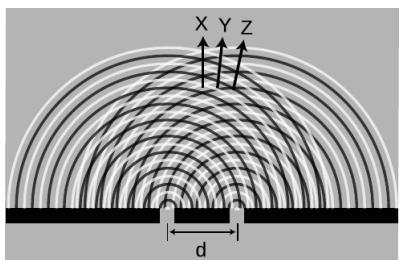
The history is interesting. Isaac Newton loved the atomic theory of matter so much that he searched enthusiastically for evidence that light was also made of tiny particles. The paths of his light particles



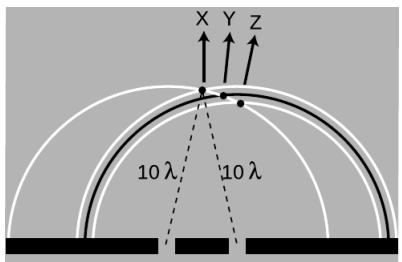
j / Thomas Young



k / Double-slit diffraction.



l / Use of Huygens' principle.



m / Constructive interference along the center-line.

would correspond to rays in our description; the only significant difference between a ray model and a particle model of light would occur if one could isolate individual particles and show that light had a “graininess” to it. Newton never did this, so although he thought of his model as a particle model, it is more accurate to say he was one of the builders of the ray model.

Almost all that was known about reflection and refraction of light could be interpreted equally well in terms of a particle model or a wave model, but Newton had one reason for strongly opposing Huygens’ wave theory. Newton knew that waves exhibited diffraction, but diffraction of light is difficult to observe, so Newton believed that light did not exhibit diffraction, and therefore must not be a wave. Although Newton’s criticisms were fair enough, the debate also took on the overtones of a nationalistic dispute between England and continental Europe, fueled by English resentment over Leibniz’s supposed plagiarism of Newton’s calculus. Newton wrote a book on optics, and his prestige and political prominence tended to discourage questioning of his model.

Thomas Young (1773-1829) was the person who finally, a hundred years later, did a careful search for wave interference effects with light and analyzed the results correctly. He observed double-slit diffraction of light as well as a variety of other diffraction effects, all of which showed that light exhibited wave interference effects, and that the wavelengths of visible light waves were extremely short. The crowning achievement was the demonstration by the experimentalist Heinrich Hertz and the theorist James Clerk Maxwell that light was an *electromagnetic* wave. Maxwell is said to have related his discovery to his wife one starry evening and told her that she was the only other person in the world who knew what starlight was.

12.5 Double-slit diffraction

Let’s now analyze double-slit diffraction, k, using Huygens’ principle. The most interesting question is how to compute the angles such as X and Z where the wave intensity is at a maximum, and the in-between angles like Y where it is minimized. Let’s measure all our angles with respect to the vertical center line of the figure, which was the original direction of propagation of the wave.

If we assume that the width of the slits is small (on the order of the wavelength of the wave or less), then we can imagine only a single set of Huygens ripples spreading out from each one, l. White lines represent peaks, black ones troughs. The only dimension of the diffracting slits that has any effect on the geometric pattern of the overlapping ripples is then the center-to-center distance, d , between the slits.

We know from our discussion of the scaling of diffraction that there must be some equation that relates an angle like θ_Z to the ratio λ/d ,

$$\frac{\lambda}{d} \leftrightarrow \theta_Z.$$

If the equation for θ_Z depended on some other expression such as $\lambda + d$ or λ^2/d , then it would change when we scaled λ and d by the same factor, which would violate what we know about the scaling of diffraction.

Along the central maximum line, X, we always have positive waves coinciding with positive ones and negative waves coinciding with negative ones. (I have arbitrarily chosen to take a snapshot of the pattern at a moment when the waves emerging from the slit are experiencing a positive peak.) The superposition of the two sets of ripples therefore results in a doubling of the wave amplitude along this line. There is constructive interference. This is easy to explain, because by symmetry, each wave has had to travel an equal number of wavelengths to get from its slit to the center line, m: Because both sets of ripples have ten wavelengths to cover in order to reach the point along direction X, they will be in step when they get there.

At the point along direction Y shown in the same figure, one wave has traveled ten wavelengths, and is therefore at a positive extreme, but the other has traveled only nine and a half wavelengths, so it is at a negative extreme. There is perfect cancellation, so points along this line experience no wave motion.

But the distance traveled does not have to be equal in order to get constructive interference. At the point along direction Z, one wave has gone nine wavelengths and the other ten. They are both at a positive extreme.

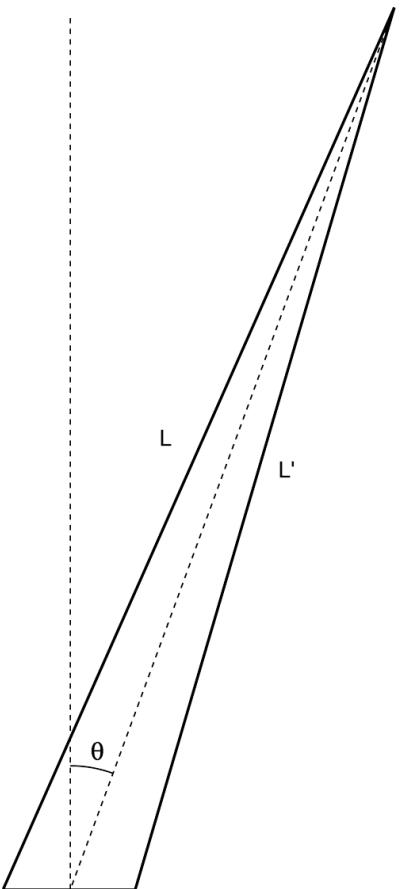
self-check B

At a point half a wavelength below the point marked along direction X, carry out a similar analysis. ▷ Answer, p. 457

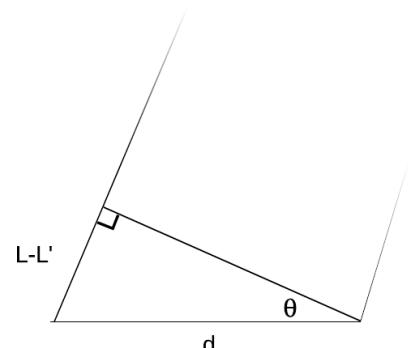
To summarize, we will have perfect constructive interference at any point where the distance to one slit differs from the distance to the other slit by an integer number of wavelengths. Perfect destructive interference will occur when the number of wavelengths of path length difference equals an integer plus a half.

Now we are ready to find the equation that predicts the angles of the maxima and minima. The waves travel different distances to get to the same point in space, n. We need to find whether the waves are in phase (in step) or out of phase at this point in order to predict whether there will be constructive interference, destructive interference, or something in between.

One of our basic assumptions in this chapter is that we will only be dealing with the diffracted wave in regions very far away from the



n / The waves travel distances L and L' from the two slits to get to the same point in space, at an angle θ from the center line.



o / A close-up view of figure n, showing how the path length difference $L - L'$ is related to d and to the angle θ .

object that diffracts it, so the triangle is long and skinny. Most real-world examples with diffraction of light, in fact, would have triangles with even skinnier proportions than this one. The two long sides are therefore very nearly parallel, and we are justified in drawing the right triangle shown in figure o, labeling one leg of the right triangle as the difference in path length, $L - L'$, and labeling the acute angle as θ . (In reality this angle is a tiny bit greater than the one labeled θ in figure n.)

The difference in path length is related to d and θ by the equation

$$\frac{L - L'}{d} = \sin \theta.$$

Constructive interference will result in a maximum at angles for which $L - L'$ is an integer number of wavelengths,

$$L - L' = m\lambda. \quad [\text{condition for a maximum;} \\ m \text{ is an integer}]$$

Here m equals 0 for the central maximum, -1 for the first maximum to its left, $+2$ for the second maximum on the right, etc. Putting all the ingredients together, we find $m\lambda/d = \sin \theta$, or

$$\frac{\lambda}{d} = \frac{\sin \theta}{m}. \quad [\text{condition for a maximum;} \\ m \text{ is an integer}]$$

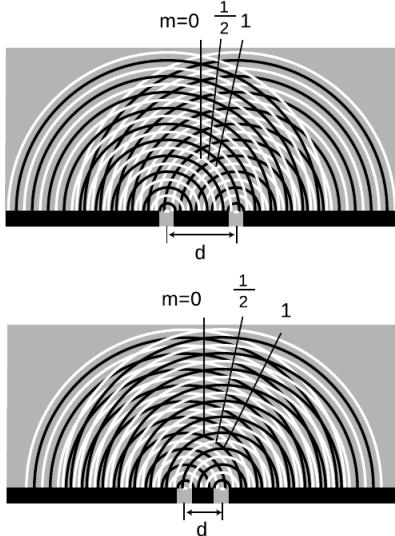
Similarly, the condition for a minimum is

$$\frac{\lambda}{d} = \frac{\sin \theta}{m}. \quad [\text{condition for a minimum;} \\ m \text{ is an integer plus } 1/2]$$

That is, the minima are about halfway between the maxima.

As expected based on scaling, this equation relates angles to the unitless ratio λ/d . Alternatively, we could say that we have proven the scaling property in the special case of double-slit diffraction. It was inevitable that the result would have these scaling properties, since the whole proof was geometric, and would have been equally valid when enlarged or reduced on a photocopying machine!

Counterintuitively, this means that a diffracting object with smaller dimensions produces a bigger diffraction pattern, p.



p / Cutting d in half doubles the angles of the diffraction fringes.

Double-slit diffraction of blue and red light

example 1

Blue light has a shorter wavelength than red. For a given double-slit spacing d , the smaller value of λ/d for leads to smaller values of $\sin \theta$, and therefore to a more closely spaced set of diffraction fringes, shown in figure q.

The correspondence principle

example 2

Let's also consider how the equations for double-slit diffraction relate to the correspondence principle. When the ratio λ/d is very small, we should recover the case of simple ray optics. Now if λ/d is small, $\sin \theta$ must be small as well, and the spacing between the diffraction fringes will be small as well. Although we have not proven it, the central fringe is always the brightest, and the fringes get dimmer and dimmer as we go farther from it. For small values of λ/d , the part of the diffraction pattern that is bright enough to be detectable covers only a small range of angles. This is exactly what we would expect from ray optics: the rays passing through the two slits would remain parallel, and would continue moving in the $\theta = 0$ direction. (In fact there would be images of the two separate slits on the screen, but our analysis was all in terms of angles, so we should not expect it to address the issue of whether there is structure within a set of rays that are all traveling in the $\theta = 0$ direction.)

Spacing of the fringes at small angles

example 3

At small angles, we can use the approximation $\sin \theta \approx \theta$, which is valid if θ is measured in radians. The equation for double-slit diffraction becomes simply

$$\frac{\lambda}{d} = \frac{\theta}{m},$$

which can be solved for θ to give

$$\theta = \frac{m\lambda}{d}.$$

The difference in angle between successive fringes is the change in θ that results from changing m by plus or minus one,

$$\Delta\theta = \frac{\lambda}{d}.$$

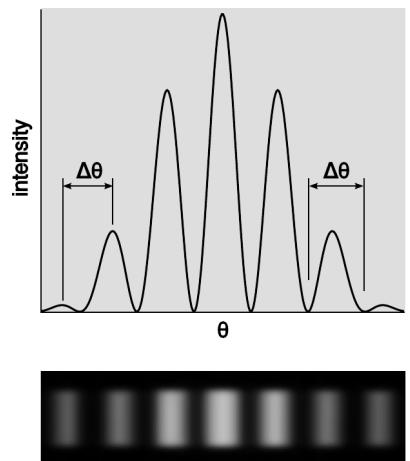
For example, if we write θ_7 for the angle of the seventh bright fringe on one side of the central maximum and θ_8 for the neighboring one, we have

$$\begin{aligned}\theta_8 - \theta_7 &= \frac{8\lambda}{d} - \frac{7\lambda}{d} \\ &= \frac{\lambda}{d},\end{aligned}$$

and similarly for any other neighboring pair of fringes.



q / Double-slit diffraction patterns of long-wavelength red light (top) and short-wavelength blue light (bottom).



r / Interpretation of the angular spacing $\Delta\theta$ in example 3. It can be defined either from maximum to maximum or from minimum to minimum. Either way, the result is the same. It does not make sense to try to interpret $\Delta\theta$ as the width of a fringe; one can see from the graph and from the image below that it is not obvious either that such a thing is well defined or that it would be the same for all fringes.

Although the equation $\lambda/d = \sin \theta/m$ is only valid for a double slit, it is still a guide to our thinking even if we are observing diffraction of light by a virus or a flea's leg: it is always true that

- (1) large values of λ/d lead to a broad diffraction pattern, and
- (2) diffraction patterns are repetitive.

In many cases the equation looks just like $\lambda/d = \sin \theta/m$ but with an extra numerical factor thrown in, and with d interpreted as some other dimension of the object, e.g., the diameter of a piece of wire.

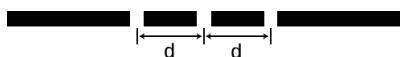
12.6 Repetition

Suppose we replace a double slit with a triple slit, s. We can think of this as a third repetition of the structures that were present in the double slit. Will this device be an improvement over the double slit for any practical reasons?

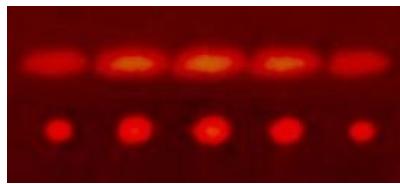
The answer is yes, as can be shown using figure u. For ease of visualization, I have violated our usual rule of only considering points very far from the diffracting object. The scale of the drawing is such that a wavelength is one cm. In u/1, all three waves travel an integer number of wavelengths to reach the same point, so there is a bright central spot, as we would expect from our experience with the double slit. In figure u/2, we show the path lengths to a new point. This point is farther from slit A by a quarter of a wavelength, and correspondingly closer to slit C. The distance from slit B has hardly changed at all. Because the paths traveled from slits A and C differ by half a wavelength, there will be perfect destructive interference between these two waves. There is still some uncanceled wave intensity because of slit B, but the amplitude will be three times less than in figure u/1, resulting in a factor of 9 decrease in brightness. Thus, by moving off to the right a little, we have gone from the bright central maximum to a point that is quite dark.

Now let's compare with what would have happened if slit C had been covered, creating a plain old double slit. The waves coming from slits A and B would have been out of phase by 0.23 wavelengths, but this would not have caused very severe interference. The point in figure u/2 would have been quite brightly lit up.

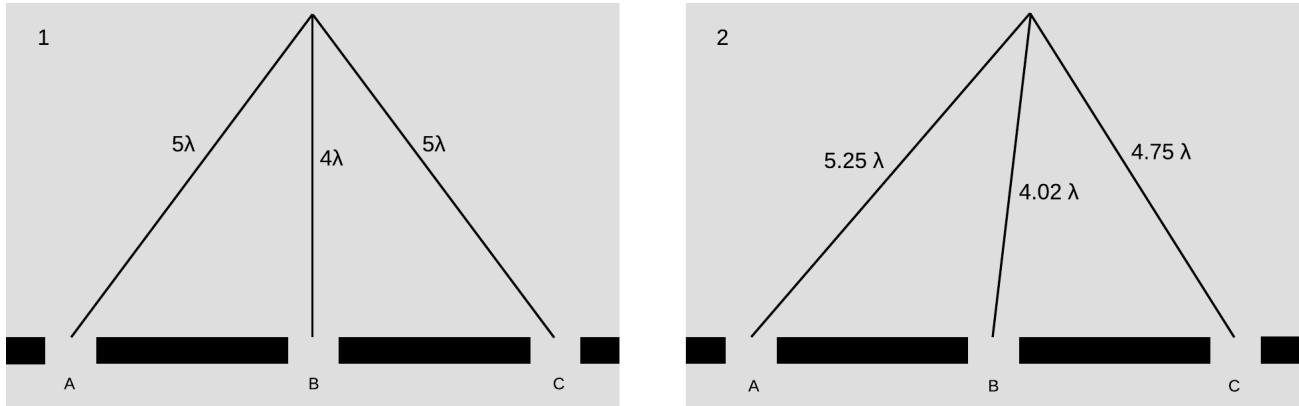
To summarize, we have found that adding a third slit narrows down the central fringe dramatically. The same is true for all the



s / A triple slit.



t / A double-slit diffraction pattern (top), and a pattern made by five slits (bottom).



u / 1. There is a bright central maximum. 2. At this point just off the central maximum, the path lengths traveled by the three waves have changed.

other fringes as well, and since the same amount of energy is concentrated in narrower diffraction fringes, each fringe is brighter and easier to see, t.

This is an example of a more general fact about diffraction: if some feature of the diffracting object is repeated, the locations of the maxima and minima are unchanged, but they become narrower.

Taking this reasoning to its logical conclusion, a diffracting object with thousands of slits would produce extremely narrow fringes. Such an object is called a diffraction grating.

12.7 Single-slit diffraction

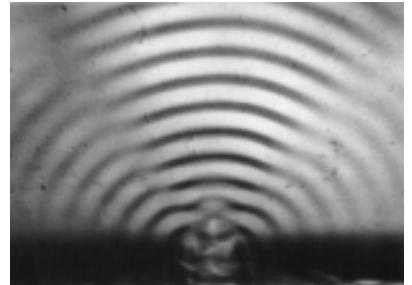
If we use only a single slit, is there diffraction? If the slit is not wide compared to a wavelength of light, then we can approximate its behavior by using only a single set of Huygens ripples. There are no other sets of ripples to add to it, so there are no constructive or destructive interference effects, and no maxima or minima. The result will be a uniform spherical wave of light spreading out in all directions, like what we would expect from a tiny lightbulb. We could call this a diffraction pattern, but it is a completely featureless one, and it could not be used, for instance, to determine the wavelength of the light, as other diffraction patterns could.

All of this, however, assumes that the slit is narrow compared to a wavelength of light. If, on the other hand, the slit is broader, there will indeed be interference among the sets of ripples spreading out from various points along the opening. Figure v shows an example with water waves, and figure w with light.

self-check C

How does the wavelength of the waves compare with the width of the slit in figure v?

▷ Answer, p. 457



v / Single-slit diffraction of water waves.



w / Single-slit diffraction of red light. Note the double width of the central maximum.



x / A pretty good simulation of the single-slit pattern of figure v, made by using three motors to produce overlapping ripples from three neighboring points in the water.

We will not go into the details of the analysis of single-slit diffraction, but let us see how its properties can be related to the general things we've learned about diffraction. We know based on scaling arguments that the angular sizes of features in the diffraction pattern must be related to the wavelength and the width, a , of the slit by some relationship of the form

$$\frac{\lambda}{a} \leftrightarrow \theta.$$

This is indeed true, and for instance the angle between the maximum of the central fringe and the maximum of the next fringe on one side equals $1.5\lambda/a$. Scaling arguments will never produce factors such as the 1.5, but they tell us that the answer must involve λ/a , so all the familiar qualitative facts are true. For instance, shorter-wavelength light will produce a more closely spaced diffraction pattern.

An important scientific example of single-slit diffraction is in telescopes. Images of individual stars, as in figure y, are a good way to examine diffraction effects, because all stars except the sun are so far away that no telescope, even at the highest magnification, can image their disks or surface features. Thus any features of a star's image must be due purely to optical effects such as diffraction. A prominent cross appears around the brightest star, and dimmer ones surround the dimmer stars. Something like this is seen in most telescope photos, and indicates that inside the tube of the telescope there were two perpendicular struts or supports. Light diffracted around these struts. You might think that diffraction could be eliminated entirely by getting rid of all obstructions in the tube, but the circles around the stars are diffraction effects arising from single-slit diffraction at the mouth of the telescope's tube! (Actually we have not even talked about diffraction through a circular opening, but the idea is the same.) Since the angular sizes of the diffracted images depend on λ/a , the only way to improve the resolution of the images is to increase the diameter, a , of the tube. This is one of the main reasons (in addition to light-gathering power) why the best telescopes must be very large in diameter.

self-check D

What would this imply about radio telescopes as compared with visible-light telescopes?

▷ Answer, p. 457

Double-slit diffraction is easier to understand conceptually than single-slit diffraction, but if you do a double-slit diffraction experiment in real life, you are likely to encounter a complicated pattern like figure aa/1, rather than the simpler one, 2, you were expecting. This is because the slits are fairly big compared to the wavelength of the light being used. We really have two different distances in our pair of slits: d , the distance between the slits, and w , the width

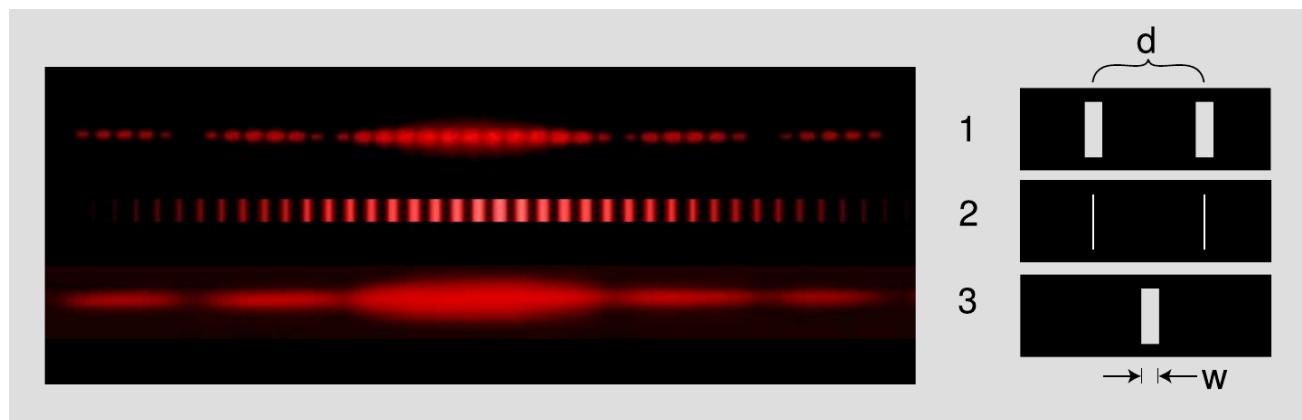


y / An image of the Pleiades star cluster. The circular rings around the bright stars are due to single-slit diffraction at the mouth of the telescope's tube.



z / A radio telescope.

of each slit. Remember that smaller distances on the object the light diffracts around correspond to larger features of the diffraction pattern. The pattern 1 thus has two spacings in it: a short spacing corresponding to the large distance d , and a long spacing that relates to the small dimension w .



aa / 1. A diffraction pattern formed by a real double slit. The width of each slit is fairly big compared to the wavelength of the light. This is a real photo. 2. This idealized pattern is not likely to occur in real life. To get it, you would need each slit to be so narrow that its width was comparable to the wavelength of the light, but that's not usually possible. This is not a real photo. 3. A real photo of a single-slit diffraction pattern caused by a slit whose width is the same as the widths of the slits used to make the top pattern.

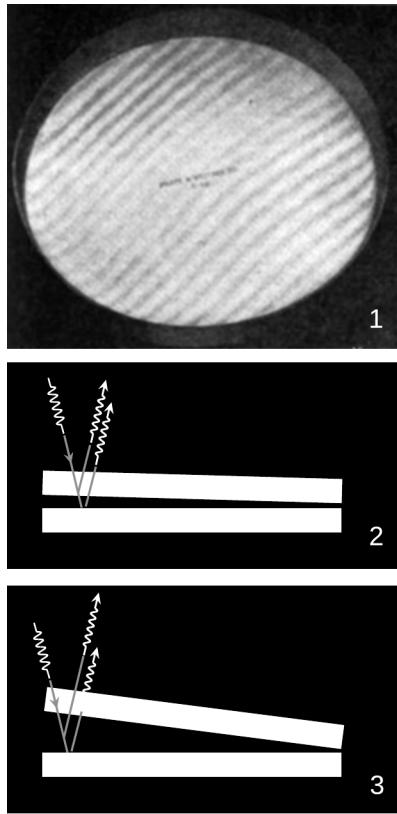
Discussion question

A Why is it optically impossible for bacteria to evolve eyes that use visible light to form images?

12.8 Coherence

Up until now, we have avoided too much detailed discussion of two facts that sometimes make interference and diffraction effects unobservable, and that historically made them more difficult to discover. First there is the fact that white light is a mixture of all the visible wavelengths. This is why, for example, the thin-film interference pattern of a soap bubble looks like a rainbow. To simplify things, we need a source of light that is monochromatic, i.e., contains only a single wavelength or a small range of wavelengths. We could do this either by filtering a white light source or by using a source of light that is intrinsically monochromatic, such as a laser or some gas discharge tubes.

But even with a monochromatic light source, we encounter a separate issue, which is that most light sources do not emit light waves that are perfect, infinitely long sine waves. Sunlight and candlelight, for example, can be thought of as being composed of separate little spurts of light, referred to as wave packets or wave trains. Each



ab / 1. Interference in an air wedge. 2. Side view. 3. If the wedge is thicker than the coherence length of the light, the interference pattern disappears.

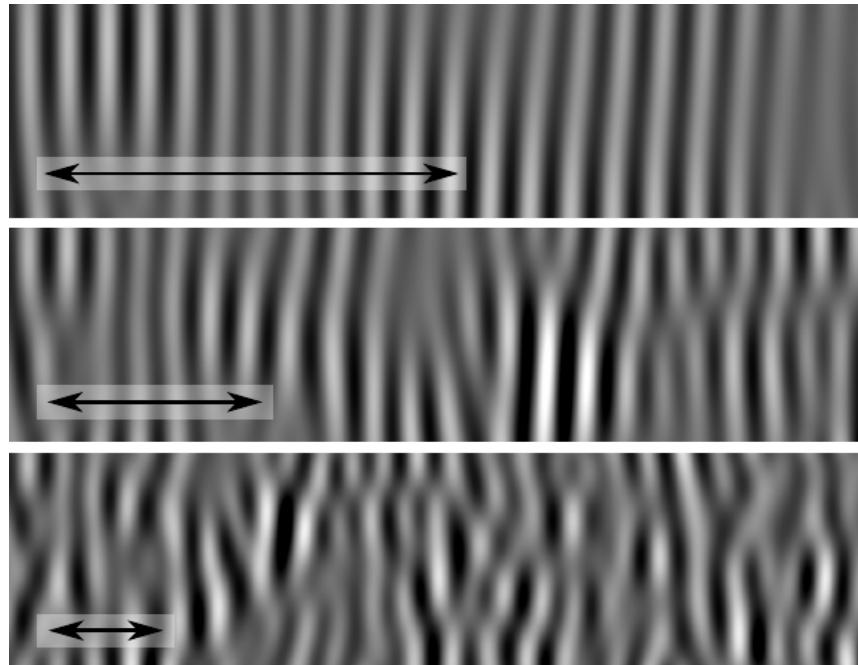
wave packet is emitted by a separate atom of the gas. It contains some number of wavelengths, and it has no fixed phase relationship to any other wave packet. The wave trains emitted by a laser are much longer, but still not infinitely long.

As an example of an experiment that can show these effects, figure ab/1 shows a thin-film interference pattern created by the air wedge between two pieces of very flat glass, where the top piece is placed at a very small angle relative to the bottom one, ab/2. The phase relationship between the two reflected waves is determined by the extra distance traveled by the ray that is reflected by the bottom plate (as well as the fact that one of the two reflections will be inverting).

If the angle is opened up too much, ab/3, we will no longer see fringes where the air layer is too thick. This is because the incident wave train has only a certain length, and the extra distance traveled is now so great that the two reflected wave trains no longer overlap in space. In general, if the incident wave trains are n wavelengths long, then we can see at most n bright and n dark fringes. The fact that about 18 fringes are visible in ab/1 shows that the light source used (let's say a sodium gas discharge tube) made wave trains at least 18 wavelengths in length.

In real-world light sources, the wave packets may not be as neat and tidy as the ones in figure ab. They may not look like sine waves with clean cut-offs at the ends, and they may overlap one another. The result will look more like the examples in figure ac. Such a wave pattern has a property called its coherence length L .

ac / Waves with three different coherence lengths, indicated by the arrows. Note that although there is a superficial similarity between these pictures and figure ab/1, they represent completely different things. Figure ab/1 is an actual photograph of interference fringes, whose brightness is proportional to the square of the amplitude. This figure is a picture of the wave's amplitude, not the squared amplitude, and is analogous to the little sine waves in ab/2 and ab/3. These are waves that are *traveling* across the page at the speed of light.



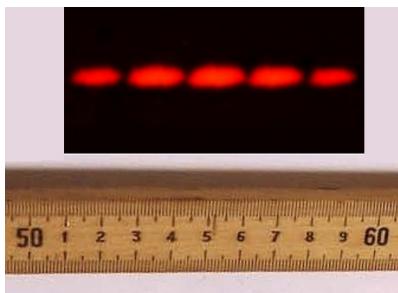
On scales small compared to L , the wave appears like a perfect sine wave. On scales large compared to L , we lose all phase correlations. For example, the middle wave in figure ac has $L \approx 5\lambda$. If we pick two points within this wave separated by a distance of λ in the left-right direction, they are likely to be very nearly in phase. But if the separation is 20λ , approximately the width of the entire figure, the phase relationship is essentially random. If the light comes from a flame or a gas discharge tube, then this lack of a phase relationship would be because the parts of the wave at these large separations from one another probably originated from different atoms in the source.

Problems

Key

- ✓ A computerized answer check is available online.
- ★ A difficult problem.

1 The figure shows a diffraction pattern made by a double slit, along with an image of a meter stick to show the scale. Sketch the diffraction pattern from the figure on your paper. Now consider the four variables in the equation $\lambda/d = \sin \theta/m$. Which of these are the same for all five fringes, and which are different for each fringe? Which variable would you naturally use in order to label which fringe was which? Label the fringes on your sketch using the values of that variable.



Problems 1 and 4.

2 Match gratings A-C with the diffraction patterns 1-3 that they produce. Explain.

A



1



B



2



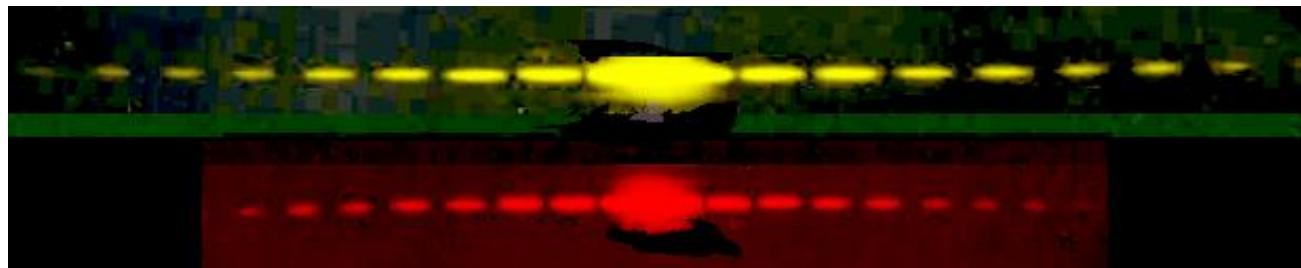
C



3



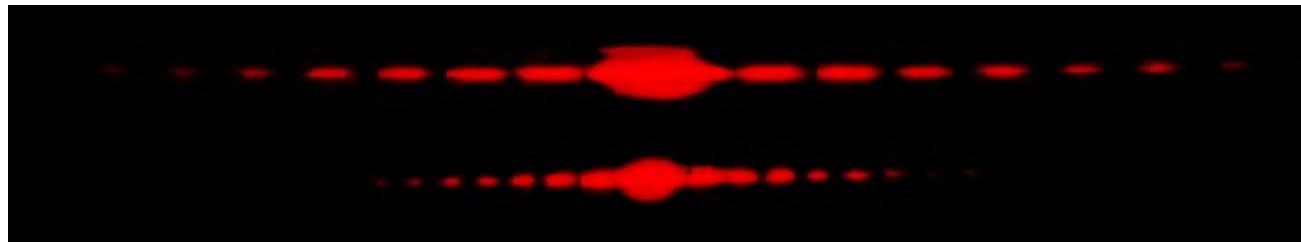
- 3** The figure below shows two diffraction patterns. The top one was made with yellow light, and the bottom one with red. Could the slits used to make the two patterns have been the same?



- 4** The figure on p. 276 shows a diffraction pattern made by a double slit, along with an image of a meter stick to show the scale. The slits were 146 cm away from the screen on which the diffraction pattern was projected. The spacing of the slits was 0.050 mm. What was the wavelength of the light? ✓

- 5** Why would blue or violet light be the best for microscopy?
▷ Solution, p. 453

- 6** The figure below shows two diffraction patterns, both made with the same wavelength of red light. (a) What type of slits made the patterns? Is it a single slit, double slits, or something else? Explain. (b) Compare the dimensions of the slits used to make the top and bottom pattern. Give a numerical ratio, and state which way the ratio is, i.e., which slit pattern was the larger one. Explain.



▷ Solution, p. 453

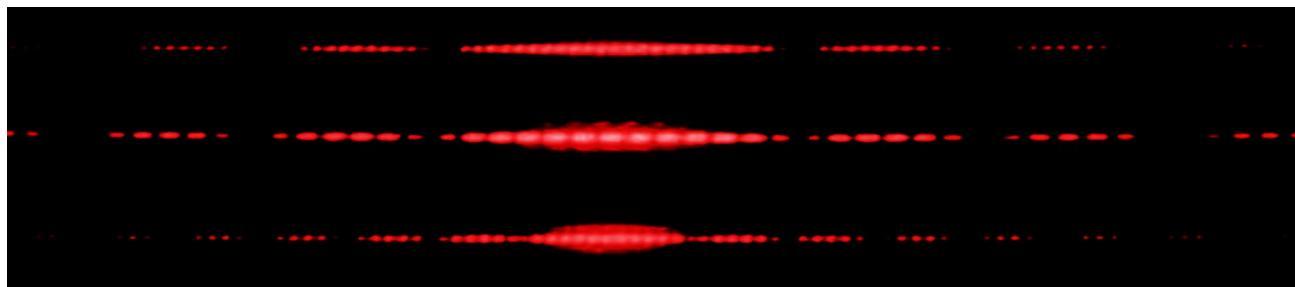
- 7** When white light passes through a diffraction grating, what is the smallest value of m for which the visible spectrum of order m overlaps the next one, of order $m + 1$? (The visible spectrum runs from about 400 nm to about 700 nm.)

Problem 8. This image of the Pleiades star cluster shows haloes around the stars due to the wave nature of light.



8 For star images such as the ones in figure y, estimate the angular width of the diffraction spot due to diffraction at the mouth of the telescope. Assume a telescope with a diameter of 10 meters (the largest currently in existence), and light with a wavelength in the middle of the visible range. Compare with the actual angular size of a star of diameter 10^9 m seen from a distance of 10^{17} m. What does this tell you? ▷ Solution, p. 453

9 The figure below shows three diffraction patterns. All were made under identical conditions, except that a different set of double slits was used for each one. The slits used to make the top pattern had a center-to-center separation $d = 0.50$ mm, and each slit was $w = 0.04$ mm wide. (a) Determine d and w for the slits used to make the pattern in the middle. (b) Do the same for the slits used to make the bottom pattern.



▷ Solution, p. 453

10 The beam of a laser passes through a diffraction grating, fans out, and illuminates a wall that is perpendicular to the original beam, lying at a distance of 2.0 m from the grating. The beam is produced by a helium-neon laser, and has a wavelength of 694.3 nm. The grating has 2000 lines per centimeter. (a) What is the distance on the wall between the central maximum and the maxima immediately to its right and left? (b) How much does your answer change when you use the small-angle approximations $\theta \approx \sin \theta \approx \tan \theta$? ✓

11 Ultrasound, i.e., sound waves with frequencies too high to be audible, can be used for imaging fetuses in the womb or for breaking up kidney stones so that they can be eliminated by the body. Consider the latter application. Lenses can be built to focus sound waves, but because the wavelength of the sound is not all that small compared to the diameter of the lens, the sound will not be concentrated exactly at the geometrical focal point. Instead, a diffraction pattern will be created with an intense central spot surrounded by fainter rings. About 85% of the power is concentrated within the central spot. The angle of the first minimum (surrounding the central spot) is given by $\sin \theta = \lambda/b$, where b is the diameter of the lens. This is similar to the corresponding equation for a single slit, but with a factor of 1.22 in front which arises from the circular shape of the aperture. Let the distance from the lens to the patient's kidney stone be $L = 20$ cm. You will want $f > 20$ kHz, so that the sound is inaudible. Find values of b and f that would result in a usable design, where the central spot is small enough to lie within a kidney stone 1 cm in diameter.

12 Under what circumstances could one get a mathematically undefined result by solving the double-slit diffraction equation for θ ? Give a physical interpretation of what would actually be observed.

▷ Solution, p. 453

13 When ultrasound is used for medical imaging, the frequency may be as high as 5-20 MHz. Another medical application of ultrasound is for therapeutic heating of tissues inside the body; here, the frequency is typically 1-3 MHz. What fundamental physical reasons could you suggest for the use of higher frequencies for imaging?

Exercise 12A: Two-source interference

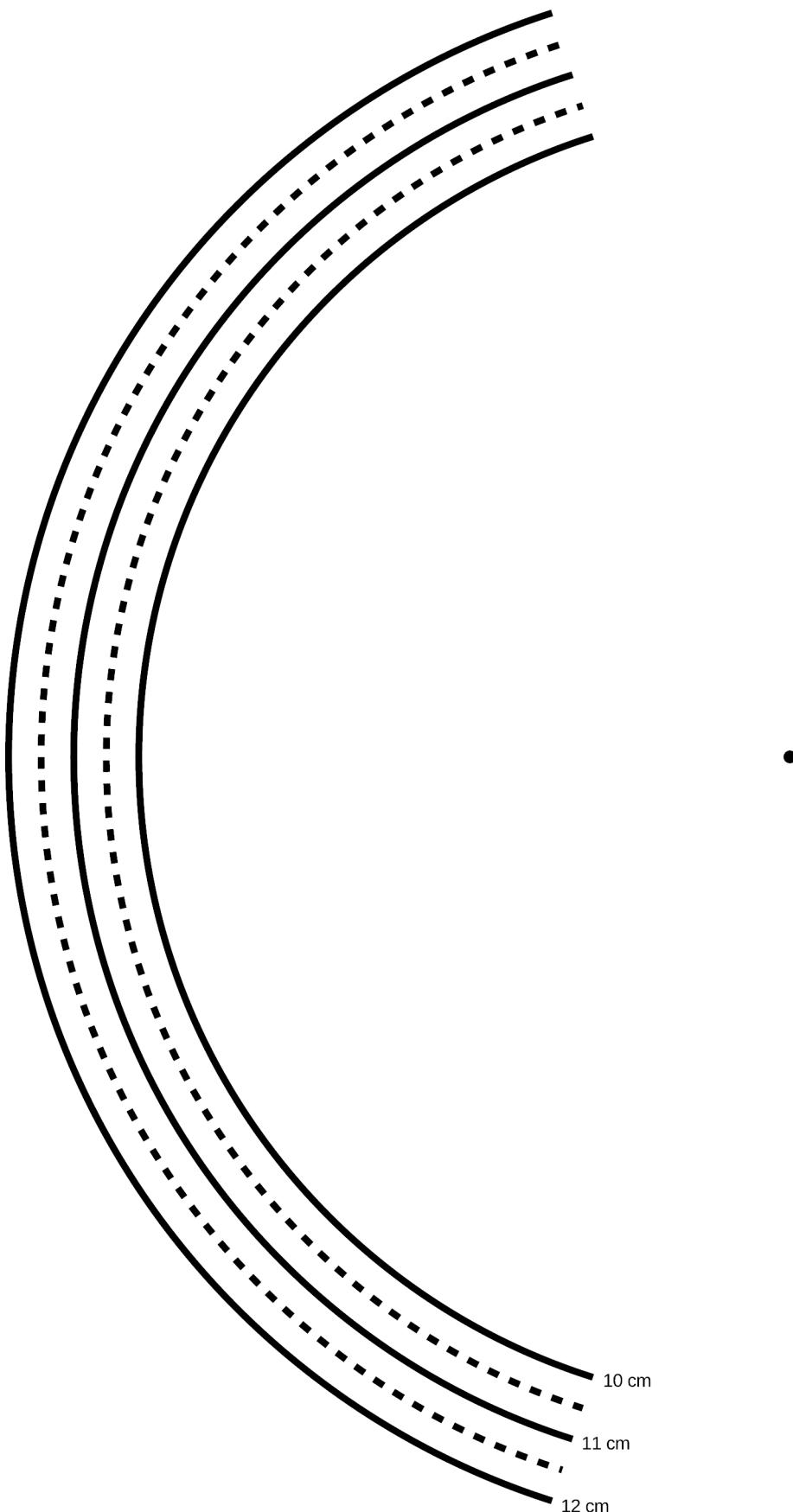
1. Two sources separated by a distance $d = 2$ cm make circular ripples with a wavelength of $\lambda = 1$ cm. On a piece of paper, make a life-size drawing of the two sources in the default setup, and locate the following points:

- A. The point that is 10 wavelengths from source #1 and 10 wavelengths from source #2.
- B. The point that is 10.5 wavelengths from #1 and 10.5 from #2.
- C. The point that is 11 wavelengths from #1 and 11 from #2.
- D. The point that is 10 wavelengths from #1 and 10.5 from #2.
- E. The point that is 11 wavelengths from #1 and 11.5 from #2.
- F. The point that is 10 wavelengths from #1 and 11 from #2.
- G. The point that is 11 wavelengths from #1 and 12 from #2.

You can do this either using a compass or by putting the next page under your paper and tracing. It is not necessary to trace all the arcs completely, and doing so is unnecessarily time-consuming; you can fairly easily estimate where these points would lie, and just trace arcs long enough to find the relevant intersections.

What do these points correspond to in the real wave pattern?

- 2. Make a fresh copy of your drawing, showing only point F and the two sources, which form a long, skinny triangle. Now suppose you were to change the setup by doubling d , while leaving λ the same. It's easiest to understand what's happening on the drawing if you move both sources outward, keeping the center fixed. Based on your drawing, what will happen to the position of point F when you double d ? How has the angle of point F changed?
- 3. What would happen if you doubled *both* λ and d compared to the standard setup?_____
- 4. Combining the ideas from parts 2 and 3, what do you think would happen to your angles if, starting from the standard setup, you doubled λ while leaving d the same?_____
- 5. Suppose λ was a millionth of a centimeter, while d was still as in the standard setup. What would happen to the angles? What does this tell you about observing diffraction of light?



Exercise 12B: Single-slit interference

Equipment:

rulers

computer with web browser

The following page is a diagram of a single slit and a screen onto which its diffraction pattern is projected. The class will make a numerical prediction of the intensity of the pattern at the different points on the screen. Each group will be responsible for calculating the intensity at one of the points. (Either 11 groups or six will work nicely – in the latter case, only points a, c, e, g, i, and k are used.) The idea is to break up the wavefront in the mouth of the slit into nine parts, each of which is assumed to radiate semicircular ripples as in Huygens' principle. The wavelength of the wave is 1 cm, and we assume for simplicity that each set of ripples has an amplitude of 1 unit when it reaches the screen.

1. For simplicity, let's imagine that we were only to use two sets of ripples rather than nine. You could measure the distance from each of the two points inside the slit to your point on the screen. Suppose the distances were both 25.0 cm. What would be the amplitude of the superimposed waves at this point on the screen?

Suppose one distance was 24.0 cm and the other was 25.0 cm. What would happen?

What if one was 24.0 cm and the other was 26.0 cm?

What if one was 24.5 cm and the other was 25.0 cm?

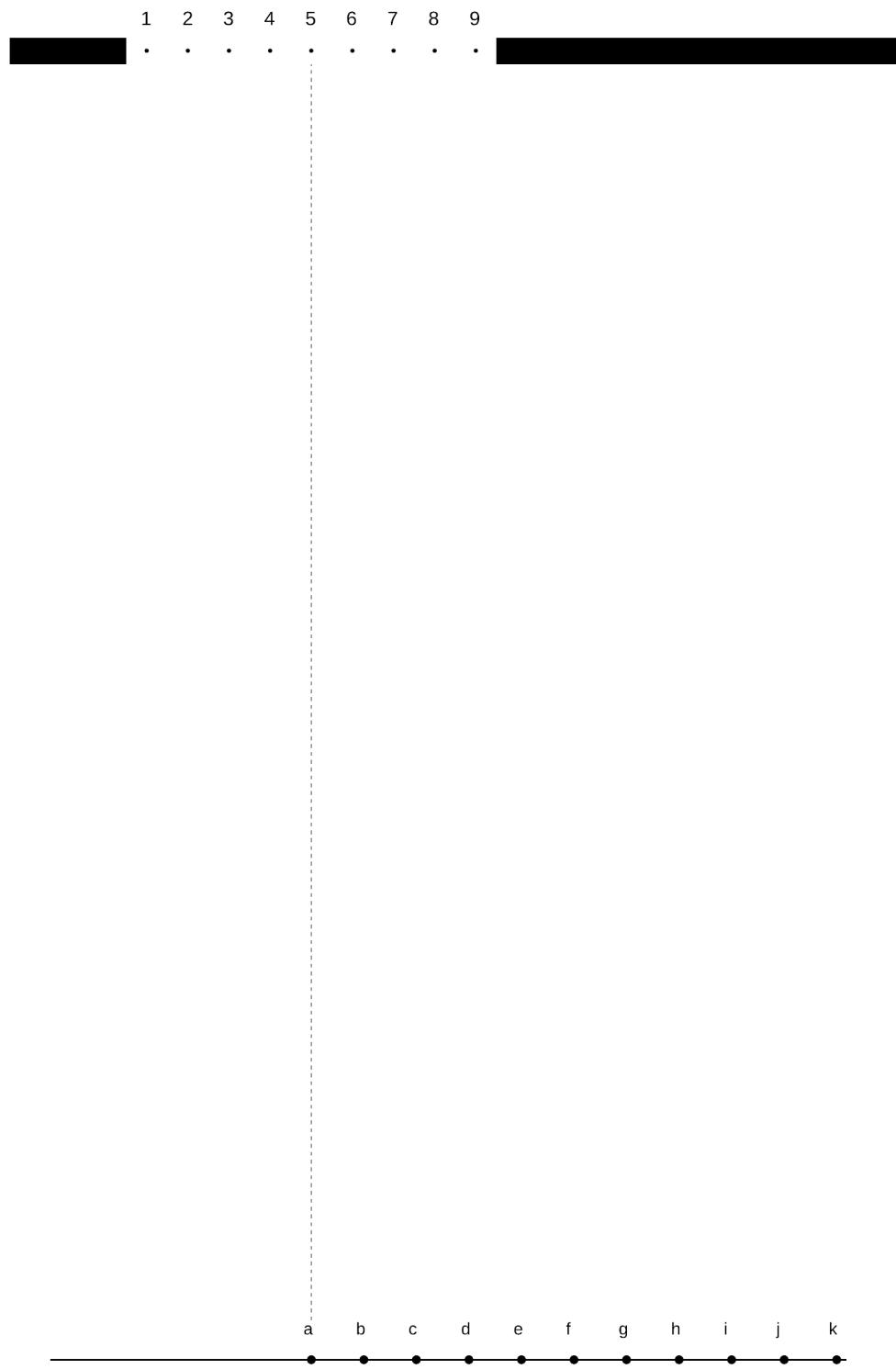
In general, what combinations of distances will lead to completely destructive and completely constructive interference?

Can you estimate the answer in the case where the distances are 24.7 and 25.0 cm?

2. Although it is possible to calculate mathematically the amplitude of the sine wave that results from superimposing two sine waves with an arbitrary phase difference between them, the algebra is rather laborious, and it becomes even more tedious when we have more than two waves to superimpose. Instead, one can simply use a computer spreadsheet or some other computer program to add up the sine waves numerically at a series of points covering one complete cycle. This is what we will actually do. You just need to enter the relevant data into the computer, then examine the results and pick off the amplitude from the resulting list of numbers. You can run the software through a web interface at <http://lightandmatter.com/cgi-bin/diffraction1.cgi>.

3. Measure all nine distances to your group's point on the screen, and write them on the board – that way everyone can see everyone else's data, and the class can try to make sense of why the results came out the way they did. Determine the amplitude of the combined wave, and write it on the board as well.

The class will discuss why the results came out the way they did.



The microscopic description of matter and quantum physics

Chapter 13

The atom and the nucleus

In chapters 7 and 9, we have used a microscopic description of matter to find out things about thermodynamics. It's remarkable that one can get so far with almost no detailed description of what matter is actually made of at the microscopic level. For example, we were able to deduce the heat capacities of solids and gases (example 2, p. 174), without even knowing anything at all about their constituent particles (atoms!). This is a good-news/bad-news situation. It's good that we can figure out all kinds of things like the heat capacity of copper without even having to know that copper is made out of atoms that have certain specific properties. But if we're actually trying to *find out* about what copper is made out of, then it's unfortunate that we can't find out much from its thermodynamic properties.

There does come a point, though, at which we need to know what's really going on at the atomic and subatomic level. In the history of technology, this coincided with the creation of radio and the exploitation of nuclear power.

13.1 The electrical nature of matter and quantization of charge

In your course on electricity and magnetism, you have seen evidence that all matter contains electrically charged parts. For example, copper is an electrical conductor, and we interpret that as evidence that it contains positively and negatively charged stuff, with at least some of that stuff being free to move. Furthermore, there is no such thing as a form of matter that is a perfect insulator. Most aren't as good as copper, but all can conduct electricity to some extent. This tells us that all matter contains charged stuff.

In 1909, Robert Millikan and coworkers published experimental results showing that when he prepared tiny droplets of oil and manipulated them under a microscope with electric fields, their electric charge seemed to come only in integer multiples of a certain amount. We notate this basic amount of charge as e and refer to it to as the fundamental charge. Millikan is now known to have fudged his data, and his result for e is statistically inconsistent with the currently accepted value, $e = 1.602 \times 10^{-19}$ C. Quantization of charge suggests that the charged *stuff* inside matter actually consists of charged

particles.

Today, the standard model of particle physics includes particles called quarks, which have fractional charges $\pm(1/3)e$ and $\pm(2/3)e$. However, single quarks are never observed, only clusters of them, and the clusters always have charges that are integer multiples of e .

We summarize these facts by saying that charge is “quantized” in units of e . Similarly, money in the US is quantized in units of cents, and discerning music listeners bewail the use of software in the recording studio that quantizes rhythm, forcing notes to land exactly on the beat rather than allowing the kinds of creative variation that used to be common in popular music.

Sometimes we will be casual and say, for example, that a proton has “one unit of charge,” or even “a charge of one,” but this means $1e$, not one coulomb.

If you mix baking soda and vinegar to get a fizzy chemical reaction, you don’t really care that the number of molecules is an integer. The chemicals are, for all practical purposes, continuous fluids, because the number of molecules is so large. Similarly, quantization of charge has no consequences for many electrical circuits, and the charge flowing through a wire acts like a continuous substance. In the SI, this is expressed by the fact that e is a very small number when measured in practical units of coulombs.

13.2 The electron

13.2.1 Cathode rays

Nineteenth-century physicists spent a lot of time trying to come up with wild, random ways to play with electricity. The best experiments of this kind were the ones that made big sparks or pretty colors of light.

One such parlor trick was the cathode ray. To produce it, you first had to hire a good glassblower and find a good vacuum pump. The glassblower would create a hollow tube and embed two pieces of metal in it, called the electrodes, which were connected to the outside via metal wires passing through the glass. Before letting him seal up the whole tube, you would hook it up to a vacuum pump, and spend several hours huffing and puffing away at the pump’s hand crank to get a good vacuum inside. Then, while you were still pumping on the tube, the glassblower would melt the glass and seal the whole thing shut. Finally, you would put a large amount of positive charge on one wire and a large amount of negative charge on the other. Metals have the property of letting charge move through them easily, so the charge deposited on one of the wires would quickly spread out because of the repulsion of each part of it for every other part. This spreading-out process would result in nearly

all the charge ending up in the electrodes, where there is more room to spread out than there is in the wire. For obscure historical reasons a negative electrode is called a cathode and a positive one is an anode.

Figure a shows the light-emitting stream that was observed. If, as shown in this figure, a hole was made in the anode, the beam would extend on through the hole until it hit the glass. Drilling a hole in the cathode, however would not result in any beam coming out on the left side, and this indicated that the stuff, whatever it was, was coming from the cathode. The rays were therefore christened “cathode rays.” (The terminology is still used today in the term “cathode ray tube” or “CRT” for the picture tube of a TV or computer monitor.)

13.2.2 Were cathode rays a form of light, or of matter?

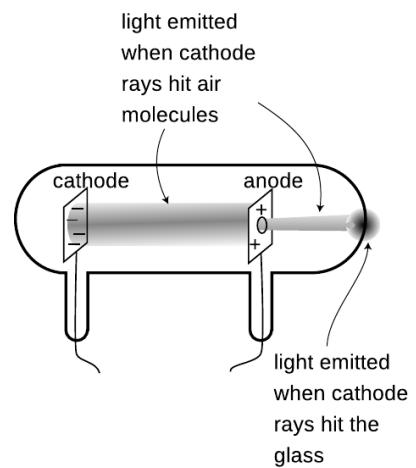
Were cathode rays a form of light, or matter? At first no one really cared what they were, but as their scientific importance became more apparent, the light-versus-matter issue turned into a controversy along nationalistic lines, with the Germans advocating light and the English holding out for matter. The supporters of the material interpretation imagined the rays as consisting of a stream of atoms ripped from the substance of the cathode.

One of our defining characteristics of matter is that material objects cannot pass through each other. Experiments showed that cathode rays could penetrate at least some small thickness of matter, such as a metal foil a tenth of a millimeter thick, implying that they were a form of light.

Other experiments, however, pointed to the contrary conclusion. Light is a wave phenomenon, and one distinguishing property of waves is demonstrated by speaking into one end of a paper towel roll. The sound waves do not emerge from the other end of the tube as a focused beam. Instead, they begin spreading out in all directions as soon as they emerge. This shows that waves do not necessarily travel in straight lines. If a piece of metal foil in the shape of a star or a cross was placed in the way of the cathode ray, then a “shadow” of the same shape would appear on the glass, showing that the rays traveled in straight lines. This straight-line motion suggested that they were a stream of small particles of matter.

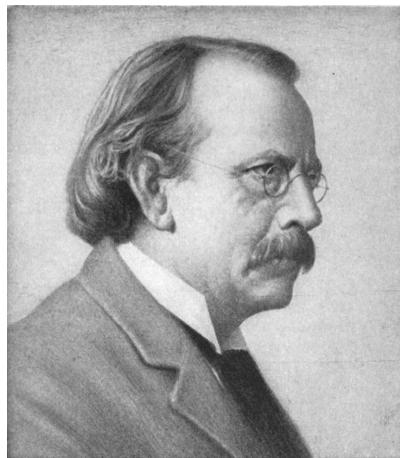
These observations were inconclusive, so what was really needed was a determination of whether the rays had mass and weight. The trouble was that cathode rays could not simply be collected in a cup and put on a scale. When the cathode ray tube is in operation, one does not observe any loss of material from the cathode, or any crust being deposited on the anode.

Nobody could think of a good way to weigh cathode rays, so the next most obvious way of settling the light/matter debate was to



a / Cathode rays observed in a vacuum tube.

check whether the cathode rays possessed electrical charge. Light was known to be uncharged. If the cathode rays carried charge, they were definitely matter and not light, and they were presumably being made to jump the gap by the simultaneous repulsion of the negative charge in the cathode and attraction of the positive charge in the anode. The rays would overshoot the anode because of their momentum. (Although electrically charged particles do not normally leap across a gap of vacuum, very large amounts of charge were being used, so the forces were unusually intense.)

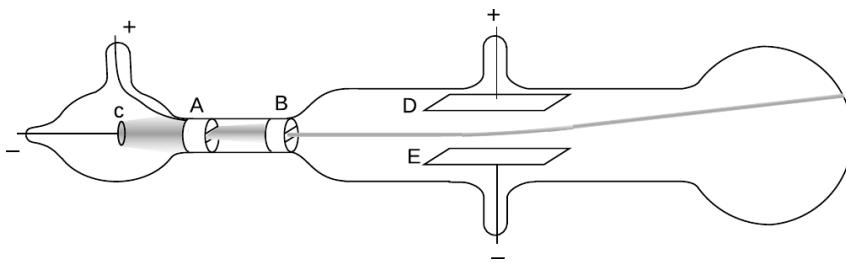


b / J.J. Thomson in the lab.

c / Thomson's experiment proving cathode rays had electric charge (redrawn from his original paper). The cathode, C, and anode, A, are as in any cathode ray tube. The rays pass through a slit in the anode, and a second slit, B, is interposed in order to make the beam thinner and eliminate rays that were not going straight. Charging plates D and E show that cathode rays have charge: they are attracted toward the positive plate D and repelled by the negative plate E.

13.2.3 Thomson's experiments

Physicist J.J. Thomson at Cambridge carried out a series of definitive experiments on cathode rays around the year 1897. By turning them slightly off course with electrical forces, c, he showed that they were indeed electrically charged, which was strong evidence that they were material. Not only that, but he proved that they had mass, and measured the ratio of their mass to their charge, m/q . Since their mass was not zero, he concluded that they were a form of matter, and presumably made up of a stream of microscopic, negatively charged particles. When Millikan published his results fourteen years later, it was reasonable to assume that the charge of one such particle equaled minus one fundamental charge, $q = -e$, and from the combination of Thomson's and Millikan's results one could therefore determine the mass of a single cathode ray particle.



The basic technique for determining m/q was simply to measure the angle through which the charged plates bent the beam. The electric force acting on a cathode ray particle while it was between the plates would be proportional to its charge,

$$F_{elec} = Eq,$$

where E is the electric field.

Application of Newton's second law, $a = F/m$, would allow m/q to be determined:

$$\frac{m}{q} = \frac{E}{a}$$

There was just one catch. Thomson needed to know the cathode ray particles' velocity in order to figure out their acceleration. At that

point, however, nobody had even an educated guess as to the speed of the cathode rays produced in a given vacuum tube. The beam appeared to leap across the vacuum tube practically instantaneously, so it was no simple matter of timing it with a stopwatch!

Thomson's clever solution was to observe the effect of both electric and magnetic forces on the beam. The magnetic force exerted by a particular magnet would depend on both the cathode ray's charge and its velocity:

$$F_{mag} = Bqv$$

Thomson played with the electric and magnetic forces until either one would produce an equal effect on the beam, allowing him to solve for the velocity,

$$v = \frac{E}{B}.$$

Knowing the velocity (which was on the order of 10% of the speed of light for his setup), he was able to find the acceleration and thus the mass-to-charge ratio m/q . Thomson's techniques were relatively crude (or perhaps more charitably we could say that they stretched the state of the art of the time), so with various methods he came up with m/q values that ranged over about a factor of two, even for cathode rays extracted from a cathode made of a single material. The best modern value is $m/q = 5.69 \times 10^{-12}$ kg/C, which is consistent with the low end of Thomson's range.

13.2.4 The cathode ray as a subatomic particle: the electron

What was significant about Thomson's experiment was not the actual numerical value of m/q , however, so much as the fact that, combined with Millikan's value of the fundamental charge, it gave a mass for the cathode ray particles that was thousands of times smaller than the mass of even the lightest atoms. Even without Millikan's results, which were 14 years in the future, Thomson recognized that the cathode rays' m/q was thousands of times smaller than the m/q ratios that had been measured for electrically charged atoms in chemical solutions. He correctly interpreted this as evidence that the cathode rays were smaller building blocks — he called them *electrons* — out of which atoms themselves were formed. This was an extremely radical claim, coming at a time when atoms had not yet been proven to exist! Even those who used the word “atom” often considered them no more than mathematical abstractions, not literal objects. The idea of searching for structure inside of “un-splittable” atoms was seen by some as lunacy, but within ten years Thomson's ideas had been amply verified by many more detailed experiments.

Discussion questions

A Thomson started to become convinced during his experiments that the “cathode rays” observed coming from the cathodes of vacuum tubes were building blocks of atoms — what we now call electrons. He then carried out observations with cathodes made of a variety of metals, and found that m/q was roughly the same in every case, considering his limited accuracy. Given his suspicion, why did it make sense to try different metals? How would the consistent values of m/q serve to test his hypothesis?

B My students have frequently asked whether the m/q that Thomson measured was the value for a single electron, or for the whole beam. Can you answer this question?

C Thomson found that the m/q of an electron was thousands of times smaller than that of charged atoms in chemical solutions. Would this imply that the electrons had more charge? Less mass? Would there be no way to tell? Explain. Remember that Millikan’s results were still many years in the future, so q was unknown.

D Can you guess any practical reason why Thomson couldn’t just let one electron fly across the gap before disconnecting the battery and turning off the beam, and then measure the amount of charge deposited on the anode, thus allowing him to measure the charge of a single electron directly?

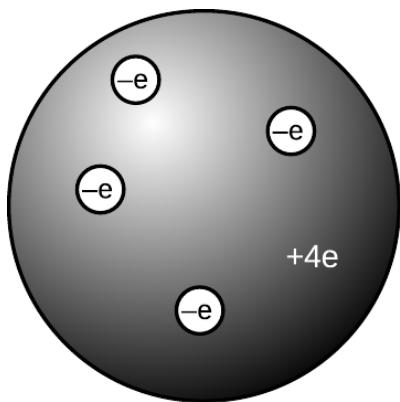
E Why is it not possible to determine m and q themselves, rather than just their ratio, by observing electrons’ motion in electric and magnetic fields?

13.3 The raisin cookie model of the atom

Based on his experiments, Thomson proposed a picture of the atom which became known as the raisin cookie model. In the neutral atom, d , there are four electrons with a total charge of $-4e$, sitting in a sphere (the “cookie”) with a charge of $+4e$ spread throughout it. It was known that chemical reactions could not change one element into another, so in Thomson’s scenario, each element’s cookie sphere had a permanently fixed radius, mass, and positive charge, different from those of other elements. The electrons, however, were not a permanent feature of the atom, and could be tacked on or pulled out to make charged ions. Although we now know, for instance, that a neutral atom with four electrons is the element beryllium, scientists at the time did not know how many electrons the various neutral atoms possessed.

This model is clearly different from the one you’ve learned in grade school or through popular culture, where the positive charge is concentrated in a tiny nucleus at the atom’s center. An equally important change in ideas about the atom has been the realization that atoms and their constituent subatomic particles behave entirely differently from objects on the human scale. For instance, we’ll see later that an electron can be in more than one place at one time.

d / The raisin cookie model of the atom with four units of charge, which we now know to be beryllium.



The raisin cookie model was part of a long tradition of attempts to make mechanical models of phenomena, and Thomson and his contemporaries never questioned the appropriateness of building a mental model of an atom as a machine with little parts inside. Today, mechanical models of atoms are still used (for instance the tinker-toy-style molecular modeling kits like the ones used by Watson and Crick to figure out the double helix structure of DNA), but scientists realize that the physical objects are only aids to help our brains' symbolic and visual processes think about atoms.

Although there was no clear-cut experimental evidence for many of the details of the raisin cookie model, physicists went ahead and started working out its implications. For instance, suppose you had a four-electron atom. All four electrons would be repelling each other, but they would also all be attracted toward the center of the "cookie" sphere. The result should be some kind of stable, symmetric arrangement in which all the forces canceled out. People sufficiently clever with math soon showed that the electrons in a four-electron atom should settle down at the vertices of a pyramid with one less side than the Egyptian kind, i.e., a regular tetrahedron. This deduction turns out to be wrong because it was based on incorrect features of the model, but the model also had many successes, a few of which we will now discuss.

Flow of electrical charge in wires *example 1*

One of my former students was the son of an electrician, and had become an electrician himself. He related to me how his father had remained refused to believe all his life that electrons really flowed through wires. If they had, he reasoned, the metal would have gradually become more and more damaged, eventually crumbling to dust.

His opinion is not at all unreasonable based on the fact that electrons are material particles, and that matter cannot normally pass through matter without making a hole through it. Nineteenth-century physicists would have shared his objection to a charged-particle model of the flow of electrical charge. In the raisin-cookie model, however, the electrons are very low in mass, and therefore presumably very small in size as well. It is not surprising that they can slip between the atoms without damaging them.

Flow of electrical charge across cell membranes *example 2*

Your nervous system is based on signals carried by charge moving from nerve cell to nerve cell. Your body is essentially all liquid, and atoms in a liquid are mobile. This means that, unlike the case of charge flowing in a solid wire, entire charged atoms can flow in your nervous system

Emission of electrons in a cathode ray tube *example 3*

Why do electrons detach themselves from the cathode of a vacuum tube? Certainly they are encouraged to do so by the re-

pulsion of the negative charge placed on the cathode and the attraction from the net positive charge of the anode, but these are not strong enough to rip electrons out of atoms by main force — if they were, then the entire apparatus would have been instantly vaporized as every atom was simultaneously ripped apart!

The raisin cookie model leads to a simple explanation. We know that heat is the energy of random motion of atoms. The atoms in any object are therefore violently jostling each other all the time, and a few of these collisions are violent enough to knock electrons out of atoms. If this occurs near the surface of a solid object, the electron may come loose. Ordinarily, however, this loss of electrons is a self-limiting process; the loss of electrons leaves the object with a net positive charge, which attracts the lost sheep home to the fold. (For objects immersed in air rather than vacuum, there will also be a balanced exchange of electrons between the air and the object.)

This interpretation explains the warm and friendly yellow glow of the vacuum tubes in an antique radio. To encourage the emission of electrons from the vacuum tubes' cathodes, the cathodes are intentionally warmed up with little heater coils.

Discussion questions

A Today many people would define an ion as an atom (or molecule) with missing electrons or extra electrons added on. How would people have defined the word "ion" before the discovery of the electron?

B Since electrically neutral atoms were known to exist, there had to be positively charged subatomic stuff to cancel out the negatively charged electrons in an atom. Based on the state of knowledge immediately after the Millikan and Thomson experiments, was it possible that the positively charged stuff had an unquantized amount of charge? Could it be quantized in units of $+e$? In units of $+2e$? In units of $+5/7e$?

13.4 The nucleus

13.4.1 Radioactivity

Becquerel's discovery of radioactivity

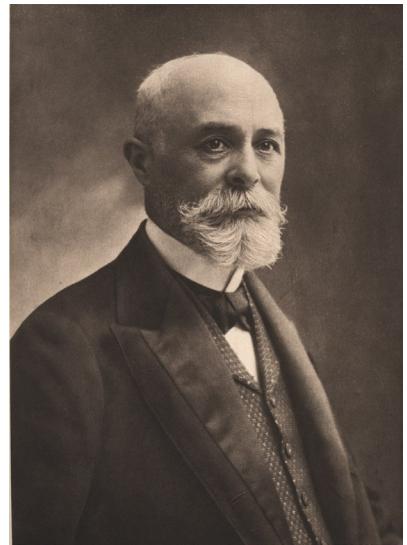
How did physicists figure out that the raisin cookie model was incorrect, and that the atom's positive charge was concentrated in a tiny, central nucleus? The story begins with the discovery of radioactivity by the French chemist Becquerel. Up until radioactivity was discovered, all the processes of nature were thought to be based on chemical reactions, which were rearrangements of combinations of atoms. Atoms exert forces on each other when they are close together, so sticking or unsticking them would either release or store electrical energy. That energy could be converted to and from other forms, as when a plant uses the energy in sunlight to make sugars and carbohydrates, or when a child eats sugar, releasing the energy

in the form of kinetic energy.

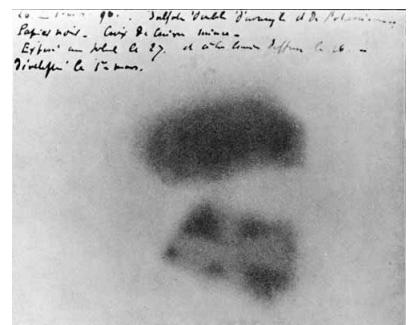
Becquerel discovered a process that seemed to release energy from an unknown new source that was not chemical. Becquerel, whose father and grandfather had also been physicists, spent the first twenty years of his professional life as a successful civil engineer, teaching physics on a part-time basis. He was awarded the chair of physics at the Musée d'Histoire Naturelle in Paris after the death of his father, who had previously occupied it. Having now a significant amount of time to devote to physics, he began studying the interaction of light and matter. He became interested in the phenomenon of phosphorescence, in which a substance absorbs energy from light, then releases the energy via a glow that only gradually goes away. One of the substances he investigated was a uranium compound, the salt $\text{U}_3\text{SO}_4 \cdot 10\text{H}_2\text{O}$. One day in 1896, cloudy weather interfered with his plan to expose this substance to sunlight in order to observe its fluorescence. He stuck it in a drawer, coincidentally on top of a blank photographic plate — the old-fashioned glass-backed counterpart of the modern plastic roll of film. The plate had been carefully wrapped, but several days later when Becquerel checked it in the darkroom before using it, he found that it was ruined, as if it had been completely exposed to light.

History provides many examples of scientific discoveries that happened this way: an alert and inquisitive mind decides to investigate a phenomenon that most people would not have worried about explaining. Becquerel first determined by further experiments that the effect was produced by the uranium salt, despite a thick wrapping of paper around the plate that blocked out all light. He tried a variety of compounds, and found that it was the uranium that did it: the effect was produced by any uranium compound, but not by any compound that didn't include uranium atoms. The effect could be at least partially blocked by a sufficient thickness of metal, and he was able to produce silhouettes of coins by interposing them between the uranium and the plate. This indicated that the effect traveled in a straight line., so that it must have been some kind of ray rather than, e.g., the seepage of chemicals through the paper. He used the word “radiations,” since the effect radiated out from the uranium salt.

At this point Becquerel still believed that the uranium atoms were absorbing energy from light and then gradually releasing the energy in the form of the mysterious rays, and this was how he presented it in his first published lecture describing his experiments. Interesting, but not earth-shattering. But he then tried to determine how long it took for the uranium to use up all the energy that had supposedly been stored in it by light, and he found that it never seemed to become inactive, no matter how long he waited. Not only that, but a sample that had been exposed to intense sunlight for a whole afternoon was no more or less effective than a sample that



e / Henri Becquerel (1852-1908).



f / Becquerel's photographic plate. In the exposure at the bottom of the image, he has found that he could absorb the radiations, casting the shadow of a Maltese cross that was placed between the plate and the uranium salts.

had always been kept inside. Was this a violation of conservation of energy? If the energy didn't come from exposure to light, where did it come from?

Three kinds of “radiations”

Unable to determine the source of the energy directly, turn-of-the-century physicists instead studied the behavior of the “radiations” once they had been emitted. Becquerel had already shown that the radioactivity could penetrate through cloth and paper, so the first obvious thing to do was to investigate in more detail what thickness of material the radioactivity could get through. They soon learned that a certain fraction of the radioactivity’s intensity would be eliminated by even a few inches of air, but the remainder was not eliminated by passing through more air. Apparently, then, the radioactivity was a mixture of more than one type, of which one was blocked by air. They then found that of the part that could penetrate air, a further fraction could be eliminated by a piece of paper or a very thin metal foil. What was left after that, however, was a third, extremely penetrating type, some of whose intensity would still remain even after passing through a brick wall. They decided that this showed there were three types of radioactivity, and without having the faintest idea of what they really were, they made up names for them. The least penetrating type was arbitrarily labeled α (alpha), the first letter of the Greek alphabet, and so on through β (beta) and finally γ (gamma) for the most penetrating type.

Radium: a more intense source of radioactivity

The measuring devices used to detect radioactivity were crude: photographic plates or even human eyeballs (radioactivity makes flashes of light in the jelly-like fluid inside the eye, which can be seen by the eyeball’s owner if it is otherwise very dark). Because the ways of detecting radioactivity were so crude and insensitive, further progress was hindered by the fact that the amount of radioactivity emitted by uranium was not really very great. The vital contribution of physicist/chemist Marie Curie and her husband Pierre was to discover the element radium, and to purify and isolate significant quantities of it. Radium emits about a million times more radioactivity per unit mass than uranium, making it possible to do the experiments that were needed to learn the true nature of radioactivity. The dangers of radioactivity to human health were then unknown, and Marie died of leukemia thirty years later. (Pierre was run over and killed by a horsecart.)

Tracking down the nature of alphas, betas, and gammas

As radium was becoming available, an apprentice scientist named Ernest Rutherford arrived in England from his native New Zealand and began studying radioactivity at the Cavendish Laboratory. The young colonial’s first success was to measure the mass-to-charge ra-

tio of beta rays. The technique was essentially the same as the one Thomson had used to measure the mass-to-charge ratio of cathode rays by measuring their deflections in electric and magnetic fields. The only difference was that instead of the cathode of a vacuum tube, a nugget of radium was used to supply the beta rays. Not only was the technique the same, but so was the result. Beta rays had the same m/q ratio as cathode rays, which suggested they were one and the same. Nowadays, it would make sense simply to use the term “electron,” and avoid the archaic “cathode ray” and “beta particle,” but the old labels are still widely used, and it is unfortunately necessary for physics students to memorize all three names for the same thing.

At first, it seemed that neither alphas or gammas could be deflected in electric or magnetic fields, making it appear that neither was electrically charged. But soon Rutherford obtained a much more powerful magnet, and was able to use it to deflect the alphas but not the gammas. The alphas had a much larger value of m/q than the betas (about 4000 times greater), which was why they had been so hard to deflect. Gammas are uncharged, and were later found to be a form of light.

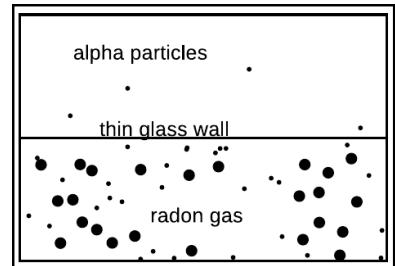
The m/q ratio of alpha particles turned out to be the same as those of two different types of ions, He^{++} (a helium atom with two missing electrons) and H_2^+ (two hydrogen atoms bonded into a molecule, with one electron missing), so it seemed likely that they were one or the other of those. The diagram shows a simplified version of Rutherford’s ingenious experiment proving that they were He^{++} ions. The gaseous element radon, an alpha emitter, was introduced into one half of a double glass chamber. The glass wall dividing the chamber was made extremely thin, so that some of the rapidly moving alpha particles were able to penetrate it. The other chamber, which was initially evacuated, gradually began to accumulate a population of alpha particles (which would quickly pick up electrons from their surroundings and become electrically neutral). Rutherford then determined that it was helium gas that had appeared in the second chamber. Thus alpha particles were proved to be He^{++} ions. The nucleus was yet to be discovered, but in modern terms, we would describe a He^{++} ion as the nucleus of a He atom.

To summarize, here are the three types of radiation emitted by radioactive elements, and their descriptions in modern terms:

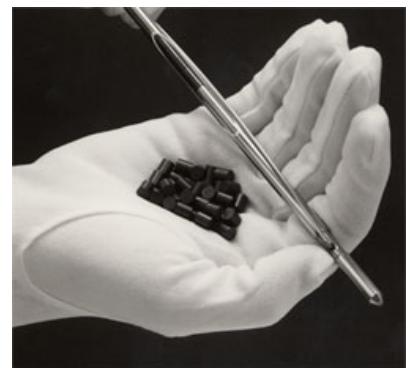
α particle	stopped by a few inches of air	He nucleus
β particle	stopped by a piece of paper	electron
γ ray	penetrates thick shielding	a type of light

Discussion question

A Most sources of radioactivity emit alphas, betas, and gammas, not just one of the three. In the radon experiment, how did Rutherford know that he was studying the alphas?



g / A simplified version of Rutherford’s 1908 experiment, showing that alpha particles were doubly ionized helium atoms.



h / These pellets of uranium fuel will be inserted into the metal fuel rod and used in a nuclear reactor. The pellets emit alpha and beta radiation, which the gloves are thick enough to stop.

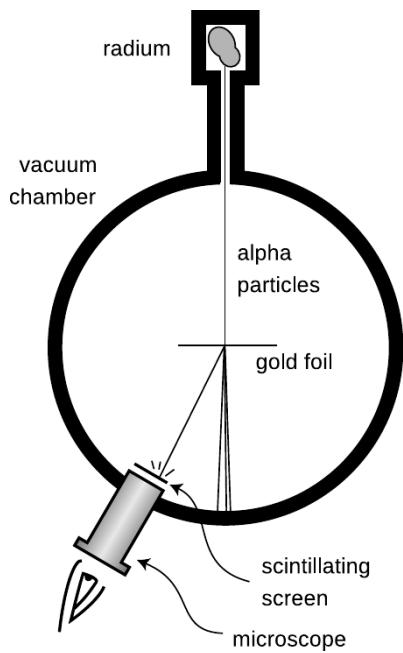


13.4.2 The planetary model

The stage was now set for the unexpected discovery that the positively charged part of the atom was a tiny, dense lump at the atom's center rather than the "cookie dough" of the raisin cookie model. By 1909, Rutherford was an established professor, and had students working under him. For a raw undergraduate named Marsden, he picked a research project he thought would be tedious but straightforward.

It was already known that although alpha particles would be stopped completely by a sheet of paper, they could pass through a sufficiently thin metal foil. Marsden was to work with a gold foil only 1000 atoms thick. (The foil was probably made by evaporating a little gold in a vacuum chamber so that a thin layer would be deposited on a glass microscope slide. The foil would then be lifted off the slide by submerging the slide in water.)

i / Ernest Rutherford (1871-1937).

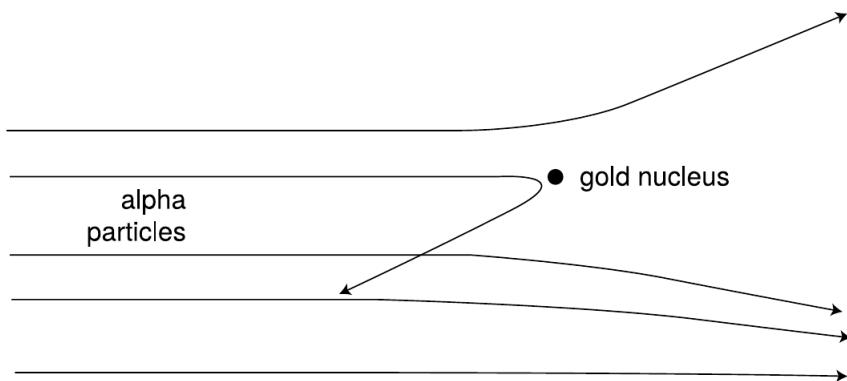


j / Marsden and Rutherford's apparatus.

Rutherford had already determined in his previous experiments the speed of the alpha particles emitted by radium, a fantastic 1.5×10^7 m/s. The experimenters in Rutherford's group visualized them as very small, very fast cannonballs penetrating the "cookie dough" part of the big gold atoms. A piece of paper has a thickness of a hundred thousand atoms or so, which would be sufficient to stop them completely, but crashing through a thousand would only slow them a little and turn them slightly off of their original paths.

Marsden's supposedly ho-hum assignment was to use the apparatus shown in figure j to measure how often alpha particles were deflected at various angles. A tiny lump of radium in a box emitted alpha particles, and a thin beam was created by blocking all the alphas except those that happened to pass out through a tube. Typically deflected in the gold by only a small amount, they would reach a screen very much like the screen of a TV's picture tube, which would make a flash of light when it was hit. Here is the first example we have encountered of an experiment in which a beam of particles is detected one at a time. This was possible because each alpha particle carried so much kinetic energy; they were moving at about the same speed as the electrons in the Thomson experiment, but had ten thousand times more mass.

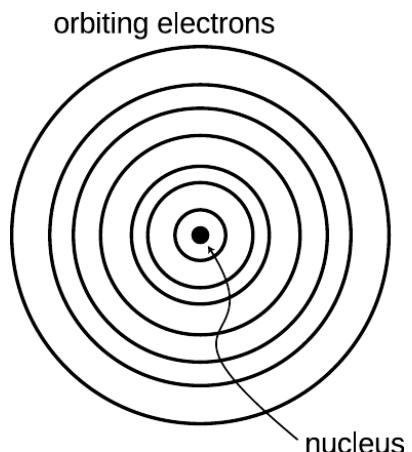
Marsden sat in a dark room, watching the apparatus hour after hour and recording the number of flashes with the screen moved to various angles. The rate of the flashes was highest when he set the screen at an angle close to the line of the alphas' original path, but if he watched an area farther off to the side, he would also occasionally see an alpha that had been deflected through a larger angle. After seeing a few of these, he got the crazy idea of moving the screen to see if even larger angles ever occurred, perhaps even angles larger than 90 degrees.



k / Alpha particles being scattered by a gold nucleus. On this scale, the gold atom is the size of a car, so all the alpha particles shown here are ones that just happened to come unusually close to the nucleus. For these exceptional alpha particles, the forces from the electrons are unimportant, because they are so much more distant than the nucleus.

The crazy idea worked: a few alpha particles were deflected through angles of up to 180 degrees, and the routine experiment had become an epoch-making one. Rutherford said, “We have been able to get some of the alpha particles coming backwards. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you.” Explanations were hard to come by in the raisin cookie model. What intense electrical forces could have caused some of the alpha particles, moving at such astronomical speeds, to change direction so drastically? Since each gold atom was electrically neutral, it would not exert much force on an alpha particle outside it. True, if the alpha particle was very near to or inside of a particular atom, then the forces would not necessarily cancel out perfectly; if the alpha particle happened to come very close to a particular electron, the $1/r^2$ form of the Coulomb force law would make for a very strong force. But Marsden and Rutherford knew that an alpha particle was 8000 times more massive than an electron, and it is simply not possible for a more massive object to rebound backwards from a collision with a less massive object while conserving momentum and energy. It might be possible in principle for a particular alpha to follow a path that took it very close to one electron, and then very close to another electron, and so on, with the net result of a large deflection, but careful calculations showed that such multiple “close encounters” with electrons would be millions of times too rare to explain what was actually observed.

At this point, Rutherford and Marsden dusted off an unpopular and neglected model of the atom, in which all the electrons orbited around a small, positively charged core or “nucleus,” just like the planets orbiting around the sun. All the positive charge



l / The planetary model of the atom.

and nearly all the mass of the atom would be concentrated in the nucleus, rather than spread throughout the atom as in the raisin cookie model. The positively charged alpha particles would be repelled by the gold atom's nucleus, but most of the alphas would not come close enough to any nucleus to have their paths drastically altered. The few that did come close to a nucleus, however, could rebound backwards from a single such encounter, since the nucleus of a heavy gold atom would be fifty times more massive than an alpha particle. It turned out that it was not even too difficult to derive a formula giving the relative frequency of deflections through various angles, and this calculation agreed with the data well enough (to within 15%), considering the difficulty in getting good experimental statistics on the rare, very large angles.

What had started out as a tedious exercise to get a student started in science had ended as a revolution in our understanding of nature. Indeed, the whole thing may sound a little too much like a moralistic fable of the scientific method with overtones of the Horatio Alger genre. The skeptical reader may wonder why the planetary model was ignored so thoroughly until Marsden and Rutherford's discovery. Is science really more of a sociological enterprise, in which certain ideas become accepted by the establishment, and other, equally plausible explanations are arbitrarily discarded? Some social scientists are currently ruffling a lot of scientists' feathers with critiques very much like this, but in this particular case, there were very sound reasons for rejecting the planetary model. As you'll learn in more detail later in this course, any charged particle that undergoes an acceleration dissipate energy in the form of light. In the planetary model, the electrons were orbiting the nucleus in circles or ellipses, which meant they were undergoing acceleration, just like the acceleration you feel in a car going around a curve. They should have dissipated energy as light, and eventually they should have lost all their energy. Atoms don't spontaneously collapse like that, which was why the raisin cookie model, with its stationary electrons, was originally preferred. There were other problems as well. In the planetary model, the one-electron atom would have to be flat, which would be inconsistent with the success of molecular modeling with spherical balls representing hydrogen and atoms. These molecular models also seemed to work best if specific sizes were used for different atoms, but there is no obvious reason in the planetary model why the radius of an electron's orbit should be a fixed number. In view of the conclusive Marsden-Rutherford results, however, these became fresh puzzles in atomic physics, not reasons for disbelieving the planetary model.

Some phenomena explained with the planetary model

The planetary model may not be the ultimate, perfect model of the atom, but don't underestimate its power. It already allows us

to visualize correctly a great many phenomena.

As an example, let's consider the distinctions among nonmetals, metals that are magnetic, and metals that are nonmagnetic. As shown in figure m, a metal differs from a nonmetal because its outermost electrons are free to wander rather than owing their allegiance to a particular atom. A metal that can be magnetized is one that is willing to line up the rotations of some of its electrons so that their axes are parallel. Recall that magnetic forces are forces made by moving charges; we have not yet discussed the mathematics and geometry of magnetic forces, but it is easy to see how random orientations of the atoms in the nonmagnetic substance would lead to cancellation of the forces.

Even if the planetary model does not immediately answer such questions as why one element would be a metal and another a nonmetal, these ideas would be difficult or impossible to conceptualize in the raisin cookie model.

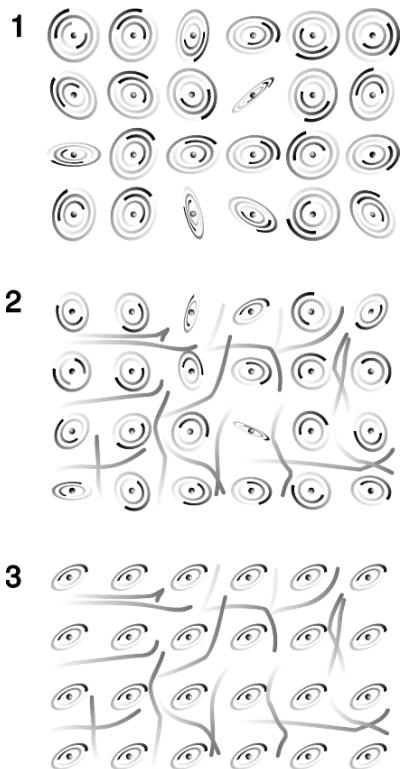
Discussion question

A In reality, charges of the same type repel one another and charges of different types are attracted. Suppose the rules were the other way around, giving repulsion between opposite charges and attraction between similar ones. What would the universe be like?

13.4.3 Atomic number

As alluded to in a discussion question in the previous section, scientists of this period had only a very approximate idea of how many units of charge resided in the nuclei of the various chemical elements. Although we now associate the number of units of nuclear charge with the element's position on the periodic table, and call it the atomic number, they had no idea that such a relationship existed. Mendeleev's table just seemed like an organizational tool, not something with any necessary physical significance. And everything Mendeleev had done seemed equally valid if you turned the table upside-down or reversed its left and right sides, so even if you wanted to number the elements sequentially with integers, there was an ambiguity as to how to do it. Mendeleev's original table was in fact upside-down compared to the modern one.

In the period immediately following the discovery of the nucleus, physicists only had rough estimates of the charges of the various nuclei. In the case of the very lightest nuclei, they simply found the maximum number of electrons they could strip off by various methods: chemical reactions, electric sparks, ultraviolet light, and so on. For example they could easily strip off one or two electrons from helium, making He^+ or He^{++} , but nobody could make He^{+++} , presumably because the nuclear charge of helium was only $+2e$. Unfortunately only a few of the lightest elements could be stripped completely, because the more electrons were stripped off, the greater the



m / The planetary model applied to a nonmetal, 1, an unmagnetized metal, 2, and a magnetized metal, 3. Note that these figures are all simplified in several ways. For one thing, the electrons of an individual atom do not all revolve around the nucleus in the same plane. It is also very unusual for a metal to become so strongly magnetized that 100% of its atoms have their rotations aligned as shown in this figure.

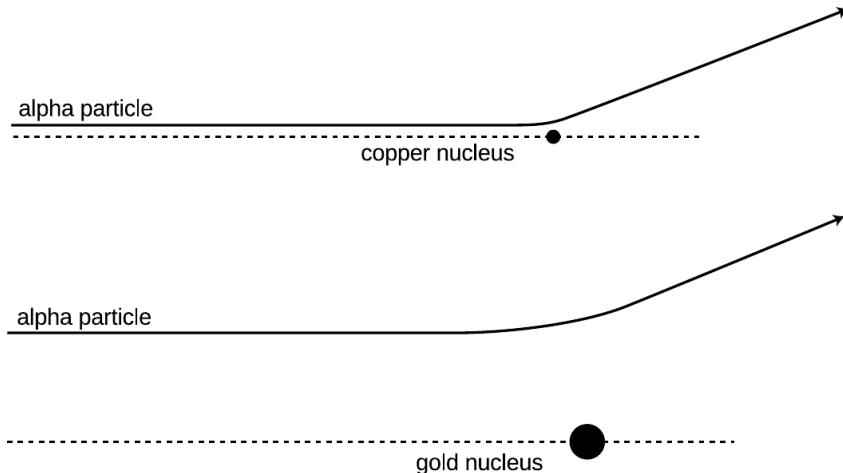
n / A modern periodic table, labeled with atomic numbers. Mendeleev's original table was upside-down compared to this one.

1 H															2 He			
3 Li	4 Be																	
11 Na	12 Mg																	
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr	
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe	
55 Cs	56 Ba	57 La	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 Ac	**	104 Rf	105 Ha	106	107	108	109	110	111	112	113	114	115	116	117	118
* ⁵⁸ Ce ⁵⁹ Pr ⁶⁰ Nd ⁶¹ Pm ⁶² Sm ⁶³ Eu ⁶⁴ Gd ⁶⁵ Tb ⁶⁶ Dy ⁶⁷ Ho ⁶⁸ Er ⁶⁹ Tm ⁷⁰ Yb ⁷¹ Lu																		
** ⁹⁰ Th ⁹¹ Pa ⁹² U ⁹³ Np ⁹⁴ Pu ⁹⁵ Am ⁹⁶ Cm ⁹⁷ Bk ⁹⁸ Cf ⁹⁹ Es ¹⁰⁰ Fm ¹⁰¹ Md ¹⁰² No ¹⁰³ Lr																		

positive net charge remaining, and the more strongly the rest of the negatively charged electrons would be held on. The heavy elements' atomic numbers could only be roughly extrapolated from the light elements, where the atomic number was about half the atom's mass expressed in units of the mass of a hydrogen atom. Gold, for example, had a mass about 197 times that of hydrogen, so its atomic number was estimated to be about half that, or somewhere around 100. We now know it to be 79.

How did we finally find out? The riddle of the nuclear charges was at last successfully attacked using two different techniques, which gave consistent results. One set of experiments, involving x-rays, was performed by the young Henry Mosely, whose scientific brilliance was soon to be sacrificed in a battle between European imperialists over who would own the Dardanelles, during that pointless conflict then known as the War to End All Wars, and now referred to as World War I.

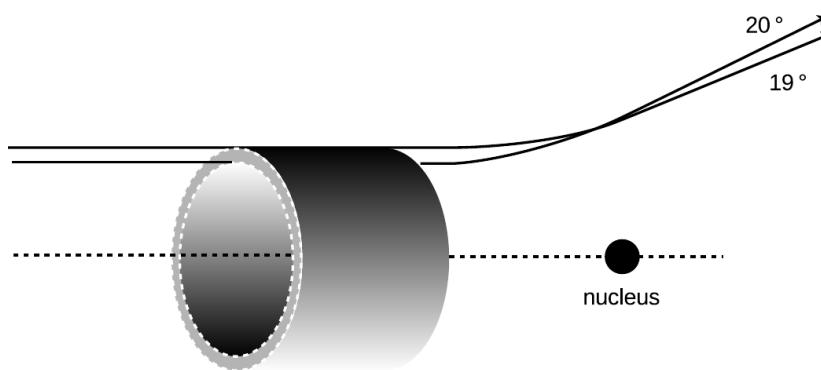
o / An alpha particle has to come much closer to the low-charged copper nucleus in order to be deflected through the same angle.



Since Mosely's analysis requires several concepts with which you are not yet familiar, we will instead describe the technique used by James Chadwick at around the same time. An added bonus of

describing Chadwick's experiments is that they presaged the important modern technique of studying *collisions* of subatomic particles. In grad school, I worked with a professor whose thesis adviser's thesis adviser was Chadwick, and he related some interesting stories about the man. Chadwick was apparently a little nutty and a complete fanatic about science, to the extent that when he was held in a German prison camp during World War II, he managed to cajole his captors into allowing him to scrounge up parts from broken radios so that he could attempt to do physics experiments.

Chadwick's experiment worked like this. Suppose you perform two Rutherford-type alpha scattering measurements, first one with a gold foil as a target as in Rutherford's original experiment, and then one with a copper foil. It is possible to get large angles of deflection in both cases, but as shown in figure p, the alpha particle must be heading almost straight for the copper nucleus to get the same angle of deflection that would have occurred with an alpha that was much farther off the mark; the gold nucleus' charge is so much greater than the copper's that it exerts a strong force on the alpha particle even from far off. The situation is very much like that of a blindfolded person playing darts. Just as it is impossible to aim an alpha particle at an individual nucleus in the target, the blindfolded person cannot really aim the darts. Achieving a very close encounter with the copper atom would be akin to hitting an inner circle on the dartboard. It's much more likely that one would have the luck to hit the outer circle, which covers a greater number of square inches. By analogy, if you measure the frequency with which alphas are scattered by copper at some particular angle, say between 19 and 20 degrees, and then perform the same measurement at the same angle with gold, you get a much higher percentage for gold than for copper.



p / An alpha particle must be headed for the ring on the front of the imaginary cylindrical pipe in order to produce scattering at an angle between 19 and 20 degrees. The area of this ring is called the "cross-section" for scattering at 19-20° because it is the cross-sectional area of a cut through the pipe.

In fact, the numerical ratio of the two nuclei's charges can be derived from this same experimentally determined ratio. Using the standard notation Z for the atomic number (charge of the nucleus

divided by e), the following equation can be proved (example 4):

$$\frac{Z_{gold}^2}{Z_{copper}^2} = \frac{\text{number of alphas scattered by gold at } 19\text{--}20^\circ}{\text{number of alphas scattered by copper at } 19\text{--}20^\circ}$$

By making such measurements for targets constructed from all the elements, one can infer the ratios of all the atomic numbers, and since the atomic numbers of the light elements were already known, atomic numbers could be assigned to the entire periodic table. According to Moseley, the atomic numbers of copper, silver and platinum were 29, 47, and 78, which corresponded well with their positions on the periodic table. Chadwick's figures for the same elements were 29.3, 46.3, and 77.4, with error bars of about 1.5 times the fundamental charge, so the two experiments were in good agreement.

The point here is absolutely not that you should be ready to plug numbers into the above equation for a homework or exam question! My overall goal in this chapter is to explain how we know what we know about atoms. An added bonus of describing Chadwick's experiment is that the approach is very similar to that used in modern particle physics experiments, and the ideas used in the analysis are closely related to the now-ubiquitous concept of a "cross-section." In the dartboard analogy, the cross-section would be the area of the circular ring you have to hit. The reasoning behind the invention of the term "cross-section" can be visualized as shown in figure p. In this language, Rutherford's invention of the planetary model came from his unexpected discovery that there was a nonzero cross-section for alpha scattering from gold at large angles, and Chadwick confirmed Moseley's determinations of the atomic numbers by measuring cross-sections for alpha scattering.

Proof of the relationship between Z and scattering example 4

The equation above can be derived by the following not very rigorous proof. To deflect the alpha particle by a certain angle requires that it acquire a certain momentum component in the direction perpendicular to its original momentum. Although the nucleus's force on the alpha particle is not constant, we can pretend that it is approximately constant during the time when the alpha is within a distance equal to, say, 150% of its distance of closest approach, and that the force is zero before and after that part of the motion. (If we chose 120% or 200%, it shouldn't make any difference in the final result, because the final result is a ratio, and the effects on the numerator and denominator should cancel each other.) In the approximation of constant force, the change in the alpha's perpendicular momentum component is then equal to $F\Delta t$. The Coulomb force law says the force is proportional to Z/r^2 . Although r does change somewhat during the time interval of interest, it's good enough to treat it as a constant number, since

we're only computing the ratio between the two experiments' results. Since we are approximating the force as acting over the time during which the distance is not too much greater than the distance of closest approach, the time interval Δt must be proportional to r , and the sideways momentum imparted to the alpha, $F\Delta t$, is proportional to $(Z/r^2)r$, or Z/r . If we're comparing alphas scattered at the same angle from gold and from copper, then Δp is the same in both cases, and the proportionality $\Delta p \propto Z/r$ tells us that the ones scattered from copper at that angle had to be headed in along a line closer to the central axis by a factor equaling $Z_{\text{gold}}/Z_{\text{copper}}$. If you imagine a "dartboard ring" that the alphas have to hit, then the ring for the gold experiment has the same proportions as the one for copper, but it is enlarged by a factor equal to $Z_{\text{gold}}/Z_{\text{copper}}$. That is, not only is the radius of the ring greater by that factor, but unlike the rings on a normal dartboard, the thickness of the outer ring is also greater in proportion to its radius. When you take a geometric shape and scale it up in size like a photographic enlargement, its area is increased in proportion to the square of the enlargement factor, so the area of the dartboard ring in the gold experiment is greater by a factor equal to $(Z_{\text{gold}}/Z_{\text{copper}})^2$. Since the alphas are aimed entirely randomly, the chances of an alpha hitting the ring are in proportion to the area of the ring, which proves the equation given above.

As an example of the modern use of scattering experiments and cross-section measurements, you may have heard of the recent experimental evidence for the existence of a particle called the top quark. Of the twelve subatomic particles currently believed to be the smallest constituents of matter, six form a family called the quarks, distinguished from the other six by the intense attractive forces that make the quarks stick to each other. (The other six consist of the electron plus five other, more exotic particles.) The only two types of quarks found in naturally occurring matter are the "up quark" and "down quark," which are what protons and neutrons are made of, but four other types were theoretically predicted to exist, for a total of six. (The whimsical term "quark" comes from a line by James Joyce reading "Three quarks for master Mark.") Until recently, only five types of quarks had been proven to exist via experiments, and the sixth, the top quark, was only theorized. There was no hope of ever detecting a top quark directly, since it is radioactive, and only exists for a zillionth of a second before evaporating. Instead, the researchers searching for it at the Fermi National Accelerator Laboratory near Chicago measured cross-sections for scattering of nuclei off of other nuclei. The experiment was much like those of Rutherford and Chadwick, except that the incoming nuclei had to be boosted to much higher speeds in a particle accelerator. The resulting encounter with a target nucleus was so violent that both nuclei were completely demolished, but, as Einstein proved, energy

can be converted into matter, and the energy of the collision creates a spray of exotic, radioactive particles, like the deadly shower of wood fragments produced by a cannon ball in an old naval battle. Among those particles were some top quarks. The cross-sections being measured were the cross-sections for the production of certain combinations of these secondary particles. However different the details, the principle was the same as that employed at the turn of the century: you smash things together and look at the fragments that fly off to see what was inside them. The approach has been compared to shooting a clock with a rifle and then studying the pieces that fly off to figure out how the clock worked.

Discussion questions

A The diagram, showing alpha particles being deflected by a gold nucleus, was drawn with the assumption that alpha particles came in on lines at many different distances from the nucleus. Why wouldn't they all come in along the same line, since they all came out through the same tube?

B Why does it make sense that, as shown in the figure, the trajectories that result in 19° and 20° scattering cross each other?

C Rutherford knew the velocity of the alpha particles emitted by radium, and guessed that the positively charged part of a gold atom had a charge of about $+100e$ (we now know it is $+79e$). Considering the fact that some alpha particles were deflected by 180° , how could he then use conservation of energy to derive an upper limit on the size of a gold nucleus? (For simplicity, assume the size of the alpha particle is negligible compared to that of the gold nucleus, and ignore the fact that the gold nucleus recoils a little from the collision, picking up a little kinetic energy.)

13.4.4 The structure of nuclei

The proton

The fact that the nuclear charges were all integer multiples of e suggested to many physicists that rather than being a pointlike object, the nucleus might contain smaller particles having individual charges of $+e$. Evidence in favor of this idea was not long in arriving. Rutherford reasoned that if he bombarded the atoms of a very light element with alpha particles, the small charge of the target nuclei would give a very weak repulsion. Perhaps those few alpha particles that happened to arrive on head-on collision courses would get so close that they would physically crash into some of the target nuclei. An alpha particle is itself a nucleus, so this would be a collision between two nuclei, and a violent one due to the high speeds involved. Rutherford hit pay dirt in an experiment with alpha particles striking a target containing nitrogen atoms. Charged particles were detected flying out of the target like parts flying off of cars in a high-speed crash. Measurements of the deflection of these particles in electric and magnetic fields showed that they had the same charge-to-mass ratio as singly-ionized hydrogen atoms. Rutherford concluded that these were the conjectured singly-charged particles

that held the charge of the nucleus, and they were later named protons. The hydrogen nucleus consists of a single proton, and in general, an element's atomic number gives the number of protons contained in each of its nuclei. The mass of the proton is about 1800 times greater than the mass of the electron.

The neutron

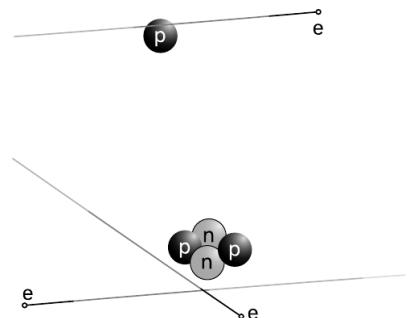
It would have been nice and simple if all the nuclei could have been built only from protons, but that couldn't be the case. If you spend a little time looking at a periodic table, you will soon notice that although some of the atomic masses are very nearly integer multiples of hydrogen's mass, many others are not. Even where the masses are close whole numbers, the masses of an element other than hydrogen is always greater than its atomic number, not equal to it. Helium, for instance, has two protons, but its mass is four times greater than that of hydrogen.

Chadwick cleared up the confusion by proving the existence of a new subatomic particle. Unlike the electron and proton, which are electrically charged, this particle is electrically neutral, and he named it the neutron. The method Chadwick used was to expose a sample of the light element beryllium to a stream of alpha particles from a lump of radium. Beryllium has only four protons, so an alpha that happens to be aimed directly at a beryllium nucleus can actually hit it rather than being stopped short of a collision by electrical repulsion. Neutrons were observed as a new form of radiation emerging from the collisions, and Chadwick correctly inferred that they were previously unsuspected components of the nucleus that had been knocked out. As described earlier, Chadwick also determined the mass of the neutron; it is very nearly the same as that of the proton.

To summarize, atoms are made of three types of particles:

	<i>charge</i>	<i>mass in units of the proton's mass</i>	<i>location in atom</i>
proton	$+e$	1	in nucleus
neutron	0	1.001	in nucleus
electron	$-e$	1/1836	orbiting nucleus

The existence of neutrons explained the mysterious masses of the elements. Helium, for instance, has a mass very close to four times greater than that of hydrogen. This is because it contains two neutrons in addition to its two protons. The mass of an atom is essentially determined by the total number of neutrons and protons. The total number of neutrons plus protons is therefore referred to as the atom's *mass number*.



q / Examples of the construction of atoms: hydrogen (top) and helium (bottom). On this scale, the electrons' orbits would be the size of a college campus.

Isotopes

We now have a clear interpretation of the fact that helium is close to four times more massive than hydrogen, and similarly for all the atomic masses that are close to an integer multiple of the mass of hydrogen. But what about copper, for instance, which had an atomic mass 63.5 times that of hydrogen? It didn't seem reasonable to think that it possessed an extra half of a neutron! The solution was found by measuring the mass-to-charge ratios of singly-ionized atoms (atoms with one electron removed). The technique is essentially that same as the one used by Thomson for cathode rays, except that whole atoms do not spontaneously leap out of the surface of an object as electrons sometimes do. Figure r shows an example of how the ions can be created and injected between the charged plates for acceleration.

Injecting a stream of copper ions into the device, we find a surprise — the beam splits into two parts! Chemists had elevated to dogma the assumption that all the atoms of a given element were identical, but we find that 69% of copper atoms have one mass, and 31% have another. Not only that, but both masses are very nearly integer multiples of the mass of hydrogen (63 and 65, respectively). Copper gets its chemical identity from the number of protons in its nucleus, 29, since chemical reactions work by electric forces. But apparently some copper atoms have $63 - 29 = 34$ neutrons while others have $65 - 29 = 36$. The atomic mass of copper, 63.5, reflects the proportions of the mixture of the mass-63 and mass-65 varieties. The different mass varieties of a given element are called *isotopes* of that element.

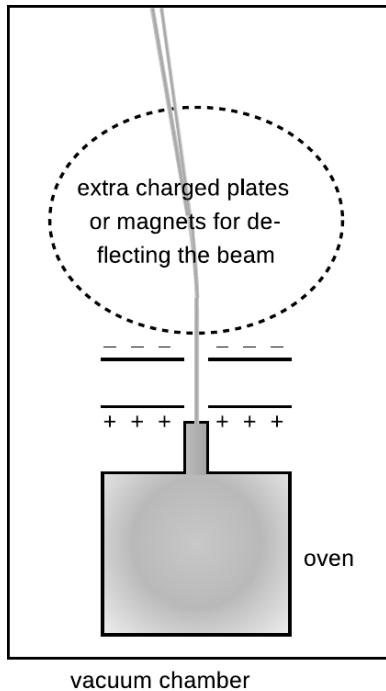
Isotopes can be named by giving the mass number as a subscript to the left of the chemical symbol, e.g., ^{65}Cu . Examples:

	protons	neutrons	mass number
^1H	1	0	$0+1 = 1$
^4He	2	2	$2+2 = 4$
^{12}C	6	6	$6+6 = 12$
^{14}C	6	8	$6+8 = 14$
^{262}Ha	105	157	$105+157 = 262$

self-check A

Why are the positive and negative charges of the accelerating plates reversed in the isotope-separating apparatus compared to the Thomson apparatus? ▷ Answer, p. 457

Chemical reactions are all about the exchange and sharing of electrons: the nuclei have to sit out this dance because the forces of electrical repulsion prevent them from ever getting close enough to make contact with each other. Although the protons do have a vitally important effect on chemical processes because of their electrical forces, the neutrons can have no effect on the atom's chemical



r / A version of the Thomson apparatus modified for measuring the mass-to-charge ratios of ions rather than electrons. A small sample of the element in question, copper in our example, is boiled in the oven to create a thin vapor. (A vacuum pump is continuously sucking on the main chamber to keep it from accumulating enough gas to stop the beam of ions.) Some of the atoms of the vapor are ionized by a spark or by ultraviolet light. Ions that wander out of the nozzle and into the region between the charged plates are then accelerated toward the top of the figure. As in the Thomson experiment, mass-to-charge ratios are inferred from the deflection of the beam.

reactions. It is not possible, for instance, to separate ^{63}Cu from ^{65}Cu by chemical reactions. This is why chemists had never realized that different isotopes existed. (To be perfectly accurate, different isotopes do behave slightly differently because the more massive atoms move more sluggishly and therefore react with a tiny bit less intensity. This tiny difference is used, for instance, to separate out the isotopes of uranium needed to build a nuclear bomb. The smallness of this effect makes the separation process a slow and difficult one, which is what we have to thank for the fact that nuclear weapons have not been built by every terrorist cabal on the planet.)

Sizes and shapes of nuclei

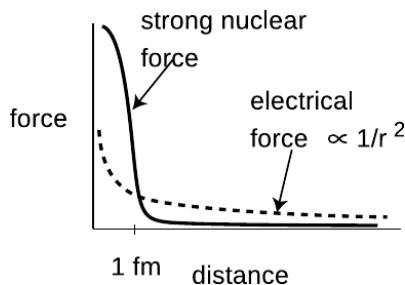
Matter is nearly all nuclei if you count by weight, but in terms of volume nuclei don't amount to much. The radius of an individual neutron or proton is very close to 1 fm ($1 \text{ fm} = 10^{-15} \text{ m}$), so even a big lead nucleus with a mass number of 208 still has a diameter of only about 13 fm, which is ten thousand times smaller than the diameter of a typical atom. Contrary to the usual imagery of the nucleus as a small sphere, it turns out that many nuclei are somewhat elongated, like an American football, and a few have exotic asymmetric shapes like pears or kiwi fruits.

Discussion questions

A Suppose the entire universe was in a (very large) cereal box, and the nutritional labeling was supposed to tell a godlike consumer what percentage of the contents was nuclei. Roughly what would the percentage be like if the labeling was according to mass? What if it was by volume?



s / A nuclear power plant at Cattenom, France. Unlike the coal and oil plants that supply most of the U.S.'s electrical power, a nuclear power plant like this one releases no pollution or greenhouse gases into the Earth's atmosphere, and therefore doesn't contribute to global warming. The white stuff puffing out of this plant is non-radioactive water vapor. Although nuclear power plants generate long-lived nuclear waste, this waste arguably poses much less of a threat to the biosphere than greenhouse gases would.

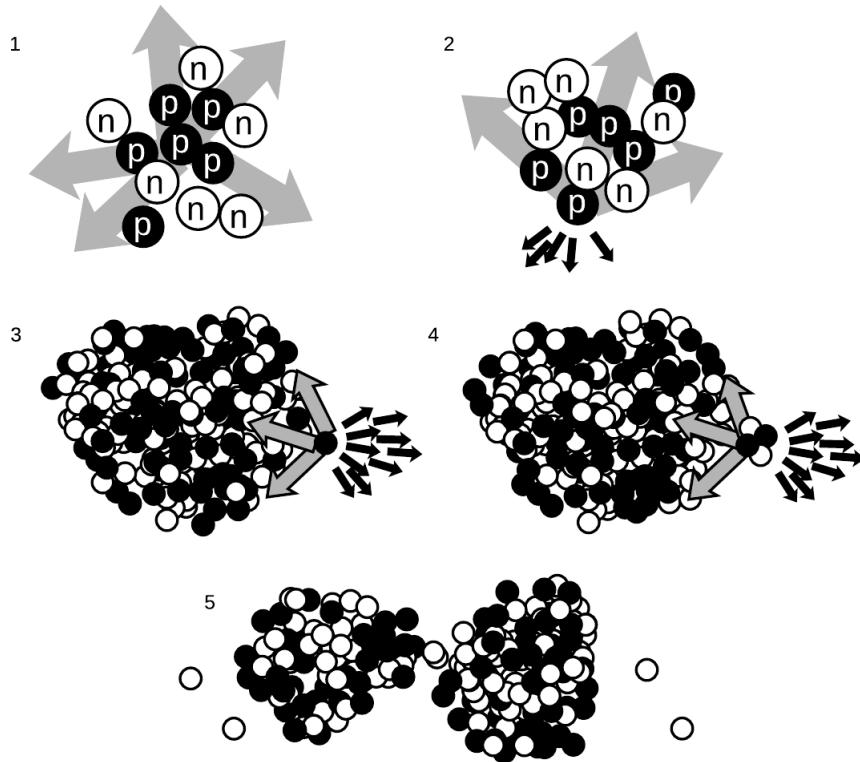


t / The strong nuclear force cuts off very sharply at a range of about 1 fm.

13.4.5 The strong nuclear force, alpha decay and fission

Once physicists realized that nuclei consisted of positively charged protons and uncharged neutrons, they had a problem on their hands. The electrical forces among the protons are all repulsive, so the nucleus should simply fly apart! The reason all the nuclei in your body are not spontaneously exploding at this moment is that there is another force acting. This force, called the *strong nuclear force*, is always attractive, and acts between neutrons and neutrons, neutrons and protons, and protons and protons with roughly equal strength. The strong nuclear force does not have any effect on electrons, which is why it does not influence chemical reactions.

Unlike electric forces, whose strengths are given by the simple Coulomb force law, there is no simple formula for how the strong nuclear force depends on distance. Roughly speaking, it is effective over ranges of ~ 1 fm, but falls off extremely quickly at larger distances (much faster than $1/r^2$). Since the radius of a neutron or proton is about 1 fm, that means that when a bunch of neutrons and protons are packed together to form a nucleus, the strong nuclear force is effective only between neighbors.



u / 1. The forces cancel. 2. The forces don't cancel. 3. In a heavy nucleus, the large number of electrical repulsions can add up to a force that is comparable to the strong nuclear attraction. 4. Alpha emission. 5. Fission.

Figure u illustrates how the strong nuclear force acts to keep

ordinary nuclei together, but is not able to keep very heavy nuclei from breaking apart. In figure u/1, a proton in the middle of a carbon nucleus feels an attractive strong nuclear force (arrows) from each of its nearest neighbors. The forces are all in different directions, and tend to cancel out. The same is true for the repulsive electrical forces (not shown). In figure u/2, a proton at the edge of the nucleus has neighbors only on one side, and therefore all the strong nuclear forces acting on it are tending to pull it back in. Although all the electrical forces from the other five protons (dark arrows) are all pushing it out of the nucleus, they are not sufficient to overcome the strong nuclear forces.

In a very heavy nucleus, u/3, a proton that finds itself near the edge has only a few neighbors close enough to attract it significantly via the strong nuclear force, but every other proton in the nucleus exerts a repulsive electrical force on it. If the nucleus is large enough, the total electrical repulsion may be sufficient to overcome the attraction of the strong force, and the nucleus may spit out a proton. Proton emission is fairly rare, however; a more common type of radioactive decay¹ in heavy nuclei is alpha decay, shown in figure u/4. The imbalance of the forces is similar, but the chunk that is ejected is an alpha particle (two protons and two neutrons) rather than a single proton.

It is also possible for the nucleus to split into two pieces of roughly equal size, u/5, a process known as fission. Note that in addition to the two large fragments, there is a spray of individual neutrons. In a nuclear fission bomb or a nuclear fission reactor, some of these neutrons fly off and hit other nuclei, causing them to undergo fission as well. The result is a chain reaction.

When a nucleus is able to undergo one of these processes, it is said to be radioactive, and to undergo radioactive decay. Some of the naturally occurring nuclei on earth are radioactive. The term “radioactive” comes from Becquerel’s image of rays radiating out from something, not from radio waves, which are a whole different phenomenon. The term “decay” can also be a little misleading, since it implies that the nucleus turns to dust or simply disappears – actually it is splitting into two new nuclei with the same total number of neutrons and protons, so the term “radioactive transformation” would have been more appropriate. Although the original atom’s electrons are mere spectators in the process of weak radioactive decay, we often speak loosely of “radioactive atoms” rather than “radioactive nuclei.”

¹Alpha decay is more common because an alpha particle happens to be a very stable arrangement of protons and neutrons.

Randomness in physics

How does an atom decide when to decay? We might imagine that it is like a termite-infested house that gets weaker and weaker, until finally it reaches the day on which it is destined to fall apart. Experiments, however, have not succeeded in detecting such “ticking clock” hidden below the surface; the evidence is that all atoms of a given isotope are absolutely identical. Why, then, would one uranium atom decay today while another lives for another million years? The answer appears to be that it is entirely random. We can make general statements about the average time required for a certain isotope to decay, or how long it will take for half the atoms in a sample to decay (its half-life), but we can never predict the behavior of a particular atom.

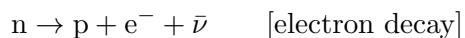
This is the first example we have encountered of an inescapable randomness in the laws of physics. If this kind of randomness makes you uneasy, you’re in good company. Einstein’s famous quote is “...I am convinced that He [God] does not play dice.” Einstein’s distaste for randomness, and his association of determinism with divinity, goes back to the Enlightenment conception of the universe as a gigantic piece of clockwork that only had to be set in motion initially by the Builder. Physics had to be entirely rebuilt in the 20th century to incorporate the fundamental randomness of physics, and this modern revolution is the topic of

chapters 15-18. In particular, we will delay the mathematical development of the half-life concept until then.

13.4.6 The weak nuclear force; beta decay

All the nuclear processes we’ve discussed so far have involved rearrangements of neutrons and protons, with no change in the total number of neutrons or the total number of protons. Now consider the proportions of neutrons and protons in your body and in the planet earth: neutrons and protons are roughly equally numerous in your body’s carbon and oxygen nuclei, and also in the nickel and iron that make up most of the earth. The proportions are about 50-50. But the only chemical elements produced in any significant quantities by the big bang were hydrogen (about 90%) and helium (about 10%). If the early universe was almost nothing but hydrogen atoms, whose nuclei are protons, where did all those neutrons come from?

The answer is that there is another nuclear force, the weak nuclear force, that is capable of transforming neutrons into protons and vice-versa. Two possible reactions are



and



(There is also a third type called electron capture, in which a proton grabs one of the atom's electrons and they produce a neutron and a neutrino.)

Whereas alpha decay and fission are just a redivision of the previously existing particles, these reactions involve the destruction of one particle and the creation of three new particles that did not exist before.

There are three new particles here that you have never previously encountered. The symbol e^+ stands for an antielectron, which is a particle just like the electron in every way, except that its electric charge is positive rather than negative. Antielectrons are also known as positrons. Nobody knows why electrons are so common in the universe and antielectrons are scarce. When an antielectron encounters an electron, they annihilate each other, producing gamma rays, and this is the fate of all the antielectrons that are produced by natural radioactivity on earth. Antielectrons are an example of antimatter. A complete atom of antimatter would consist of antiprotons, antielectrons, and antineutrons. Although individual particles of antimatter occur commonly in nature due to natural radioactivity and cosmic rays, only a few complete atoms of antihydrogen have ever been produced artificially.

The notation ν stands for a particle called a neutrino, and $\bar{\nu}$ means an antineutrino. Neutrinos and antineutrinos have no electric charge (hence the name).

We can now list all four of the known fundamental forces of physics:

- gravity
- electromagnetism
- strong nuclear force
- weak nuclear force

The other forces we have learned about, such as friction and the normal force, all arise from electromagnetic interactions between atoms, and therefore are not considered to be fundamental forces of physics.

Decay of ^{212}Pb

example 5

As an example, consider the radioactive isotope of lead ^{212}Pb . It contains 82 protons and 130 neutrons. It decays by the process $n \rightarrow p + e^- + \bar{\nu}$. The newly created proton is held inside the nucleus by the strong nuclear force, so the new nucleus contains 83 protons and 129 neutrons. Having 83 protons makes it the element bismuth, so it will be an atom of ^{212}Bi .

In a reaction like this one, the electron flies off at high speed (typically close to the speed of light), and the escaping electrons are the things that make large amounts of this type of radioactivity dangerous. The outgoing electron was the first thing that tipped off scientists in the early 1900s to the existence of this type of radioactivity. Since they didn't know that the outgoing particles were electrons, they called them beta particles, and this type of radioactive decay was therefore known as beta decay. A clearer but less common terminology is to call the two processes electron decay and positron decay.

The neutrino or antineutrino emitted in such a reaction pretty much ignores all matter, because its lack of charge makes it immune to electrical forces, and it also remains aloof from strong nuclear interactions. Even if it happens to fly off going straight down, it is almost certain to make it through the entire earth without interacting with any atoms in any way. It ends up flying through outer space forever. The neutrino's behavior makes it exceedingly difficult to detect, and when beta decay was first discovered nobody realized that neutrinos even existed. We now know that the neutrino carries off some of the energy produced in the reaction, but at the time it seemed that the total energy afterwards (not counting the unsuspected neutrino's energy) was greater than the total energy before the reaction, violating conservation of energy. Physicists were getting ready to throw conservation of energy out the window as a basic law of physics when indirect evidence led them to the conclusion that neutrinos existed.

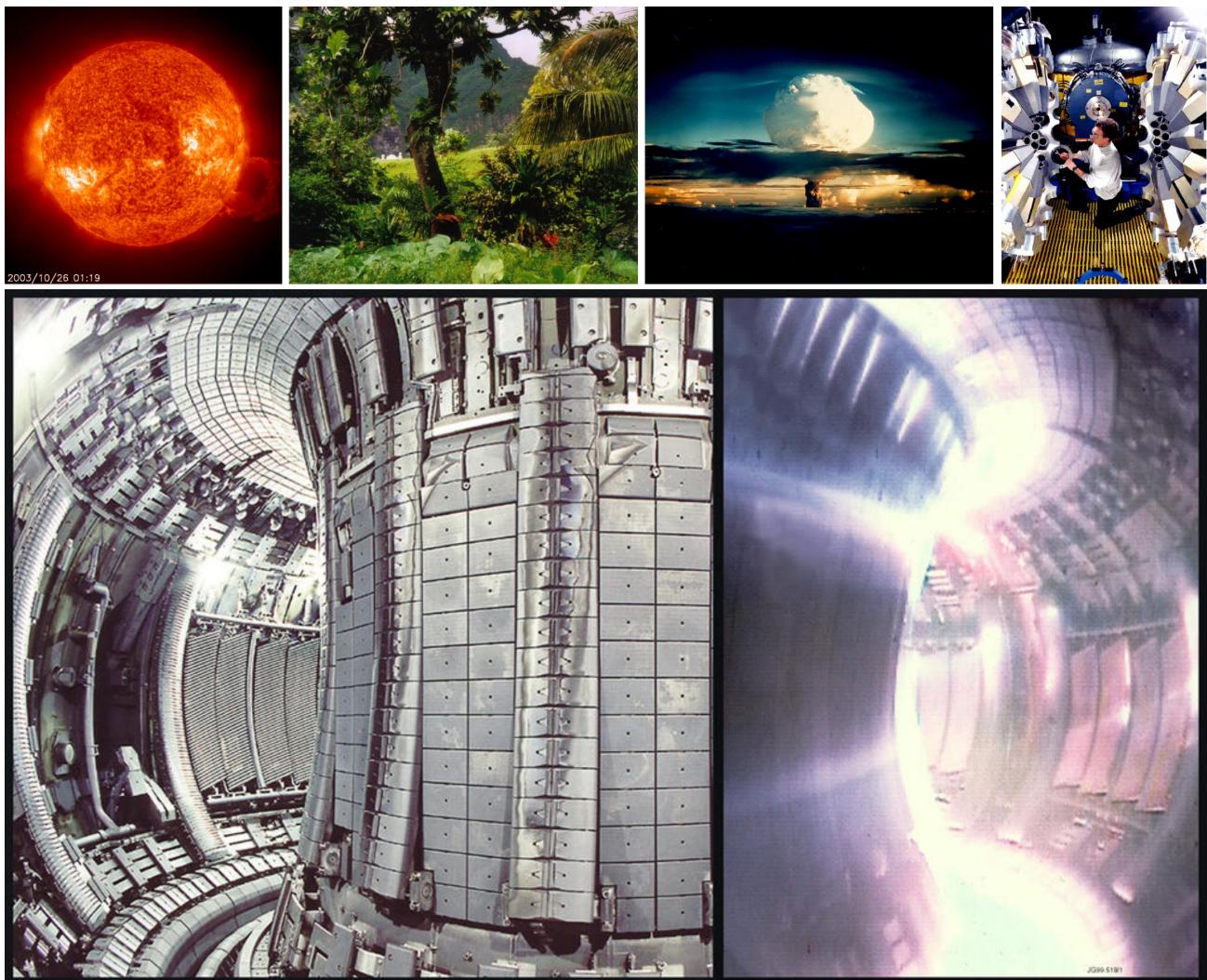
Discussion questions

A In the reactions $n \rightarrow p + e^- + \bar{\nu}$ and $p \rightarrow n + e^+ + \nu$, verify that charge is conserved. In beta decay, when one of these reactions happens to a neutron or proton within a nucleus, one or more gamma rays may also be emitted. Does this affect conservation of charge? Would it be possible for some extra electrons to be released without violating charge conservation?

B When an antielectron and an electron annihilate each other, they produce two gamma rays. Is charge conserved in this reaction?

13.4.7 Fusion

As we have seen, heavy nuclei tend to fly apart because each proton is being repelled by every other proton in the nucleus, but is only attracted by its nearest neighbors. The nucleus splits up into two parts, and as soon as those two parts are more than about 1 fm apart, the strong nuclear force no longer causes the two fragments to attract each other. The electrical repulsion then accelerates them, causing them to gain a large amount of kinetic energy. This release of kinetic energy is what powers nuclear reactors and fission bombs.



v / 1. Our sun's source of energy is nuclear fusion, so nuclear fusion is also the source of power for all life on earth, including, 2, this rain forest in Fatu-Hiva. 3. The first release of energy by nuclear fusion through human technology was the 1952 Ivy Mike test at the Enewetak Atoll. 4. This array of gamma-ray detectors is called GAMMASPHERE. During operation, the array is closed up, and a beam of ions produced by a particle accelerator strikes a target at its center, producing nuclear fusion reactions. The gamma rays can be studied for information about the structure of the fused nuclei, which are typically varieties not found in nature. 5. Nuclear fusion promises to be a clean, inexhaustible source of energy. However, the goal of commercially viable nuclear fusion power has remained elusive, due to the engineering difficulties involved in magnetically containing a plasma (ionized gas) at a sufficiently high temperature and density. This photo shows the experimental JET reactor, with the device opened up on the left, and in action on the right.

It might seem, then, that the lightest nuclei would be the most stable, but that is not the case. Let's compare an extremely light nucleus like ${}^4\text{He}$ with a somewhat heavier one, ${}^{16}\text{O}$. A neutron or proton in ${}^4\text{He}$ can be attracted by the three others, but in ${}^{16}\text{O}$, it might have five or six neighbors attracting it. The ${}^{16}\text{O}$ nucleus is therefore more stable.

It turns out that the most stable nuclei of all are those around nickel and iron, having about 30 protons and 30 neutrons. Just as a nucleus that is too heavy to be stable can release energy by splitting apart into pieces that are closer to the most stable size, light nuclei can release energy if you stick them together to make bigger nuclei that are closer to the most stable size. Fusing one nucleus with another is called nuclear fusion. Nuclear fusion is what powers our sun and other stars.

13.4.8 Nuclear energy and binding energies

In the same way that chemical reactions can be classified as exothermic (releasing energy) or endothermic (requiring energy to react), so nuclear reactions may either release or use up energy. The energies involved in nuclear reactions are greater by a huge factor. Thousands of tons of coal would have to be burned to produce as much energy as would be produced in a nuclear power plant by one kg of fuel.

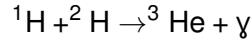
Although nuclear reactions that use up energy (endothermic reactions) can be initiated in accelerators, where one nucleus is rammed into another at high speed, they do not occur in nature, not even in the sun. The amount of kinetic energy required is simply not available.

To find the amount of energy consumed or released in a nuclear reaction, you need to know how much nuclear interaction energy, U_{nuc} , was stored or released. Experimentalists have determined the amount of nuclear energy stored in the nucleus of every stable element, as well as many unstable elements. This is the amount of mechanical work that would be required to pull the nucleus apart into its individual neutrons and protons, and is known as the nuclear binding energy.

A reaction occurring in the sun

example 6

The sun produces its energy through a series of nuclear fusion reactions. One of the reactions is



The excess energy is almost all carried off by the gamma ray (not by the kinetic energy of the helium-3 atom). The binding energies in units of pJ (picojoules) are:

$$^1\text{H} \quad 0 \text{ J}$$

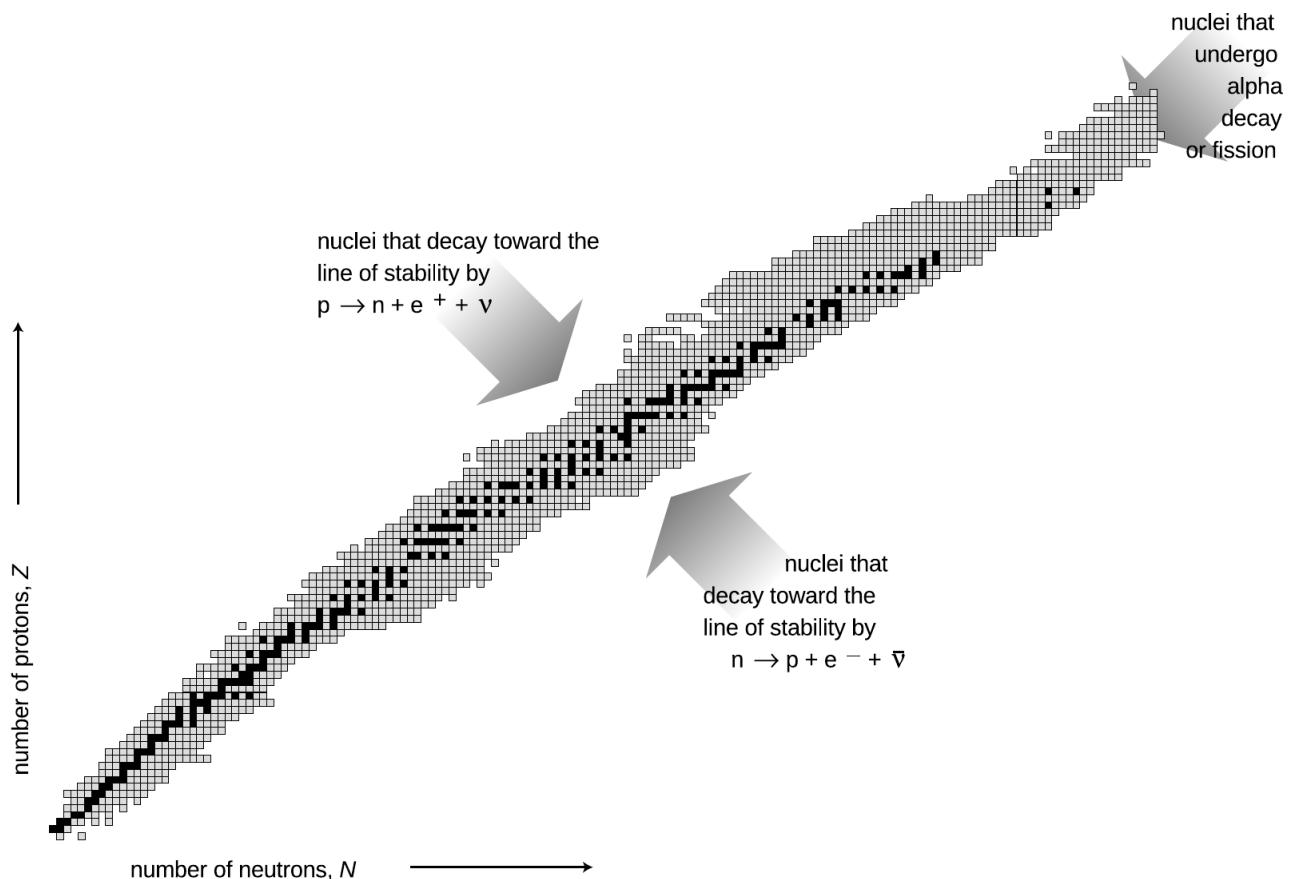
$$^2\text{H} \quad 0.35593 \text{ pJ}$$

$$^3\text{He} \quad 1.23489 \text{ pJ}$$

The total initial nuclear energy is $0 \text{ pJ} + 0.35593 \text{ pJ}$, and the final nuclear energy is 1.23489 pJ , so by conservation of energy, the gamma ray must carry off 0.87896 pJ of energy. The gamma ray is then absorbed by the sun and converted to heat.

self-check B

Why is the binding energy of ${}^1\text{H}$ exactly equal to zero? \triangleright Answer, p. 458



w / The known nuclei, represented on a chart of proton number versus neutron number. Note the two nuclei in the bottom row with zero protons.

Figure w is a compact way of showing the vast variety of the nuclei. Each box represents a particular number of neutrons and protons. The black boxes are nuclei that are stable, i.e., that would require an input of energy in order to change into another. The gray boxes show all the unstable nuclei that have been studied experimentally. Some of these last for billions of years on the average before decaying and are found in nature, but most have much shorter average lifetimes, and can only be created and studied in the laboratory.

The curve along which the stable nuclei lie is called the line of stability. Nuclei along this line have the most stable proportion of neutrons to protons. For light nuclei the most stable mixture is about 50-50, but we can see that stable heavy nuclei have two or three times more neutrons than protons. This is because the electrical repulsions of all the protons in a heavy nucleus add up



x / A map showing levels of radiation near the site of the Chernobyl nuclear accident.

to a powerful force that would tend to tear it apart. The presence of a large number of neutrons increases the distances among the protons, and also increases the number of attractions due to the strong nuclear force.

13.4.9 Biological effects of ionizing radiation

Units used to measure exposure

As a science educator, I find it frustrating that nowhere in the massive amount of journalism devoted to nuclear safety does one ever find any numerical statements about the amount of radiation to which people have been exposed. Anyone capable of understanding sports statistics or weather reports ought to be able to understand such measurements, as long as something like the following explanatory text was inserted somewhere in the article:

Radiation exposure is measured in units of Sieverts (Sv). The average person is exposed to about $2000 \mu\text{Sv}$ (microSieverts) each year from natural background sources.

With this context, people would be able to come to informed conclusions. For example, figure x shows a scary-looking map of the levels of radiation in the area surrounding the 1986 nuclear accident at Chernobyl, Ukraine, the most serious that has ever occurred. At the boundary of the most highly contaminated (bright red) areas, people would be exposed to about $13,000 \mu\text{Sv}$ per year, or about four times the natural background level. In the pink areas, which are still densely populated, the exposure is comparable to the natural level found in a high-altitude city such as Denver.

What is a Sievert? It measures the amount of energy per kilogram deposited in the body by ionizing radiation, multiplied by a “quality factor” to account for the different health hazards posed by alphas, betas, gammas, neutrons, and other types of radiation. Only ionizing radiation is counted, since nonionizing radiation simply heats one’s body rather than killing cells or altering DNA. For instance, alpha particles are typically moving so fast that their kinetic energy is sufficient to ionize thousands of atoms, but it is possible for an alpha particle to be moving so slowly that it would not have enough kinetic energy to ionize even one atom.

Unfortunately, most people don’t know much about radiation and tend to react to it based on unscientific cultural notions. These may, as in figure y, be based on fictional tropes silly enough to require the suspension of disbelief by the audience, but they can also be more subtle. People of my kids’ generation are more familiar with the 2011 Fukushima nuclear accident than with the much more serious Chernobyl accident. The news coverage of Fukushima showed scary scenes of devastated landscapes and distraught evacuees, implying that people had been killed and displaced by the release of radiation from the reaction. In fact, there were no deaths at all due

to the radiation released at Fukushima, and no excess cancer deaths are statistically predicted in the future. The devastation and the death toll of 16,000 were caused by the earthquake and tsunami, which were also what damaged the plant.

Effects of exposure

Notwithstanding the pop culture images like figure z, it is not possible for a multicellular animal to become “mutated” as a whole. In most cases, a particle of ionizing radiation will not even hit the DNA, and even if it does, it will only affect the DNA of a single cell, not every cell in the animal’s body. Typically, that cell is simply killed, because the DNA becomes unable to function properly. Once in a while, however, the DNA may be altered so as to make that cell cancerous. For instance, skin cancer can be caused by UV light hitting a single skin cell in the body of a sunbather. If that cell becomes cancerous and begins reproducing uncontrollably, she will end up with a tumor twenty years later.

Other than cancer, the only other dramatic effect that can result from altering a single cell’s DNA is if that cell happens to be a sperm or ovum, which can result in nonviable or mutated offspring. Men are relatively immune to reproductive harm from radiation, because their sperm cells are replaced frequently. Women are more vulnerable because they keep the same set of ova as long as they live.

Effects of high doses of radiation

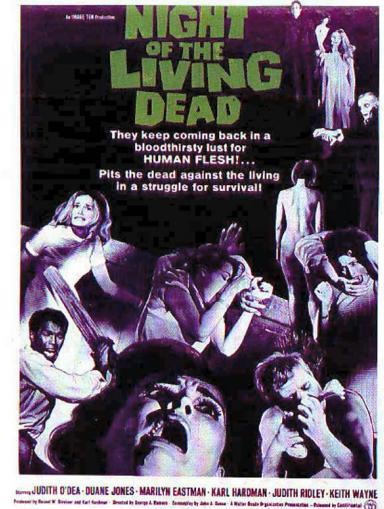
A whole-body exposure of 5,000,000 μSv will kill a person within a week or so. Luckily, only a small number of humans have ever been exposed to such levels: one scientist working on the Manhattan Project, some victims of the Nagasaki and Hiroshima explosions, and 31 workers at Chernobyl. Death occurs by massive killing of cells, especially in the blood-producing cells of the bone marrow.

Effects of low doses radiation

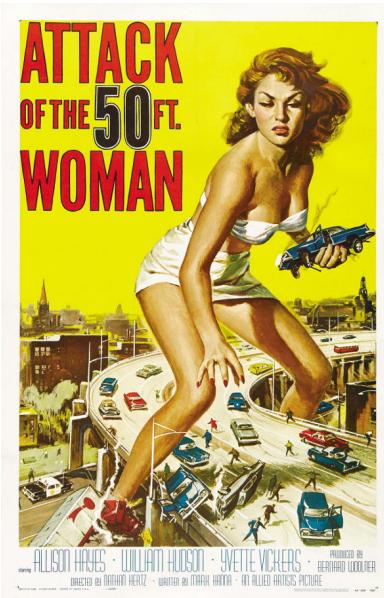
Lower levels, on the order of 1,000,000 μSv , were inflicted on some people at Nagasaki and Hiroshima. No acute symptoms result from this level of exposure, but certain types of cancer are significantly more common among these people. It was originally expected that the radiation would cause many mutations resulting in birth defects, but very few such inherited effects have been observed.

A great deal of time has been spent debating the effects of very low levels of ionizing radiation. The following table gives some sample figures.

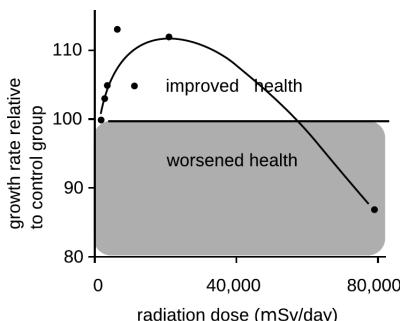
THEY WON'T STAY DEAD!



y / In this classic zombie flick, a newscaster speculates that the dead have been reanimated due to radiation brought back to earth by a space probe.



z / Radiation doesn't mutate entire multicellular organisms.



aa / A typical example of radiation hormesis: the health of mice is improved by low levels of radiation. In this study, young mice were exposed to fairly high levels of x-rays, while a control group of mice was not exposed. The mice were weighed, and their rate of growth was taken as a measure of their health. At levels below about $50,000 \mu\text{Sv}$, the radiation had a beneficial effect on the health of the mice, presumably by activating cellular damage control mechanisms. The two highest data points are statistically significant at the 99% level. The curve is a fit to a theoretical model. Redrawn from T.D. Luckey, *Hormesis with Ionizing Radiation*, CRC Press, 1980.

maximum <i>beneficial</i> dose per day	$\sim 10,000 \mu\text{Sv}$
CT scan	$\sim 10,000 \mu\text{Sv}$
natural background per year	2,000-7,000 μSv
health guidelines for exposure to a fetus	1,000 μSv
flying from New York to Tokyo	150 μSv
chest x-ray	50 μSv

Note that the largest number, on the first line of the table, is the maximum *beneficial* dose. The most useful evidence comes from experiments in animals, which can intentionally be exposed to significant and well measured doses of radiation under controlled conditions. Experiments show that low levels of radiation activate cellular damage control mechanisms, increasing the health of the organism. For example, exposure to radiation up to a certain level makes mice grow faster; makes guinea pigs' immune systems function better against diphteria; increases fertility in trout and mice; improves fetal mice's resistance to disease; increases the life-spans of flour beetles and mice; and reduces mortality from cancer in mice. This type of effect is called radiation hormesis.

There is also some evidence that in humans, small doses of radiation increase fertility, reduce genetic abnormalities, and reduce mortality from cancer. The human data, however, tend to be very poor compared to the animal data. Due to ethical issues, one cannot do controlled experiments in humans. For example, one of the best sources of information has been from the survivors of the Hiroshima and Nagasaki bomb blasts, but these people were also exposed to high levels of carcinogenic chemicals in the smoke from their burning cities; for comparison, firefighters have a heightened risk of cancer, and there are also significant concerns about cancer from the 9/11 attacks in New York. The direct empirical evidence about radiation hormesis in humans is therefore not good enough to tell us anything unambiguous,² and the most scientifically reasonable approach is to assume that the results in animals also hold for humans: small doses of radiation in humans are beneficial, rather than harmful. However, a variety of cultural and historical factors have led to a situation in which public health policy is based on the assumption, known as "linear no-threshold" (LNT), that even tiny doses of radiation are harmful, and that the risk they carry is proportional to the dose. In other words, law and policy are made based on the assumption that the effects of radiation on humans are dramatically different than its effects on mice and guinea pigs. Even with the unrealistic assumption of LNT, one can still evaluate risks by comparing with natural background radiation. For example, we can see that

²For two opposing viewpoints, see Tubiana et al., "The Linear No-Threshold Relationship Is Inconsistent with Radiation Biologic and Experimental Data," Radiology, 251 (2009) 13 and Little et al., "Risks Associated with Low Doses and Low Dose Rates of Ionizing Radiation: Why Linearity May Be (Almost) the Best We Can Do," Radiology, 251 (2009) 6.

the effect of a chest x-ray is about a hundred times smaller than the effect of spending a year in Colorado, where the level of natural background radiation from cosmic rays is higher than average, due to the high altitude. Dropping the implausible LNT assumption, we can see that the impact on one's health of spending a year in Colorado is likely to be *positive*, because the excess radiation is below the maximum beneficial level.

The green case for nuclear power

In the late twentieth century, antinuclear activists largely succeeded in bringing construction of new nuclear power plants to a halt in the U.S. Ironically, we now know that the burning of fossil fuels, which leads to global warming, is a far more grave threat to the environment than even the Chernobyl disaster. A team of biologists writes: "During recent visits to Chernobyl, we experienced numerous sightings of moose (*Alces alces*), roe deer (*Capreolus capreolus*), Russian wild boar (*Sus scrofa*), foxes (*Vulpes vulpes*), river otter (*Lutra canadensis*), and rabbits (*Lepus europaeus*) ... Diversity of flowers and other plants in the highly radioactive regions is impressive and equals that observed in protected habitats outside the zone ... The observation that typical human activity (industrialization, farming, cattle raising, collection of firewood, hunting, etc.) is more devastating to biodiversity and abundance of local flora and fauna than is the worst nuclear power plant disaster validates the negative impact the exponential growth of human populations has on wildlife."³

Nuclear power is the only source of energy that is sufficient to replace any significant percentage of energy from fossil fuels on the rapid schedule demanded by the speed at which global warming is progressing. People worried about the downside of nuclear energy might be better off putting their energy into issues related to nuclear weapons: the poor stewardship of the former Soviet Union's warheads; nuclear proliferation in unstable states such as Pakistan; and the poor safety and environmental history of the superpowers' nuclear weapons programs, including the loss of several warheads in plane crashes, and the environmental disaster at the Hanford, Washington, weapons plant.

Protection from radiation

People do sometimes work with strong enough radioactivity that there is a serious health risk. Typically the scariest sources are those used in cancer treatment and in medical and biological research.



ab / Wild Przewalski's horses prosper in the Chernobyl area.



ac / Fossil fuels have done incomparably more damage to the environment than nuclear power ever has. Polar bears' habitat is rapidly being destroyed by global warming.

³Baker and Chesson, Env. Toxicology and Chem. 19 (1231) 2000. Similar effects have been seen at the Bikini Atoll, the site of a 1954 hydrogen bomb test. Although some species have disappeared from the area, the coral reef is in many ways healthier than similar reefs elsewhere, because humans have tended to stay away for fear of radiation (Richards et al., Marine Pollution Bulletin 56 (2008) 503).

Also, a dental technician, for example, needs to take precautions to avoid accumulating a large radiation dose from giving dental x-rays to many patients. There are three general ways to reduce exposure: time, distance, and shielding. This is why a dental technician doing x-rays wears a lead apron (shielding) and steps outside of the x-ray room while running an exposure (distance). Reducing the time of exposure dictates, for example, that a person working with a hot cancer-therapy source would minimize the amount of time spent near it.

Shielding against alpha and beta particles is trivial to accomplish. (Alphas can't even penetrate the skin.) Gammas and x-rays interact most strongly with materials that are dense and have high atomic numbers, which is why lead is so commonly used. But other materials will also work. For example, the reason that bones show up so clearly on x-ray images is that they are dense and contain plenty of calcium, which has a higher atomic number than the elements found in most other body tissues, which are mostly made of water.

Neutrons are difficult to shield against. Because they are electrically neutral, they don't interact intensely with matter in the same way as alphas and betas. They only interact if they happen to collide head-on with a nucleus, and that doesn't happen very often because nuclei are tiny targets. Kinematically, a collision can transfer kinetic energy most efficiently when the target is as low in mass as possible compared to the projectile. For this reason, substances that contain a lot of hydrogen make the best shielding against neutrons. Blocks of paraffin wax from the supermarket are often used for this purpose.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 ^{241}Pu decays either by electron decay or by alpha decay. (A given ^{241}Pu nucleus may do either one; it's random.) What are the isotopes created as products of these two modes of decay?

2 As discussed in more detail in section 13.4, a nucleus contains protons, which have positive charge, and neutrons, which have zero charge. If only the electrical force existed, a nucleus would immediately fly apart due to electrical repulsion. However, there is also another force, called the strong nuclear force, which keeps this from happening. Suppose that a proton in a lead nucleus wanders out to the surface of the nucleus, and experiences a strong nuclear force of about 8 kN from the nearby neutrons and protons pulling it back in. Compare this numerically to the repulsive electrical force from the other protons, and verify that the net force is attractive. A lead nucleus is very nearly spherical, is about 6.5 fm in radius, and contains 82 protons, each with a charge of $+e$, where $e = 1.60 \times 10^{-19} \text{ C}$. ✓

3 The nuclear process of beta decay by electron capture is described parenthetically on page 313. The reaction is $\text{p} + \text{e}^- \rightarrow \text{n} + \nu$.
(a) Show that charge is conserved in this reaction.
(b) Conversion between energy and mass is discussed in sec. 3.6.1, p. 75. Based on these ideas, explain why electron capture doesn't occur in hydrogen atoms. (If it did, matter wouldn't exist!)

4 Potassium 40 is the strongest source of naturally occurring beta radioactivity in our environment. It decays according to



The energy released in the decay is 1.33 MeV, where 1 eV is defined as the fundamental charge e multiplied by one volt. The energy is shared randomly among the products, subject to the constraint imposed by conservation of energy-momentum, which dictates that very little of the energy is carried by the recoiling calcium nucleus. Determine the maximum energy of the calcium, and compare with the typical energy of a chemical bond, which is a few eV. If the potassium is part of a molecule, do we expect the molecule to survive? Carry out the calculation first by assuming that the electron is ultrarelativistic, then without the approximation, and comment on the how good the approximation is. *

Exercise 13: Nuclear decay

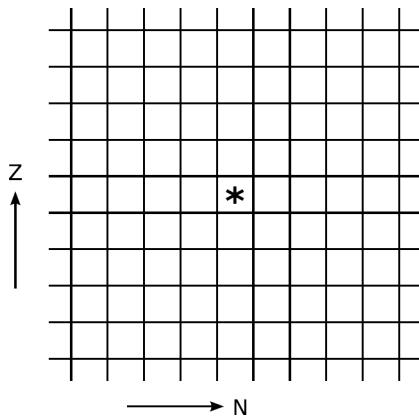
1. Consulting a periodic table, find the N , Z , and A of the following:

	N	Z	A
${}^4\text{He}$			
${}^{244}\text{Pu}$			

2. Consider the following five decay processes:

- α decay
- γ decay
- $\text{p} \rightarrow \text{n} + \text{e}^+ + \nu$ (β^+ decay)
- $\text{n} \rightarrow \text{p} + \text{e}^- + \bar{\nu}$ (β^- decay)
- $\text{p} + \text{e}^- \rightarrow \text{n} + \nu$ (electron capture)

What would be the action of each of these on the chart of the nuclei? The * represents the original nucleus.



3. (a) Suppose that ${}^{244}\text{Pu}$ undergoes perfectly symmetric fission, and also emits two neutrons. Find the daughter isotope.

(b) Is the daughter stable, or is it neutron-rich or -poor relative to the line of stability? (To estimate what's stable, you can use a large chart of the nuclei, or, if you don't have one handy, consult a periodic table and use the average atomic mass as an approximation to the stable value of A .)

(c) Consulting the chart of the nuclei (fig. w on p. 317), explain why it turns out this way.

(d) If the daughter is unstable, which process from question #2 would you expect it to decay by?

Chapter 14

Probability distributions and a first glimpse of quantum physics

14.1 Probability distributions

In ch. 7, we considered random variables such as the number of gas molecules r on the right-hand side of the box in figure e, p. 170. This variable is discrete rather than continuous, so we can speak meaningfully of the probability that the integer r has some particular value. On the other hand, the time t at which a particular unstable nucleus decays is a *continuous* variable. For such a variable, there is an infinite number of possible values, and the probability of any particular value is typically zero.

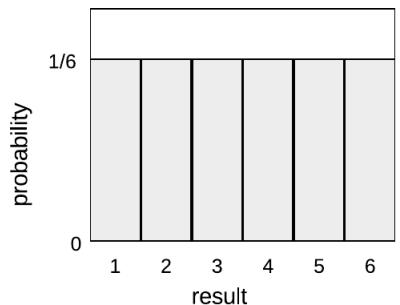
How do we handle this mathematically? Let's start finite and sneak up on the infinity.

Consider a throw of a die. If the die is honest, then we expect all six values to be equally likely. Since all six probabilities must add up to 1, then the probability of any particular value must be $1/6$. We can summarize this in a graph, a. Areas under the curve can be interpreted as total probabilities. For instance, the area under the curve from 1 to 3 is $1/6 + 1/6 + 1/6 = 1/2$, so the probability of getting a result from 1 to 3 is $1/2$. The function shown on the graph is called the probability distribution.

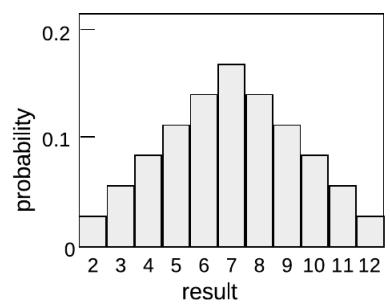
Figure b shows the probabilities of various results obtained by rolling two dice and adding them together, as in the game of craps. The probabilities are not all the same. There is a small probability of getting a two, for example, because there is only one way to do it, by rolling a one and then another one. The probability of rolling a seven is high because there are six different ways to do it: 1+6, 2+5, etc.

If the number of possible outcomes is large but finite, for example the number of hairs on a dog, the graph would start to look like a smooth curve rather than a ziggurat.

In these examples, the probability that the result will fall within some *range* is proportional to the area under the bar graph. In other words, we're talking about an integral. Passing to the case of



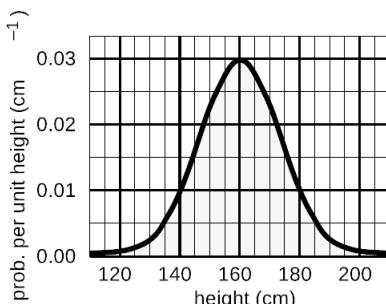
a / Probability distribution for the result of rolling a single die.



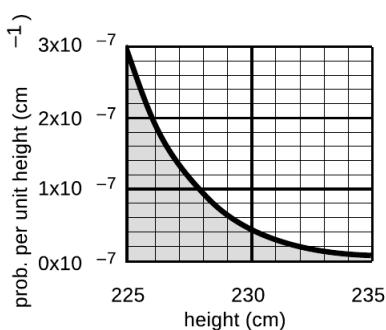
b / Rolling two dice and adding them up.

a continuous variable, we use this as our definition of the concept of a *probability distribution*. If x is a random number, the probability distribution $D(x)$ is defined so that the probability that x lies between a and b is equal to

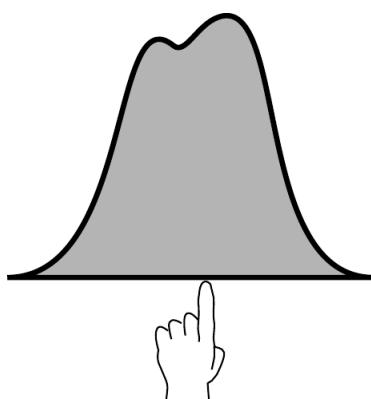
$$P(a \leq x \leq b) = \int_a^b D(x) dx.$$



c / A probability distribution for height of human adults (not real data).



d / Example 1.



e / The average of a probability distribution.

You've probably heard about "the bell curve," and seen people draw it with a pencil. When they do this, the function they're drawing is an example of one of these probability distributions. If you've heard of the idea that an electron in an atom is like a probability cloud, what is being described qualitatively is actually the function D (which in this case depends on three coordinates, x , y , and z).

Suppose that x has some units such as seconds. Then dx , which represents a small change in x , also has units of seconds, and since P is unitless, it follows that D has units of s^{-1} . That is, D represents the probability per unit of time. The same kind of thing occurs for random variables with other units: whatever units x has, D has the inverse of those units.

Recall that normalization (p. 169) is the requirement that the total probability for x to have *some* value must be one,

$$\int_{-\infty}^{\infty} D(x) dx = 1.$$

For a random variable that is discrete rather than continuous, we just do a sum rather than an integral, $\sum P(x) = 1$.

Figure c shows another example, a probability distribution for people's height. This kind of bell-shaped curve is quite common.

self-check A

Compare the number of people with heights in the range of 130–135 cm to the number in the range 135–140. ▷ Answer, p. 458

Looking for tall basketball players

example 1

▷ A certain country with a large population wants to find very tall people to be on its Olympic basketball team and strike a blow against western imperialism. Out of a pool of 10^8 people who are the right age and gender, how many are they likely to find who are over 225 cm (7 feet 4 inches) in height? Figure d gives a close-up of the "tail" of the distribution shown previously in figure c.

▷ The shaded area under the curve represents the probability that a given person is tall enough. Each rectangle represents a probability of $0.2 \times 10^{-7} \text{ cm}^{-1} \times 1 \text{ cm} = 2 \times 10^{-8}$. There are about 35 rectangles covered by the shaded area, so the probability of having a height greater than 225 cm is 7×10^{-7} , or just under one in a million. Using the rule for calculating averages, the average, or expected number of people this tall is $(10^8) \times (7 \times 10^{-7}) = 70$.

The average value of x is given by

$$(\text{average of } x) = \langle x \rangle = \bar{x} = \int_a^b x D(x) dx.$$

The notation $\langle \dots \rangle$ means “the average of \dots ,” and the bar in \bar{x} means the same thing. We can think of the average of a probability distribution geometrically as the horizontal position at which it could be balanced if it was constructed out of cardboard, figure e. For a discrete variable, we again just switch the integral to a probability-weighted sum, $\bar{x} = \sum x P(x)$.

The average is not the only possible way to say what is a typical value for a quantity that can vary randomly; another possible definition is the median, defined as the value that is exceeded with 50% probability. When discussing incomes of people living in a certain town, the average could be very misleading, since it can be affected massively if a single resident of the town is Mark Zuckerberg.

14.2 The variance and standard deviation

If the next Martian you meet asks you, “How tall is an adult human?,” you will probably reply with a statement about the average human height, such as “Oh, about 5 feet 6 inches.” If you wanted to explain a little more, you could say, “But that’s only an average. Most people are somewhere between 5 feet and 6 feet tall.” Without bothering to draw the relevant bell curve for your new extraterrestrial acquaintance, you’ve summarized the relevant information by giving an average and a typical range of variation.

Just as an average is not the only way of defining a central value of a distribution, there are many possible ways of measuring the amount of variation about that center. But a method that is common and has nice mathematical properties is the following. We define the *variance* of a probability distribution as follows:

$$(\text{variance of } x) = \langle (x - \bar{x})^2 \rangle.$$

In other words, we consider the difference between x and its average value \bar{x} , and we take the average of the square of that difference. If x always had exactly its average value, then $x - \bar{x}$ would always be zero, and the variance would be zero. It would not make sense to define the variance without the square, because, for example, a symmetrical probability distribution would have a variance of zero — the negative values of $x - \bar{x}$ would cancel the positive ones.

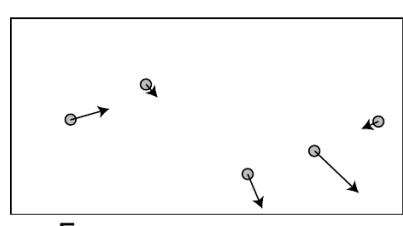
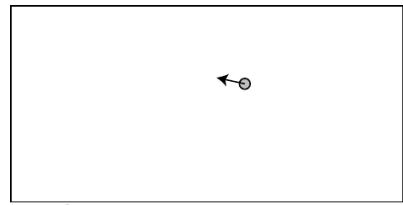
The big mathematical advantage of the variance is that it is additive: the variance of $x + y$ is the same as the sum of the variances, provided that x and y are not correlated to one another. For instance, if Susan has two bartending jobs, bringing in two different incomes x and y , then this week’s variation $x - \bar{x}$ in her tips at

Hipster Lounge X is probably not related to the variation $y - \bar{y}$ in what she gets from the tip jar at Y Bar and Grill. One can then plug in to the definition of the variance and show that the variances do add; this works out because the cross-term $\langle(x - \bar{x})(y - \bar{y})\rangle$ is zero.

The only unfortunate thing about the variance is that its units aren't the same as the units of the variable such as x . For example, if x has units of dollars, then the variance of x has units of dollars squared. For this reason, we define the standard deviation

$$\text{(standard deviation of } x) = \sigma_x = \sqrt{\langle(x - \bar{x})^2\rangle}.$$

When people give error bars in science experiments, or ranges of error in an opinion poll, they are usually quoting a standard deviation. If someone gives you a number like 137 ± 5 kg, then the 5 certainly can't be the variance, since the variance would have units of kg^2 , not kg. The standard deviation's name comes from its interpretation as a typical or standard amount by which x deviates from \bar{x} . In the context of AC circuits, you have probably encountered the idea of an r.m.s. (root-mean-square) value, which is exactly a standard deviation (in the case where $\bar{x} = 0$).



f / A two-sided box with one atom (top) and five atoms (bottom).

As an example, consider the simplest possible case of the gas atoms in the two-sided box. Let the total number of atoms be one (figure f, top), so that the number r of atoms on the right-hand side is either 0 or 1, with equal probability. By normalization, each of these probabilities is $1/2$, and $\bar{r} = 1/2$ as well. A calculation ([2336](#)) shows that the standard deviation is also $1/2$, which makes sense: $1/2$ is not just a typical value for how much r differs from \bar{r} , it is *always* the size of that deviation.

We now have an easy way to estimate the sizes of fluctuations in r when the number of atoms is larger. (On p. 172 we did this by a technique that was a lot more work.) Say there are $n = 5$ atoms, as in the bottom of figure f. If this is an ideal gas, then the atoms don't interact with each other often enough to matter, and there should be no correlation between finding one atom on the right and another atom there. Therefore the variances add. The variance in r contributed by one atom is $(1/2)^2 = 1/4$. Therefore the total variance for 5 atoms is $5/4$, and the standard deviation of r is $\sqrt{5}/2$.

This is the justification for our claim on p. 172 that when n is large, the fluctuations in r are negligible compared to r . For the relative size of the fluctuations, we should typically have $\sigma_r/\bar{r} = (\sqrt{n}/2)/(n/2) = 1/\sqrt{n}$. When n is 10^{22} , for example, the relative size of the fluctuations should be 10^{-11} , which is much too small to measure in any experiment. This is similar to the idea of the “law of averages,” which decrees that the casino always makes a profit by the end of the month.

14.3 Errors in random counts: Poisson statistics

If you do a lab experiment as part of this course in which you count radioactive decays with a Geiger counter, the number of counts N in a fixed time period will have some standard deviation σ_N . This is an example of a more generally occurring situation in statistics, which is that we have a large number of things that may happen, each with some small probability, and we count them up. The total number of them that do happen, N , is called a Poisson (“Pwa-SAN”) random variable. For example, the number of houses burglarized in Fullerton this year is a Poisson random variable. When you count the number of nuclear decays in a certain time interval, the result is Poisson. The helpful thing to know is that when a Poisson variable has an average value N , its statistical uncertainty is \sqrt{N} . So for example if your Geiger counter counts 100 clicks in one minute, this is 100 ± 10 . We could anticipate based on almost the same reasoning as in section 14.2, p. 328, that the standard deviation would be proportional to \sqrt{N} . It just so happens that the constant of proportionality for a Poisson random variable equals one.

14.4 Exponential decay

14.4.1 Half-life

Most people know that radioactivity “lasts a certain amount of time,” but that simple statement leaves out a lot. As an example, consider the following medical procedure used to diagnose thyroid function. A very small quantity of the isotope ^{131}I , produced in a nuclear reactor, is fed to or injected into the patient. The body’s biochemical systems treat this artificial, radioactive isotope exactly the same as ^{127}I , which is the only naturally occurring type. (Nutritionally, iodine is a necessary trace element. Iodine taken into the body is partly excreted, but the rest becomes concentrated in the thyroid gland. Iodized salt has had iodine added to it to prevent the nutritional deficiency known as goiters, in which the iodine-starved thyroid becomes swollen.) As the ^{131}I undergoes beta decay, it emits electrons, neutrinos, and gamma rays. The gamma rays can be measured by a detector passed over the patient’s body. As the radioactive iodine becomes concentrated in the thyroid, the amount of gamma radiation coming from the thyroid becomes greater, and that emitted by the rest of the body is reduced. The rate at which the iodine concentrates in the thyroid tells the doctor about the health of the thyroid.

If you ever undergo this procedure, someone will presumably explain a little about radioactivity to you, to allay your fears that you will turn into the Incredible Hulk, or that your next child will have an unusual number of limbs. Since iodine stays in your thyroid

for a long time once it gets there, one thing you'll want to know is whether your thyroid is going to become radioactive forever. They may just tell you that the radioactivity "only lasts a certain amount of time," but we can now carry out a quantitative derivation of how the radioactivity really will die out.

Let $P(t)$ be the probability that an iodine atom will survive without decaying for a period of at least t . It has been experimentally measured that half all ^{131}I atoms decay in 8 hours, so we have

$$P(8 \text{ hr}) = 0.5.$$

Now using the law of independent probabilities, the probability of surviving for 16 hours equals the probability of surviving for the first 8 hours multiplied by the probability of surviving for the second 8 hours,

$$\begin{aligned} P(16 \text{ hr}) &= 0.50 \times 0.50 \\ &= 0.25. \end{aligned}$$

Similarly we have

$$\begin{aligned} P(24 \text{ hr}) &= 0.50 \times 0.5 \times 0.5 \\ &= 0.125. \end{aligned}$$

Generalizing from this pattern, the probability of surviving for any time t that is a multiple of 8 hours is

$$P(t) = 0.5^{t/8 \text{ hr}}.$$

We now know how to find the probability of survival at intervals of 8 hours, but what about the points in time in between? What would be the probability of surviving for 4 hours? Well, using the law of independent probabilities again, we have

$$P(8 \text{ hr}) = P(4 \text{ hr}) \times P(4 \text{ hr}),$$

which can be rearranged to give

$$\begin{aligned} P(4 \text{ hr}) &= \sqrt{P(8 \text{ hr})} \\ &= \sqrt{0.5} \\ &= 0.707. \end{aligned}$$

This is exactly what we would have found simply by plugging in $P(t) = 0.5^{t/8 \text{ hr}}$ and ignoring the restriction to multiples of 8 hours. Since 8 hours is the amount of time required for half of the atoms to decay, it is known as the half-life, written $t_{1/2}$. The general rule is then the exponential decay equation

$$P(t) = 2^{-t/t_{1/2}}. \quad [\text{probability of survival for time } t]$$

14.4.2 Calculations for exponential decay

We'll see that all our formulas come out simpler if we state them in terms of the average lifetime τ rather than the half-life $t_{1/2}$. These are related by $\tau = t_{1/2} / \ln 2$.

Also, it's a little awkward doing exponentials with base 2. Usually we prefer to get everything in terms of base e . This has advantages, for example, when we do calculus, because it's easy to differentiate or integrate e^x , but hard to do those things with 2^x . Using the identity $2^x = \exp(\ln 2^x) = e^{x \ln 2}$, we find that the probability of survival is given by

$$P(t) = e^{-t/\tau}. \quad [\text{probability of survival for time } t]$$

We would like to know the probability distribution $D(t)$ for the time at which decay occurs. Since the survival probability is $P(t) = \int_t^\infty D(t') dt'$, the fundamental theorem of calculus gives $D(t) = -dP/dt$, or

$$D(t) = \frac{1}{\tau} e^{-t/\tau}. \quad [\text{prob. dist. of the time of decay } t]$$

We can see that the units of this equation make sense, since the probability distribution for a random variable with units of seconds must itself have units of inverse seconds.

If you're fiddling around with a hunk of plutonium and want to know how badly your chromosomes are getting nuked, then you're interested in the rate of decay, i.e., the number of decays per second, dN/dt . As the cumulative number of decays goes up, the number of survivors goes down, so $dN = N_0 D(t) dt$, where N_0 is the initial number of nuclei in your sample, and

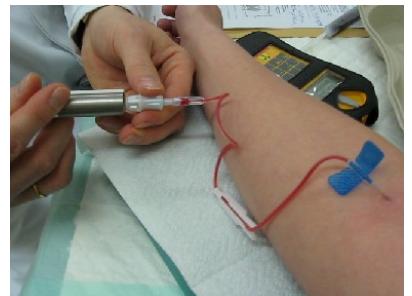
$$\frac{dN}{dt} = \frac{N_0}{\tau} e^{-t/\tau}. \quad [\text{rate of decay}]$$

Note that the three functions above are all basically the same exponential function, just with different constant factors out in front. This is because they're integrals and derivatives of each other, and integrating or differentiating e^{bx} gives back the same function with a factor of b or $1/b$.

A heart test

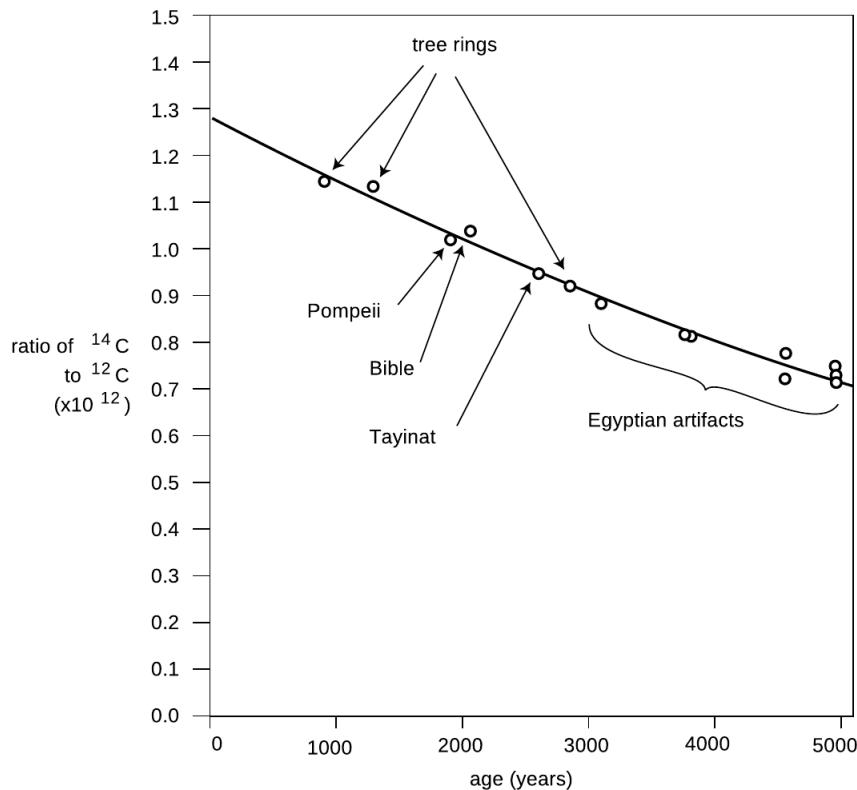
example 2

- ▷ In a common heart test, the patient is injected with a molecule containing ^{99}Tc (technetium-99) atoms in an excited state. This state decays by emitting a gamma ray, with a half-life of 6.01 hours. The molecules bind to red blood cells, so a gamma-ray video camera can see the flow of blood through the chambers of the heart. Once the chemical for the injection has been prepared, it has to be used fairly promptly. Suppose that usable medical results require that at least 40% of the ^{99}Tc nuclei remain in their excited state. What is the shelf life of the chemical?



g / A patient being injected with a radiological marker containing an excited state of ^{99}Tc . The syringe is surrounded with a thin layer of shielding in order to reduce the radiation exposure to the technician's hands. The gamma rays are relatively low in energy, so not much shielding is needed in order to stop almost all of them.

▷ Once we start doing math, it's easier to work with the mean lifetime, which in this case is $\tau = t_{1/2} / \ln 2 = 8.67$ hr. We have $P(t) = e^{-t/\tau}$, so taking logs of both sides gives $\ln P = -t/\tau$, and $t = -\tau(\ln 0.40) = 7.9$ hr. This is a little more than the half-life, which makes sense, because 0.4 is a little less than 0.5.



h / Calibration of the ^{14}C dating method using tree rings and artifacts whose ages were known from other methods. Redrawn from Emilio Segrè, **Nuclei and Particles**, 1965.

^{14}C Dating

example 3

Almost all the carbon on Earth is ^{12}C , but not quite. The isotope ^{14}C , with a half-life of 5600 years, is produced by cosmic rays in the atmosphere. It decays naturally, but is replenished at such a rate that the fraction of ^{14}C in the atmosphere remains constant, at 1.3×10^{-12} . Living plants and animals take in both ^{12}C and ^{14}C from the atmosphere and incorporate both into their bodies. Once the living organism dies, it no longer takes in C atoms from the atmosphere, and the proportion of ^{14}C gradually falls off as it undergoes radioactive decay. This effect can be used to find the age of dead organisms, or human artifacts made from plants or animals. Figure h on page 332 shows the exponential decay curve of ^{14}C in various objects. Similar methods, using longer-lived isotopes, provided the first firm proof that the earth was billions of

years old, not a few thousand as some had claimed on religious grounds.

Discussion questions

A In the medical procedure involving ^{99}Tc , example 2, why is it the gamma rays that are detected, not the electrons or neutrinos that are also emitted?

B For 1 s, Fred holds in his hands 1 kg of radioactive stuff with a half-life of 1000 years. Ginger holds 1 kg of a different substance, with a half-life of 1 min, for the same amount of time. Did they place themselves in equal danger, or not?

C Does the half-life depend on how much of the substance you have? Does the expected time until the sample decays completely depend on how much of the substance you have?

14.5 A first glimpse of quantum physics

Let's step back and think for a moment about the subversive physics assumptions behind all of our derivations about radioactive decay curves. Back around 1900, when the nucleus and radioactivity were first discovered, there were no clear principles underlying this sort of thing. In observations, it *seemed* like nuclear decay was random, and observations *seemed* to show that a nucleus's probability of decaying during a certain time interval was statistically independent of its previous history. Nobody had any idea *why* these things were true. One might expect that any answers to questions like these would be very technical, and would have to await a deeper understanding of the nucleus.

Actually these facts about nuclear decay require even less detailed technical knowledge of nuclear physics than you have from sec. 13.4. They arise from certain more basic facts of life concerning quantum physics. Let's preview these facts. I've stated these below in nonmathematical and sometimes somewhat facetious language, but the phrases, or variations on them, are the ones actually used by physicists. In pop culture, there is a tendency to over-sell quantum physics as if it were even more mysterious than it really is, or as if it had more implications than it really does for philosophy, religion, consciousness, and so on. Because of these issues, I've added footnotes below each statement to try to clarify where they really come from, what physicists mean by them, and which ones have more rigorous mathematical formulations.

1. *Totalitarian principle*: Everything not forbidden is compulsory. That is, if a process *can* take place without violating a conservation law, it *will* take place, with some probability.¹
2. *Ground state*: Every system has a lower bound on its energy. This is some number such that no state has any energy lower than that number.² In many cases, there is exactly one state, called the ground state, that has the lowest energy.
3. *State fundamentalism*: There is nothing more that can be known about a system than its state.³

The first principle is arguably *the* fundamental difference between quantum physics and classical physics (i.e., all the physical theories that came before). In classical physics, pigs can't fly, particles can't transmute themselves into other particles, and a marble locked inside a box can't get out unless we open the lid. In quantum physics, we have to ask ourselves whether there is some conservation law that prevents these things from happening, and if there isn't such a law, we expect that the process will happen, albeit possibly at a rate that is too low to measure. I suspect that flying pigs probably do violate conservation of energy (sorry, pigs), but particles do transmute themselves in various ways, and the marble definitely can get out of the box through a process called quantum tunneling, although we can estimate the rate, and it is very low. (See example 2, p. 398, for a crude estimate for the marble's chances of escape.)

The second principle, a lower bound on energies, seems to be true because we see forms of matter in our universe that seem to be either relatively stable or even (apparently) completely stable. Examples include the proton (if isolated rather than inside a nucleus) and black holes. If there was always some lower-energy state to decay to, then based on the totalitarian principle, every system would have something that it could decay to without violating conservation of energy, and therefore no system would be stable.

The third principle says that, whatever these “states” are, that’s all there is. There is nothing else to know, no higher reality, no deeper insight to be gained, nothing else that can be measured or

¹The facetious wording’s first published application to the description of quantum physics was in a footnote in a 1956 paper by Murray Gell-Mann. There is no rigorous version of this principle that forms a foundational principle of quantum physics, but ever since the birth of quantum mechanics, physicists have found it to be an excellent informal guide to reasoning.

²This is a mathematically rigorous statement. Although it is not typically included in formal axiomatizations of quantum physics, any theory that violates this principle is normally discarded as unrealistic. Specialists call it the spectrum condition.

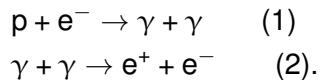
³This is a more vaguely defined philosophical statement that does not have a complete and rigorous mathematical formulation. People working on the philosophy and foundations of quantum mechanics use a variety of related phrases, such as “wavefunction fundamentalism” and “state monism.”

observed about the system. This principle is what allows us to assert that a nucleus's probability of decay is independent of its history, as assumed in our derivation of the exponential decay equation. When we see a nucleus that's been sitting around for a while, it's normally in its ground state. Therefore there is nothing else to know about it besides the fact that it's in its ground state. It can't have cracks and strain that show it's about to decay, nor can we see that it must be a really tough little nucleus because it's survived for such a long time. (A more detailed description of exponential decay is given in example 12, p. 417.) Another application of this principle is discussed in sec. 17.8.1, p. 413.

Two processes involving positrons

example 4

Around 1930, Paul Dirac proposed that each of the following processes might occur:



In process (1), a proton and an electron annihilate, creating two gamma rays. This would seem to imply that the hydrogen atom would be unstable with respect to radioactive decay, which seems like a daring prediction, although perhaps Dirac thought the rate at which the annihilation would occur would be so low that it would not yet have been noticed in laboratory experiments. He wrote, "There appears to be no reason why such processes should not actually occur somewhere in the world. They would be consistent with all the general laws of Nature." This sounds like a use of the totalitarian principle.

But in fact, process (1) has never been observed. Because physicists tend to believe in the totalitarian principle, they need a conservation law to explain why it doesn't occur. The process is consistent, however, with both conservation of energy-momentum (p. 148) and conservation of charge. Therefore, we invent a new conservation law. In fact, we have *two* conservation laws at this point that forbid this decay, of which I'll discuss only one. This one is called conservation of lepton number. "Lepton" is a general term that refers to particles like electrons and positrons, as well as some other, similar particles that are unstable. The electron has a lepton number of +1, while the lepton numbers of the other particles in this process are zero.

Process (2) actually does occur — it is the time-reversed version of the process of electron-positron annihilation (p. 77). It was in fact this process that allowed the original experimental discovery of the positron in 1932. This process obeys conservation of energy-momentum, conservation of charge, and conservation of lepton number, because the antielectron has lepton number -1.

Notes for chapter 14

328 Standard deviation of r with one atom

For a single atom in a box, the standard deviation of the number of atoms on the right is $1/2$.

We have a single atom in a box, with $r = 1$ if the atom is in the right half and $r = 0$ otherwise. Because r is discrete, the variance $\langle(r - \bar{r})^2\rangle$ can be computed as a probability-weighted sum,

$$\begin{aligned}\text{(variance of } r\text{)} &= \langle \left(r - \frac{1}{2}\right)^2 \rangle \\ &= P(r = 0) \left(0 - \frac{1}{2}\right)^2 \\ &\quad + P(r = 1) \left(1 - \frac{1}{2}\right)^2 \\ &= \frac{1}{4}.\end{aligned}$$

The standard deviation of r is then $\sqrt{1/4} = 1/2$, as expected.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 If a radioactive substance has a half-life of one year, does this mean that it will be completely decayed after two years? Explain.

2 A nuclear physicist is studying a nuclear reaction caused in an accelerator experiment, with a beam of ions from the accelerator striking a thin metal foil and causing nuclear reactions when a nucleus from one of the beam ions happens to hit one of the nuclei in the target. After the experiment has been running for a few hours, a few billion radioactive atoms have been produced, embedded in the target. She does not know what nuclei are being produced, but she suspects they are an isotope of some heavy element such as Pb, Bi, Fr or U. Following one such experiment, she takes the target foil out of the accelerator, sticks it in front of a detector, measures the activity every 5 min, and makes a graph (figure). The isotopes she thinks may have been produced are:

isotope half-life (minutes)

^{211}Pb 36.1

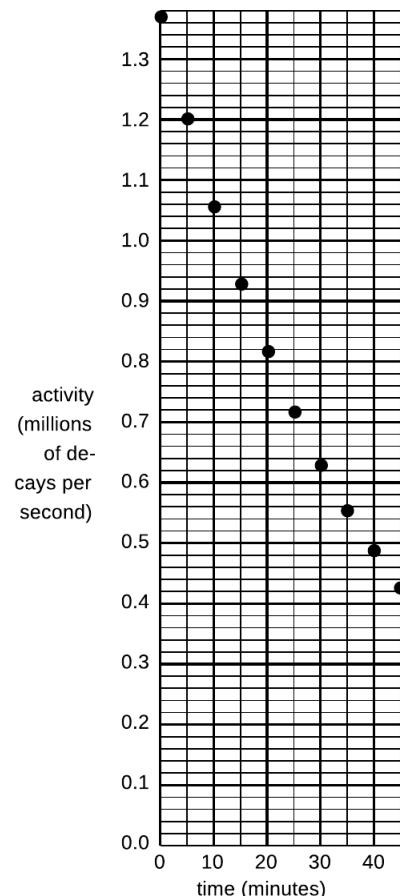
^{214}Pb 26.8

^{214}Bi 19.7

^{223}Fr 21.8

^{239}U 23.5

Which one is it? [Part b of this problem has been deleted.]



Problem 2.

3 Refer to the probability distribution for people's heights in figure c on page 326.

- (a) Show that the graph is properly normalized.
- (b) Estimate the fraction of the population having heights between 140 and 150 cm. ✓

4 A blindfolded person fires a gun at a circular target of radius b , and is allowed to continue firing until a shot actually hits it. Any part of the target is equally likely to get hit. We measure the random distance r from the center of the circle to where the bullet went in.

- (a) Show that the probability distribution of r must be of the form $D(r) = kr$, where k is some constant. (Of course we have $D(r) = 0$ for $r > b$.)
- (b) Determine k by requiring D to be properly normalized. ✓
- (c) Find the average value of r . ✓
- (d) Interpreting your result from part c, how does it compare with $b/2$? Does this make sense? Explain.

5 We are given some atoms of a certain radioactive isotope, with half-life $t_{1/2}$. We pick one atom at random, and observe it for one half-life, starting at time zero. If it decays during that one-half-life period, we record the time t at which the decay occurred. If it doesn't, we reset our clock to zero and keep trying until we get an atom that cooperates. The final result is a time $0 \leq t \leq t_{1/2}$, with a distribution that looks like the usual exponential decay curve, but with its tail chopped off.

- (a) Find the distribution $D(t)$, with the proper normalization. ✓
- (b) Find the average value of t . ✓
- (c) Interpreting your result from part b, how does it compare with $t_{1/2}/2$? Does this make sense? Explain.

6 The speed, v , of an atom in an ideal gas has a probability distribution of the form $D(v) = bve^{-cv^2}$, where $0 \leq v < \infty$, c relates to the temperature, and b is determined by normalization.

- (a) Sketch the distribution.
- (b) Find b in terms of c . ✓
- (c) Find the average speed in terms of c , eliminating b . (Don't try to do the indefinite integral, because it can't be done in closed form. The relevant definite integral can be found in tables or done with computer software.) ✓

7 All helium on earth is from the decay of naturally occurring heavy radioactive elements such as uranium. Each alpha particle that is emitted ends up claiming two electrons, which makes it a helium atom. If the original ^{238}U atom is in solid rock (as opposed to the earth's molten regions), the He atoms are unable to diffuse out of the rock. This problem involves dating a rock using the known decay properties of uranium 238. Suppose a geologist finds a sample of hardened lava, melts it in a furnace, and finds that it contains 1230 mg of uranium and 2.3 mg of helium. ^{238}U decays by alpha emission, with a half-life of 4.5×10^9 years. The subsequent chain of alpha and electron (beta) decays involves much shorter half-lives, and terminates in the stable nucleus ^{206}Pb . Almost all natural uranium is ^{238}U , and the chemical composition of this rock indicates that there were no decay chains involved other than that of ^{238}U .

- (a) How many alphas are emitted per decay chain? [Hint: Use conservation of mass.]
- (b) How many electrons are emitted per decay chain? [Hint: Use conservation of charge.]
- (c) How long has it been since the lava originally hardened? ✓

8 In the year 2010, Fullerton, California, had 342 car thefts. In the following year, the number was 359. Was this a statistically significant increase?

9 New isotopes are continually being produced and studied. A common method is that experimenters produce a beam of nuclei in an accelerator, and the beam strikes a target such as a thin metal foil. If a beam nucleus happens to hit a target nucleus, nuclear fusion can occur. Once the fused nucleus is formed, it is common for several neutrons to boil off, and the number of neutrons lost can be random, so that more than one isotope can be produced in the same experiment.

Liza carries out such an experiment and observes beta particles being emitted afterward, meaning that she has produced an isotope that is radioactive. She counts the number of betas observed in her detector for the first three hours after the isotope has been produced, with the following results:

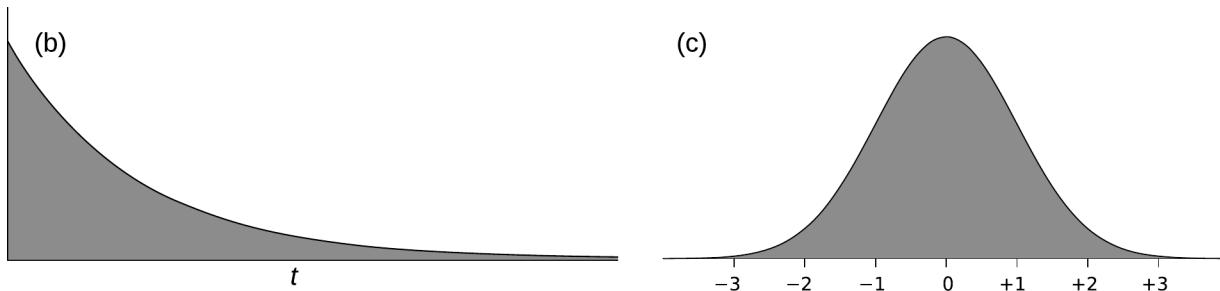
first hour	770,336
second hour	662,901
third hour	582,813

Is this a decay curve that could be statistically consistent with a single half-life, i.e., with the production of a single isotope?

Exercise 14: Probability distributions

Questions 1-3 involve the useful concept of the *cumulative distribution* (not introduced in the text). Let $P(x) = \int_{-\infty}^x D(x')dx'$ be the probability of finding a value for the random number that is less than or equal to x . The function P is referred to as the cumulative distribution. (In the context of radioactive decay, the survival probability referred to in the text is actually $1 - P$ for this definition of P .)

1. Using the fundamental theorem of calculus, express D in terms of P . Use Leibniz (“d”) notation.
2. Suppose that x has some units such as kilograms. Use one of the relations between D and P to determine the units of both functions, and then check that it also works out according to the other relation.
3. Sketch the functions D and P for the following random variables: (a) a random real number that has a uniform probability of lying anywhere in the interval from 0 to 1; (b) the time t at which an atom of a radioactive isotope decays, shown by the exponential curve below; (c) a random number that follows the standard “bell curve,” shown below, with an average value of 0 and a standard deviation of 1; (d) the result of a die roll. You will not find that both D and P are well defined in all cases.



Pause. I'll walk you through the following problem on the board. A certain electron wave confined to a box of length L has a probability distribution given by

$$D(x) = \begin{cases} A \sin^2(2\pi x/L) & \text{if } 0 \leq x \leq L \\ 0 & \text{elsewhere} \end{cases}$$

Sketch the function. Infer the units of A . What would go wrong if the sine function wasn't squared? Determine A from the requirement of normalization.

4. The earth is constantly exposed to neutrinos from outer space. Most neutrinos pass through the entire planet without interacting, but a small fraction of them are absorbed. These absorption events are distributed uniformly within the spherical volume of the earth, which is of radius b . Let r be the distance from the center at which one of these events occurs, so that $0 \leq r \leq b$.
 - (a) Sketch the geometrical situation, showing an infinitesimally thin shell that stretches from r to $r + dr$.
 - (b) Show that the probability distribution $D(r)$ is of the form kr^p , where k and p are constants, and determine p . Hint: consider the volume of the shell you sketched in part a.
 - (c) Determine k .
 - (d) Find the average value of r .

Pause. Let x be a random variable with a known distribution $D(x)$, and let $y = f(x)$ be a function of x . I'll show you how to use the chain rule to determine the probability distribution $D^*(y)$, where the star indicates that D^* is a different function from D .

5. A drunk guy in a bar offers you the following bet. He holds a sharp razor blade facing up, and you toss a stick of uncooked spaghetti up in the air so that when it comes down, the blade breaks it at a random point. The spaghetti is 1 foot long, and the break point is a random variable x . He offers to pay you $y = 1/x$ in units of dollars. You want to determine whether this is a sucker bet and the guy is a scammer. Find the amount of money you should be willing to pay in return for the expected payout.

Chapter 15

Light as a particle

The only thing that interferes with my learning is my education.

Albert Einstein

Radioactivity is random, but do the laws of physics exhibit randomness in other contexts besides radioactivity? Yes. Radioactive decay was just a good playpen to get us started with concepts of randomness, because all atoms of a given isotope are identical. By stocking the playpen with an unlimited supply of identical atom-toys, nature helped us to realize that their future behavior could be different regardless of their original identicality. We are now ready to leave the playpen, and see how randomness fits into the structure of physics at the most fundamental level.

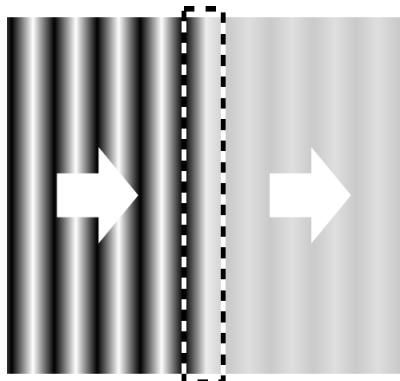
The laws of physics describe light and matter, and the quantum revolution rewrote both descriptions. Radioactivity was a good example of matter's behaving in a way that was inconsistent with classical physics, but if we want to get under the hood and understand how nonclassical things happen, it will be easier to focus on light rather than matter. A radioactive atom such as uranium-235 is after all an extremely complex system, consisting of 92 protons, 143 neutrons, and 92 electrons. Light, however, can be a simple sine wave.

However successful the classical wave theory of light had been — allowing the creation of radio and radar, for example — it still failed to describe many important phenomena. An example that is currently of great interest is the way the ozone layer protects us from the dangerous short-wavelength ultraviolet part of the sun's spectrum. In the classical description, light is a wave. When a wave passes into and back out of a medium, its frequency is unchanged, and although its wavelength is altered while it is in the medium, it returns to its original value when the wave reemerges. Luckily for us, this is not at all what ultraviolet light does when it passes through the ozone layer, or the layer would offer no protection at all!

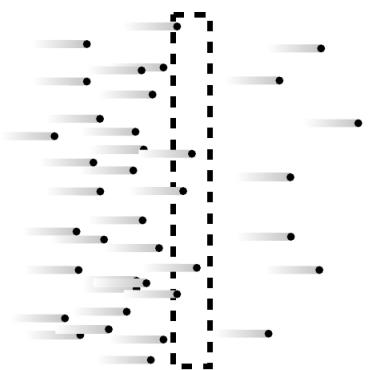
15.1 Evidence for light as a particle

For a long time, physicists tried to explain away the problems with the classical theory of light as arising from an imperfect understanding of atoms and the interaction of light with individual atoms and

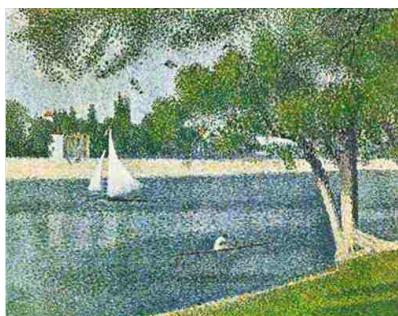
molecules. The ozone paradox, for example, could have been attributed to the incorrect assumption that one could think of the ozone layer as a smooth, continuous substance, when in reality it was made of individual ozone molecules. It wasn't until 1905 that Albert Einstein threw down the gauntlet, proposing that the problem had nothing to do with the details of light's interaction with atoms and everything to do with the fundamental nature of light itself.



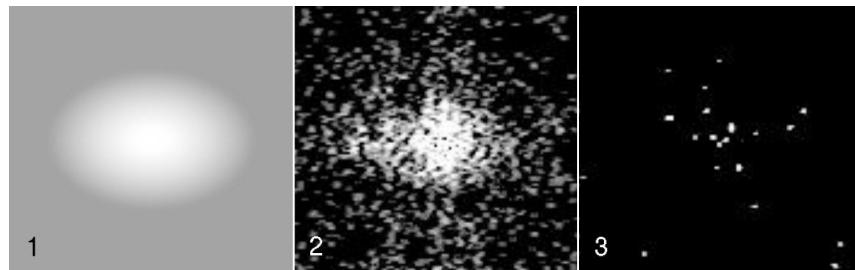
b / A wave is partially absorbed.



c / A stream of particles is partially absorbed.



d / Einstein and Seurat: twins separated at birth? *Seine Grande Jatte* by Georges Seurat (19th century).



a / Digital camera images of dimmer and dimmer sources of light. The dots are records of individual photons.

In those days the data were sketchy, the ideas vague, and the experiments difficult to interpret; it took a genius like Einstein to cut through the thicket of confusion and find a simple solution. Today, however, we can get right to the heart of the matter with a piece of ordinary consumer electronics, the digital camera. Instead of film, a digital camera has a computer chip with its surface divided up into a grid of light-sensitive squares, called "pixels." Compared to a grain of the silver compound used to make regular photographic film, a digital camera pixel is activated by an amount of light energy orders of magnitude smaller. We can learn something new about light by using a digital camera to detect smaller and smaller amounts of light, as shown in figure a. Figure a/1 is fake, but a/2 and a/3 are real digital-camera images made by Prof. Lyman Page of Princeton University as a classroom demonstration. Figure a/1 is what we would see if we used the digital camera to take a picture of a fairly dim source of light. In figures a/2 and a/3, the intensity of the light was drastically reduced by inserting semitransparent absorbers like the tinted plastic used in sunglasses. Going from a/1 to a/2 to a/3, more and more light energy is being thrown away by the absorbers.

The results are drastically different from what we would expect based on the wave theory of light. If light was a wave and nothing but a wave, b, then the absorbers would simply cut down the wave's amplitude across the whole wavefront. The digital camera's entire chip would be illuminated uniformly, and weakening the wave with an absorber would just mean that every pixel would take a long time to soak up enough energy to register a signal.

But figures a/2 and a/3 show that some pixels take strong hits while others pick up no energy at all. Instead of the wave picture, the image that is naturally evoked by the data is something more like a hail of bullets from a machine gun, c. Each “bullet” of light apparently carries only a tiny amount of energy, which is why detecting them individually requires a sensitive digital camera rather than an eye or a piece of film.

Although Einstein was interpreting different observations, this is the conclusion he reached in his 1905 paper: that the pure wave theory of light is an oversimplification, and that the energy of a beam of light comes in finite chunks rather than being spread smoothly throughout a region of space.

We now think of these chunks as particles of light, and call them “photons,” although Einstein avoided the word “particle,” and the word “photon” was invented later. Regardless of words, the trouble was that waves and particles seemed like inconsistent categories. The reaction to Einstein’s paper could be kindly described as vigorously skeptical. Even twenty years later, Einstein wrote, “There are therefore now two theories of light, both indispensable, and — as one must admit today despite twenty years of tremendous effort on the part of theoretical physicists — without any logical connection.” In the remainder of this section we will learn how the seeming paradox was eventually resolved.

Discussion questions

A Suppose someone rebuts the digital camera data in figure a, claiming that the random pattern of dots occurs not because of anything fundamental about the nature of light but simply because the camera’s pixels are not all exactly the same — some are just more sensitive than others. How could we test this interpretation?

B Discuss how the correspondence principle applies to the observations and concepts discussed in this section.

15.2 How much light is one photon?

15.2.1 The photoelectric effect

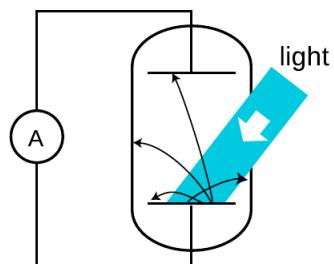
We have seen evidence that light energy comes in little chunks, so the next question to be asked is naturally how much energy is in one chunk. The most straightforward experimental avenue for addressing this question is a phenomenon known as the photoelectric effect. The photoelectric effect occurs when a photon strikes the surface of a solid object and knocks out an electron. It occurs continually all around you. It is happening right now at the surface of your skin and on the paper or computer screen from which you are reading these words. It does not ordinarily lead to any observable electrical effect, however, because on the average free electrons are wandering back in just as frequently as they are being ejected. (If an object did somehow lose a significant number of electrons, its growing net positive charge would begin attracting the electrons back more and more strongly.)

Figure e shows a practical method for detecting the photoelectric effect. Two very clean parallel metal plates (the electrodes of a capacitor) are sealed inside a vacuum tube, and only one plate is exposed to light. Because there is a good vacuum between the plates, any ejected electron that happens to be headed in the right direction will almost certainly reach the other capacitor plate without colliding with any air molecules.

The illuminated (bottom) plate is left with a net positive charge, and the unilluminated (top) plate acquires a negative charge from the electrons deposited on it. There is thus an electric field between the plates, and it is because of this field that the electrons' paths are curved, as shown in the diagram. However, since vacuum is a good insulator, any electrons that reach the top plate are prevented from responding to the electrical attraction by jumping back across the gap. Instead they are forced to make their way around the circuit, passing through an ammeter. The ammeter allows a measurement of the strength of the photoelectric effect.

15.2.2 An unexpected dependence on frequency

The photoelectric effect was discovered serendipitously by Heinrich Hertz in 1887, as he was experimenting with radio waves. He was not particularly interested in the phenomenon, but he did notice that the effect was produced strongly by ultraviolet light and more weakly by lower frequencies. Light whose frequency was lower than a certain critical value did not eject any electrons at all. (In fact this was all prior to Thomson's discovery of the electron, so Hertz would not have described the effect in terms of electrons — we are discussing everything with the benefit of hindsight.) This dependence on frequency didn't make any sense in terms of the classical wave theory of light. A light wave consists of electric and magnetic



e / Apparatus for observing the photoelectric effect. A beam of light strikes a capacitor plate inside a vacuum tube, and electrons are ejected (black arrows).

fields. The stronger the fields, i.e., the greater the wave's amplitude, the greater the forces that would be exerted on electrons that found themselves bathed in the light. It should have been amplitude (brightness) that was relevant, not frequency. The dependence on frequency not only proves that the wave model of light needs modifying, but with the proper interpretation it allows us to determine how much energy is in one photon, and it also leads to a connection between the wave and particle models that we need in order to reconcile them.

To make any progress, we need to consider the physical process by which a photon would eject an electron from the metal electrode. A metal contains electrons that are free to move around. Ordinarily, in the interior of the metal, such an electron feels attractive forces from atoms in every direction around it. The forces cancel out. But if the electron happens to find itself at the surface of the metal, the attraction from the interior side is not balanced out by any attraction from outside. In popping out through the surface the electron therefore loses some amount of energy E_s , which depends on the type of metal used.

Suppose a photon strikes an electron, annihilating itself and giving up all its energy to the electron. (We now know that this is what always happens in the photoelectric effect, although it had not yet been established in 1905 whether or not the photon was completely annihilated.) The electron will (1) lose kinetic energy through collisions with other electrons as it plows through the metal on its way to the surface; (2) lose an amount of kinetic energy equal to E_s as it emerges through the surface; and (3) lose more energy on its way across the gap between the plates, due to the electric field between the plates. Even if the electron happens to be right at the surface of the metal when it absorbs the photon, and even if the electric field between the plates has not yet built up very much, E_s is the bare minimum amount of energy that it must receive from the photon if it is to contribute to a measurable current. The reason for using very clean electrodes is to minimize E_s and make it have a definite value characteristic of the metal surface, not a mixture of values due to the various types of dirt and crud that are present in tiny amounts on all surfaces in everyday life.

We can now interpret the frequency dependence of the photoelectric effect in a simple way: apparently the amount of energy possessed by a photon is related to its frequency. A low-frequency red or infrared photon has an energy less than E_s , so a beam of them will not produce any current. A high-frequency blue or violet photon, on the other hand, packs enough of a punch to allow an electron to make it to the other plate. At frequencies higher than the minimum, the photoelectric current continues to increase with the frequency of the light because of effects (1) and (3).

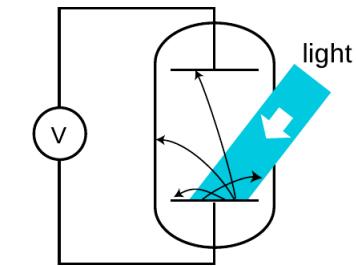


f / The hamster in her hamster ball is like an electron emerging from the metal (tiled kitchen floor) into the surrounding vacuum (wood floor). The wood floor is higher than the tiled floor, so as she rolls up the step, the hamster will lose a certain amount of kinetic energy, analogous to E_s . If her kinetic energy is too small, she won't even make it up the step.

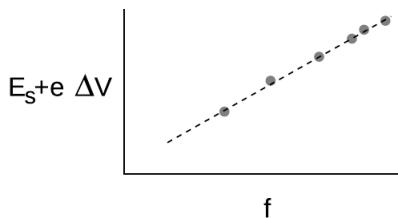
15.2.3 Numerical relationship between energy and frequency

Figure g shows an experiment that is used sometimes in college laboratory courses to probe the relationship between the energy and frequency of a photon. The idea is simply to illuminate one plate of the vacuum tube with light of a single wavelength and monitor the voltage difference between the two plates as they charge up. Since the resistance of a voltmeter is very high (much higher than the resistance of an ammeter), we can assume to a good approximation that electrons reaching the top plate are stuck there permanently, so the voltage will keep on increasing for as long as electrons are making it across the vacuum tube.

At a moment when the voltage difference has reached a value ΔV , the minimum energy required by an electron to make it out of the bottom plate and across the gap to the other plate is $E_s + e\Delta V$. As ΔV increases, we eventually reach a point at which $E_s + e\Delta V$ equals the energy of one photon. No more electrons can cross the gap, and the reading on the voltmeter stops rising. The quantity $E_s + e\Delta V$ now tells us the energy of one photon. If we determine this energy for a variety of wavelengths, h , we find the following simple relationship between the energy of a photon and the frequency of the light:



g / A different way of studying the photoelectric effect.



h / The quantity $E_s + e\Delta V$ indicates the energy of one photon. It is found to be proportional to the frequency of the light.

$$E = hf,$$

where h is a constant with the value $6.63 \times 10^{-34} \text{ J} \cdot \text{s}$. Note how the equation brings the wave and particle models of light under the same roof: the left side is the energy of one *particle* of light, while the right side is the frequency of the same light, interpreted as a *wave*. The constant h is known as Planck's constant, for historical reasons explained in the footnote beginning on the preceding page.

self-check A

How would you extract h from the graph in figure h? What if you didn't even know E_s in advance, and could only graph $e\Delta V$ versus f ? ▷ Answer, p. 458

Since the energy of a photon is hf , a beam of light can only have energies of hf , $2hf$, $3hf$, etc. Its energy is quantized — there is no such thing as a fraction of a photon. Quantum physics gets its name from the fact that it quantizes quantities like energy, momentum, and angular momentum that had previously been thought to be smooth, continuous and infinitely divisible.

Photons from a lightbulb

example 1

▷ Roughly how many photons are emitted by a 100 watt lightbulb in 1 second?

▷ People tend to remember wavelengths rather than frequencies for visible light. The bulb emits photons with a range of frequencies and wavelengths, but let's take 600 nm as a typical wavelength for purposes of estimation. The energy of a single photon is

$$\begin{aligned}E_{\text{photon}} &= hf \\&= hc/\lambda\end{aligned}$$

A power of 100 W means 100 joules per second, so the number of photons is

$$\begin{aligned}(100 \text{ J})/E_{\text{photon}} &= (100 \text{ J})/(hc/\lambda) \\&\approx 3 \times 10^{20}\end{aligned}$$

This hugeness of this number is consistent with the correspondence principle. The experiments that established the classical theory of optics weren't wrong. They were right, within their domain of applicability, in which the number of photons was so large as to be indistinguishable from a continuous beam.

Measuring the wave

example 2

When surfers are out on the water waiting for their chance to catch a wave, they're interested in both the height of the waves and when the waves are going to arrive. In other words, they observe both the amplitude and phase of the waves, and it doesn't matter to them that the water is granular at the molecular level. The correspondence principle requires that we be able to do the same thing for electromagnetic waves, since the classical theory of electricity and magnetism was all stated and verified experimentally in terms of the fields **E** and **B**, which are the amplitude of an electromagnetic wave. The phase is also necessary, since the induction effects predicted by Maxwell's equation would flip their signs depending on whether an oscillating field is on its way up or on its way back down.

This is a more demanding application of the correspondence principle than the one in example 1, since amplitudes and phases constitute more detailed information than the over-all intensity of a beam of light. Eyeball measurements can't detect this type of information, since the eye is much bigger than a wavelength, but for example an AM radio receiver can do it with radio waves, since the wavelength for a station at 1000 kHz is about 300 meters, which is much larger than the antenna. The correspondence principle demands that we be able to explain this in terms of the photon theory, and this requires not just that we have a large

number of photons emitted by the transmitter per second, as in example 1, but that even by the time they spread out and reach the receiving antenna, there should be many photons overlapping each other within a space of one cubic wavelength. Problem 9 on p. 362 verifies that the number is in fact extremely large.

Momentum of a photon

example 3

▷ According to the theory of relativity, the momentum of a beam of light is given by $p = E/c$. Apply this to find the momentum of a single photon in terms of its frequency, and in terms of its wavelength.

▷ Combining the equations $p = E/c$ and $E = hf$, we find

$$\begin{aligned} p &= E/c \\ &= \frac{h}{c}f. \end{aligned}$$

To reexpress this in terms of wavelength, we use $c = f\lambda$:

$$\begin{aligned} p &= \frac{h}{c} \cdot \frac{c}{\lambda} \\ &= \frac{h}{\lambda} \end{aligned}$$

The second form turns out to be simpler.

Discussion questions

A The photoelectric effect only ever ejects a very tiny percentage of the electrons available near the surface of an object. How well does this agree with the wave model of light, and how well with the particle model? Consider the two different distance scales involved: the wavelength of the light, and the size of an atom, which is on the order of 10^{-10} or 10^{-9} m.

B What is the significance of the fact that Planck's constant is numerically very small? How would our everyday experience of light be different if it was not so small?

C How would the experiments described above be affected if a single electron was likely to get hit by more than one photon?

D Draw some representative trajectories of electrons for $\Delta V = 0$, ΔV less than the maximum value, and ΔV greater than the maximum value.

E Explain based on the photon theory of light why ultraviolet light would be more likely than visible or infrared light to cause cancer by damaging DNA molecules. How does this relate to discussion question C?

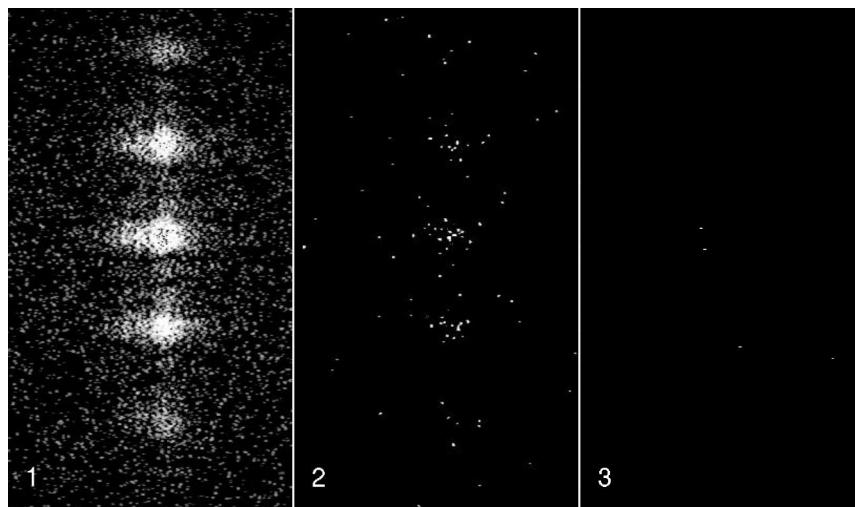
F Does $E = hf$ imply that a photon changes its energy when it passes from one transparent material into another substance with a different index of refraction?

15.3 Wave-particle duality

How can light be both a particle and a wave? We are now ready to resolve this seeming contradiction. Often in science when something seems paradoxical, it's because we (1) don't define our terms carefully, or (2) don't test our ideas against any specific real-world situation. Let's define particles and waves as follows:

- Waves exhibit superposition, and specifically interference phenomena.
- Particles can only exist in whole numbers, not fractions.

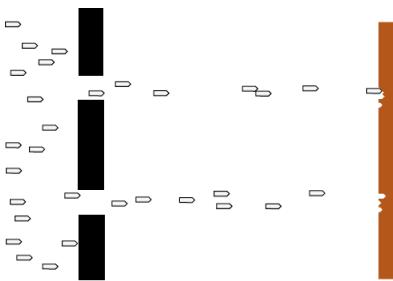
As a real-world check on our philosophizing, there is one particular experiment that works perfectly. We set up a double-slit interference experiment that we know will produce a diffraction pattern if light is an honest-to-goodness wave, but we detect the light with a detector that is capable of sensing individual photons, e.g., a digital camera. To make it possible to pick out individual dots due to individual photons, we must use filters to cut down the intensity of the light to a very low level, just as in the photos by Prof. Page on p. 344. The whole thing is sealed inside a light-tight box. The results are shown in figure i. (In fact, the similar figures in on page 344 are simply cutouts from these figures.)



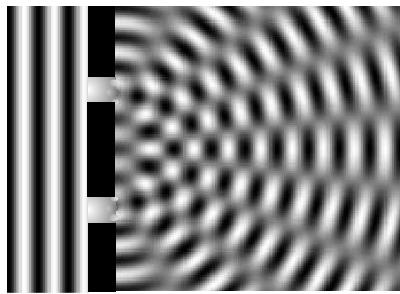
i / Wave interference patterns photographed by Prof. Lyman Page with a digital camera. Laser light with a single well-defined wavelength passed through a series of absorbers to cut down its intensity, then through a set of slits to produce interference, and finally into a digital camera chip. (A triple slit was actually used, but for conceptual simplicity we discuss the results in the main text as if it was a double slit.) In panel 2 the intensity has been reduced relative to 1, and even more so for panel 3.

Neither the pure wave theory nor the pure particle theory can explain the results. If light was only a particle and not a wave, there would be no interference effect. The result of the experiment would be like firing a hail of bullets through a double slit, j. Only two spots directly behind the slits would be hit.

If, on the other hand, light was only a wave and not a particle, we would get the same kind of diffraction pattern that would happen



j / Bullets pass through a double slit.



k / A water wave passes through a double slit.



l / A single photon can go through both slits.

with a water wave, k. There would be no discrete dots in the photo, only a diffraction pattern that shaded smoothly between light and dark.

Applying the definitions to this experiment, light must be both a particle and a wave. It is a wave because it exhibits interference effects. At the same time, the fact that the photographs contain discrete dots is a direct demonstration that light refuses to be split into units of less than a single photon. There can only be whole numbers of photons: four photons in figure i/3, for example.

15.3.1 A wrong interpretation: photons interfering with each other

One possible interpretation of wave-particle duality that occurred to physicists early in the game was that perhaps the interference effects came from photons interacting with each other. By analogy, a water wave consists of moving water molecules, and interference of water waves results ultimately from all the mutual pushes and pulls of the molecules. This interpretation has been conclusively disproved by forming interference patterns with light so dim that no more than one photon is in flight at a time. In figure i/3, for example, the intensity of the light has been cut down so much by the absorbers that if it was in the open, the average separation between photons would be on the order of a kilometer! Although most light sources tend to emit photons in bunches, experiments have been done with light sources that really do emit single photons at wide time intervals, and the same type of interference pattern is observed, showing that a single photon can interfere with *itself*.

The concept of a photon's path is undefined.

If a single photon can demonstrate double-slit interference, then which slit did it pass through? The unavoidable answer must be that it passes through both! This might not seem so strange if we think of the photon as a wave, but it is highly counterintuitive if we try to visualize it as a particle. The moral is that we should not think in terms of the path of a photon. Like the fully human and fully divine Jesus of Christian theology, a photon is supposed to be 100% wave and 100% particle. If a photon had a well defined path, then it would not demonstrate wave superposition and interference effects, contradicting its wave nature. (In sec. section 16.4 we will discuss the Heisenberg uncertainty principle, which gives a numerical way of approaching this issue.)

15.3.2 The probability interpretation

The correct interpretation of wave-particle duality is suggested by the random nature of the experiment we've been discussing: even though every photon wave/particle is prepared and released in the same way, the location at which it is eventually detected by the

digital camera is different every time. The idea of the probability interpretation of wave-particle duality is that the location of the photon-particle is random, but the probability that it is in a certain location is higher where the photon-wave's amplitude is greater.

More specifically, the probability distribution of the particle must be proportional to the *square* of the wave's amplitude,

$$(\text{probability distribution}) \propto (\text{amplitude})^2.$$

This follows from the correspondence principle and from the fact that a wave's energy density is proportional to the square of its amplitude. If we run the double-slit experiment for a long enough time, the pattern of dots fills in and becomes very smooth as would have been expected in classical physics. To preserve the correspondence between classical and quantum physics, the amount of energy deposited in a given region of the picture over the long run must be proportional to the square of the wave's amplitude. The amount of energy deposited in a certain area depends on the number of photons picked up, which is proportional to the probability of finding any given photon there.

A microwave oven

example 4

► The figure shows two-dimensional (top) and one-dimensional (bottom) representations of the standing wave inside a microwave oven. Gray represents zero field, and white and black signify the strongest fields, with white being a field that is in the opposite direction compared to black. Compare the probabilities of detecting a microwave photon at points A, B, and C.

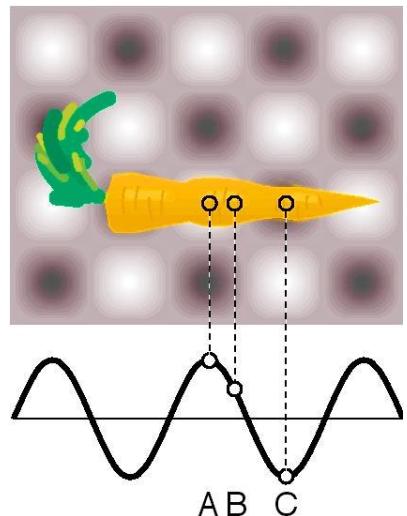
► A and C are both extremes of the wave, so the probabilities of detecting a photon at A and C are equal. It doesn't matter that we have represented C as negative and A as positive, because it is the square of the amplitude that is relevant. The amplitude at B is about 1/2 as much as the others, so the probability of detecting a photon there is about 1/4 as much.

Discussion questions

A Referring back to the example of the carrot in the microwave oven, show that it would be nonsensical to have probability be proportional to the field itself, rather than the square of the field.

B Einstein did not try to reconcile the wave and particle theories of light, and did not say much about their apparent inconsistency. Einstein basically visualized a beam of light as a stream of bullets coming from a machine gun. In the photoelectric effect, a photon "bullet" would only hit one atom, just as a real bullet would only hit one person. Suppose someone reading his 1905 paper wanted to interpret it by saying that Einstein's so-called particles of light are simply short wave-trains that only occupy a small region of space. Comparing the wavelength of visible light (a few hundred nm) to the size of an atom (on the order of 0.1 nm), explain why this poses a difficulty for reconciling the particle and wave theories.

C Can a white photon exist?



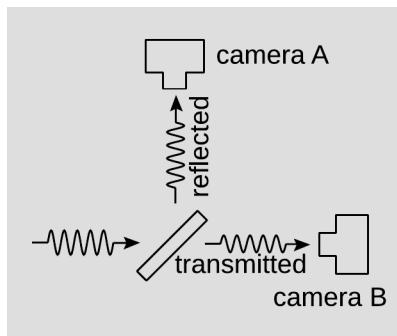
m / Example 4.

D In double-slit diffraction of photons, would you get the same pattern of dots on the digital camera image if you covered one slit? Why should it matter whether you give the photon two choices or only one?

15.4 Nonlocality and entanglement

15.4.1 Nonlocality

People sometimes say that quantum mechanics is the set of rules for describing the world of the very small, but this is a false generalization, like saying that terriers are untrainable. How do we define our measure of how small is small? The only distance scales we've discussed have been wavelengths, and there is no upper limit on wavelengths. The wavelength of an FM radio photon is bigger than my terrier, who is very obedient to Newton's laws. The only scale built in to the structure of quantum mechanics is Planck's constant, and Planck's constant has units of joules per hertz, not meters, so it can't be converted into a distance. Quantum mechanics is, as far as we can tell, a valid tool for describing systems at scales from quarks to galaxies.



n / A photon hits a piece of glass that reflects half of the light and transmits the other half.

So quantum behavior can occur at any scale, even large ones. For an example that may be a little disturbing, consider the arrangement shown in figure n. A single photon comes in from the left and encounters a diagonal piece of glass. The glass reflects half the light and transmits half of it. The photon is a wave, and this is expected wave behavior. But the photon is also a particle, and we can't have half a particle. Therefore either camera A will detect a whole photon and B will see none, or it will be the other way around. If we repeat the experiment many times, we might come up with a list of results like this:

A	B
no	yes
yes	no
yes	no
no	yes
no	yes
yes	no
no	yes
yes	no

An instant before the moment of detection, the photon is a wave pattern that just happens to consist of two widely separated pieces, each carrying half the energy. The situation seems perfectly symmetric, but then a moment later we find that B has detected the photon and A hasn't. If B's detection of the photon is random, then how does the information get to A that it had better *not* detect it? This seems as though there is some sort of conspiracy being carried out over arbitrarily large distances and with no time delay. It's as though the two parts of the wave are a pair of criminal suspects who

would like to line up their stories but are being kept in separate jail cells so that they can't communicate. If the part of the wave at B is going to be detected (at full strength, carrying 100% of the energy $E = hf$), how does the part at A get the message that it should fade away like the Cheshire cat? This coordination would have to occur over very large distances — real-world experiments of this type have been done over distances of a thousand kilometers, with the photons traveling either through outer space or through fiber-optic cables. Einstein derisively referred to this apparent coordination as “spooky action at a distance.”

Niels Bohr and two collaborators proposed in 1924 the seemingly reasonable solution that there *can't* be any such coordination. Then the random detection of the photon by camera A and camera B would be independent. Independent probabilities multiply, so there would be a probability of $(1/2)(1/2) = 1/4$ that both cameras would see photons. This would violate conservation of energy, since the original energy $E = hf$ would have been detected twice, and the universe would have gained $1hf$ worth of total energy. But Bohr pointed out that there would also be the same probability that neither camera would detect a photon, in which case the change in the universe's energy would be $-1hf$. On the average, energy would be conserved. According to Bohr's theory, conservation of energy and momentum would not be absolute laws of physics but only rules that would be true on the average.

The experimentalists Geiger and Bothe immediately set out to test this prediction. They performed an experiment analogous to the one in figure n, but with x-rays rather than visible light. Their results, published in 1926, showed that if one detector saw the x-ray photon, the other did not, so that energy was always conserved at the microscopic level, not just on the average. We *never* observe an outcome in which both A and B detect a photon, or one in which neither detects it. That is, the occurrence of event A (camera A sees a photon) and event B (camera B sees one) are both random, but they are not independent.

15.4.2 Entanglement

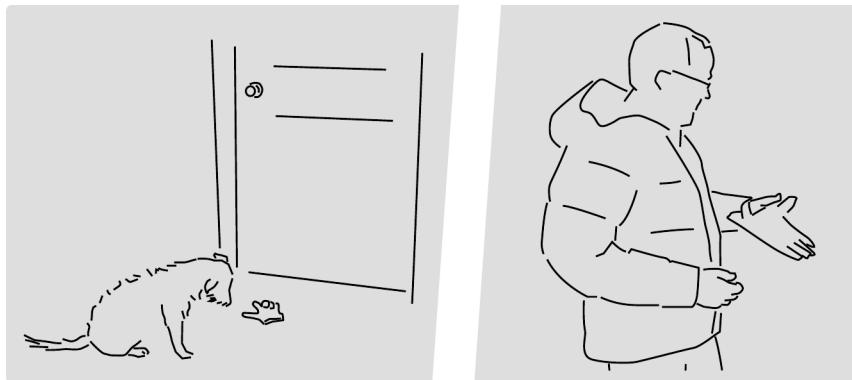
At a 1927 conference in Brussels, Einstein protested that this was a problem, because the two detectors could in principle make their observations simultaneously, and it would then seem that some influence or communication was being transmitted between them faster than the speed of light. “It seems to me,” he complained, “that this difficulty cannot be overcome unless the description of the process in terms of the . . . wave is supplemented by some detailed specification of the [trajectory of the particle]. . . . If one works only with . . . waves, the interpretation . . . , I think, contradicts the postulate of relativity.”

The experimental fact ends up being that the spooky action at a

distance exists, and it does go faster than light. In 2012, Guerreiro *et al.*¹ carried out a very direct and conceptually simple enactment of exactly the experiment in figure n, with electronic timing precise enough to prove that the detection events at A and B were separated from each other by too great a distance to have been linked by any influence traveling at $\leq c$. These findings are summarized by saying that quantum mechanics is *nonlocal*. A single wave-particle can be spread out over an arbitrarily large region of space, but its interactions that transfer energy and momentum are always correlated over these distances in such a way that the conservation laws are maintained.

What Einstein had not originally appreciated was that these correlations do not violate relativity because they do not actually transport any energy, or even any information, between A and B. For example, if Alice is at detector A, and Bob is at B, a million kilometers away, Alice can detect the photon and know immediately that Bob did not detect it. She learns something seemingly instantaneously about Bob — Bob is probably sad and disappointed right now. But because Bob does not have any control over the result, he cannot use this fact to send a message to Alice, so there is no transmission of information. Alice and Bob's states are said to be *entangled*.

o / Entanglement is like finding that you only have your left glove, so that you must have left your right glove at home. There is a gain in information, but no sudden transmission of information from the dog to you.



By analogy, suppose that you head off to work on a winter day in New York. As you step out of the subway station into the cold air, you reach into your pockets for your gloves, but you find that you only have your left glove. Oh, you must have dropped your right glove on the floor while you were petting your adorable terrier on the way out the door. The presence of your left glove tells you that your right glove must be at home. But there has been no spooky action at a distance. You have simply recovered some information about a region of space that lies at some distance from you.

Einstein and Bohr had strong physical intuitions that led them

¹arxiv.org/abs/1204.1712. The paper is very readable.

to incorrect predictions about experiments, and these predictions were the fruits of inappropriate mental pictures of what was going on. If we take the principles of quantum mechanics seriously, then the correct picture is the following. Before the photon in figure n hits the glass diagonal, the state of things is the following.

A photon is headed to the right.

Our photon is then partially reflected and partially transmitted. Now we have a superposition of two wave patterns:

$$c \begin{cases} \text{The photon has been} \\ \text{reflected upward.} \end{cases} + c' \begin{cases} \text{The photon has con-} \\ \text{tinued to the right.} \end{cases},$$

where the amplitudes c and c' are equal in absolute value.²

Let's say that the cameras are at equal distances from the glass diagonal, so that their chances to detect the photon occur simultaneously.³ After detection, we have this:

$$c \begin{cases} \text{Camera A detected a} \\ \text{photon and B didn't.} \end{cases} + c' \begin{cases} \text{B detected a photon} \\ \text{and A didn't.} \end{cases},$$

Here we have made the nontrivial assumption that material objects like cameras obey the same wave-superposition rules as photons. This turns out to be true. Cameras are made out of things like electrons, and as we'll see in chapter 16, things like electrons are also wave-particles, and they obey all the same wave-particle rules as photons. The states of the two cameras are now entangled.

You can see where this is going. Alice had been standing by camera A, watching anxiously, while Bob, a million kilometers away, was breathlessly observing camera B.

$$c \begin{cases} \text{Alice saw a photon} \\ \text{and Bob didn't. They} \\ \text{consider this result to} \\ \text{have been random.} \end{cases} + c' \begin{cases} \text{Bob saw a photon and} \\ \text{Alice didn't. They} \\ \text{consider this result to} \\ \text{have been random.} \end{cases}$$

It doesn't *seem* to Alice and Bob as though their brains are in a superposition of two states. They *feel* as though they have only

²Conservation of energy requires $c^2 = 1/2$ and $c'^2 = 1/2$, even in classical physics. We could have, for example, $c = 1/\sqrt{2}$ and $c' = -1/\sqrt{2}$. Such a possible difference in signs wouldn't concern us in this example. It would only be relevant if there were some later opportunity for the two parts of the wave to recombine and superimpose on one another, producing interference effects.

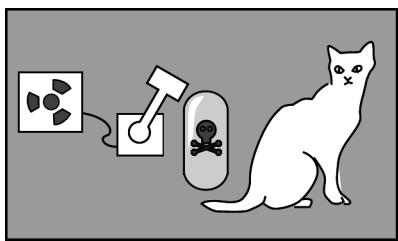
³According to special relativity, this simultaneity holds only in one frame of reference, say the lab frame. But if simultaneity does hold in one frame, then we can also say that in *all* frames, the distance between the two events is "spacelike," i.e., they are too far apart to have been connected by any causal influence propagating at $\leq c$.

experienced the one possibility that actually happened, not a mixture of both at the same time. And yet this picture of the physics explains very nicely how the deterministic laws of physics produce a result that *seems* to them to have been random.

If Alice and Bob have been split into two ghostlike halves of themselves, then conceivably these half-selves could undergo interference, as in the double-slit experiment. But there are practical reasons why we cannot actually detect such interference effects. For one thing, Alice and Bob are macroscopic objects, with energies E on the order of many joules. Because Planck's constant is small, their wave frequencies $f = E/h$ are extremely high, and their wavelengths incredibly short (on the order of 10^{-34} m!). We have seen that diffraction becomes undetectable when wavelengths are too short. Furthermore, there is a phenomenon called decoherence, discussed further in sec. 16.5.1, p. 381, in which interactions with the environment tend to rapidly randomize the wave-phases of large objects. When phases are randomized, interference and diffraction effects become undetectable.

Historically, it seemed absurd to the originators of quantum mechanics to imagine a macroscopic object in a superposition of states. The most celebrated example is called the Schrödinger's cat experiment. Luckily for the cat, there probably was no actual experiment — it was simply a “thought experiment” that the German theorist Schrödinger discussed with his colleagues. Schrödinger wrote:

One can even construct quite burlesque cases. A cat is shut up in a steel container, together with the following diabolical apparatus (which one must keep out of the direct clutches of the cat): In a Geiger tube [radiation detector] there is a tiny mass of radioactive substance, so little that in the course of an hour perhaps one atom of it disintegrates, but also with equal probability not even one; if it does happen, the counter [detector] responds and ... activates a hammer that shatters a little flask of prussic acid [filling the chamber with poison gas]. If one has left this entire system to itself for an hour, then one will say to himself that the cat is still living, if in that time no atom has disintegrated. The first atomic disintegration would have poisoned it.



p / Schrödinger's cat.

It seemed ridiculous to Schrödinger that at the end of the hour, “The uncertainty originally restricted to the atomic domain has been transformed into a macroscopic uncertainty...,” and the cat would be in a superposed state.

In modern language, people like Einstein and Schrödinger didn't feel comfortable with nonlocality, or with entanglement of subatomic particles, and they felt even less comfortable with applying these

concepts to macroscopic objects. Today, entanglement has been demonstrated using objects that clearly deserve to be called macroscopic. For example, in 2012, K.C. Lee *et al.* created a version of the experiment in figure n in which the cameras were replaced by small diamonds, about 1 mm in size. They were separated by 15 cm, which is a macroscopic distance. When a photon hit one of the diamonds, it produced a vibration in the crystal lattice. This vibration was localized to a relatively small region within the diamond, but this region was still large enough that one has to admit that it qualifies as macroscopic. Its atoms had a total weight of about 0.1 nanograms, which is a quantity big enough to weigh on a state-of-the-art balance, and the region was about 0.01 mm in size, which would make it visible with a magnifying glass.

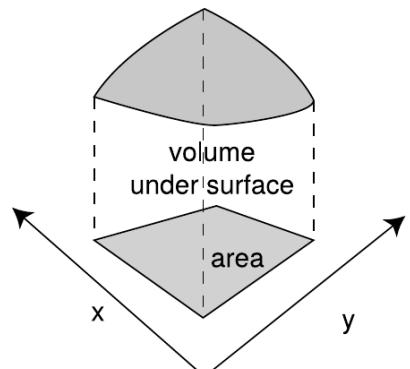
The quantum states of the two diamonds became entangled: if one had detected the photon, the other hadn't. This entangled state was maintained for only about 7 picoseconds before decoherence destroyed the phase relationship between one diamond and the other. But Lee was able to use additional photons to "read out" the quantum states in only 0.5 ps, before decoherence occurred, and verify that there were wave interference effects in which one diamond's quantum-mechanical wave had a definite phase relationship with the other's. Although these experiments are difficult, they suggest that there is no obstruction in principle to observing quantum-mechanical effects such as superposition in arbitrarily large objects.

15.5 Photons in three dimensions

Up until now I've been sneaky and avoided a full discussion of the three-dimensional aspects of the probability interpretation. The example of the carrot in the microwave oven, for example, reduced to a one-dimensional situation because we were considering three points along the same line and because we were only comparing ratios of probabilities.

A typical example of a probability distribution in section 14.1 was the distribution of heights of human beings. The thing that varied randomly, height, h , had units of meters, and the probability distribution was a graph of a function $D(h)$. The units of the probability distribution had to be m^{-1} (inverse meters) so that areas under the curve, interpreted as probabilities, would be unitless: $(\text{area}) = (\text{height})(\text{width}) = \text{m}^{-1} \cdot \text{m}$.

Now suppose we have a two-dimensional problem, e.g., the probability distribution for the place on the surface of a digital camera chip where a photon will be detected. The point where it is detected would be described with two variables, x and y , each having units of meters. The probability distribution will be a function of both variables, $D(x, y)$. A probability is now visualized as the volume



q / Probability is the volume under a surface defined by $D(x, y)$.

under the surface described by the function $D(x, y)$, as shown in figure q. The units of D must be m^{-2} so that probabilities will be unitless: (probability) = (depth)(length)(width) = $\text{m}^{-2} \cdot \text{m} \cdot \text{m}$. In terms of calculus, we have $P = \int D \, dx \, dy$.

Generalizing finally to three dimensions, we find by analogy that the probability distribution will be a function of all three coordinates, $D(x, y, z)$, and will have units of m^{-3} . It is unfortunately impossible to visualize the graph unless you are a mutant with a natural feel for life in four dimensions. If the probability distribution is nearly constant within a certain volume of space v , the probability that the photon is in that volume is simply vD . If not, then we can use an integral, $P = \int D \, dx \, dy \, dz$.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 When light is reflected from a mirror, perhaps only 80% of the energy comes back. One could try to explain this in two different ways: (1) 80% of the photons are reflected, or (2) all the photons are reflected, but each loses 20% of its energy. Based on your everyday knowledge about mirrors, how can you tell which interpretation is correct? [Based on a problem from PSSC Physics.]

2 Suppose we want to build an electronic light sensor using an apparatus like the one described in section 15.2 on p. 346. How would its ability to detect different parts of the spectrum depend on the type of metal used in the capacitor plates?

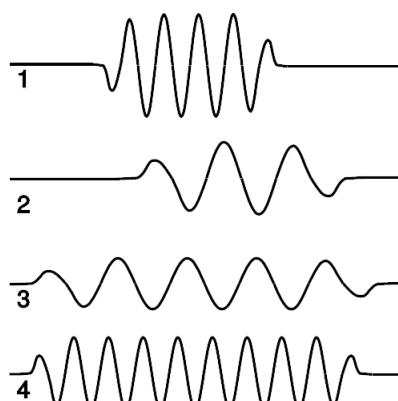
3 The photoelectric effect can occur not just for metal cathodes but for any substance, including living tissue. Ionization of DNA molecules can cause cancer or birth defects. If the energy required to ionize DNA is on the same order of magnitude as the energy required to produce the photoelectric effect in a metal, which of the following types of electromagnetic waves might pose such a hazard? Explain.

- 60 Hz waves from power lines
- 100 MHz FM radio
- microwaves from a microwave oven
- visible light
- ultraviolet light
- x-rays

4 (a) Rank-order the photons according to their wavelengths, frequencies, and energies. If two are equal, say so. Explain all your answers.

(b) Photon 3 was emitted by a xenon atom going from its second-lowest-energy state to its lowest-energy state. Which of photons 1, 2, and 4 are capable of exciting a xenon atom from its lowest-energy state to its second-lowest-energy state? Explain.

5 The beam of a 100 W overhead projector covers an area of $1 \text{ m} \times 1 \text{ m}$ when it hits the screen 3 m away. Estimate the number of photons that are in flight at any given time. (Since this is only an estimate, we can ignore the fact that the beam is not parallel.) ✓



Problem 4.

6 In the photoelectric effect, electrons are observed with virtually no time delay (~ 10 ns), even when the light source is very weak. (A weak light source does however only produce a small number of ejected electrons.) The purpose of this problem is to show that the lack of a significant time delay contradicted the classical wave theory of light, so throughout this problem you should put yourself in the shoes of a classical physicist and pretend you don't know about photons at all. At that time, it was thought that the electron might have a radius on the order of 10^{-15} m. (Recent experiments have shown that if the electron has any finite size at all, it is far smaller.)

(a) Estimate the power that would be soaked up by a single electron in a beam of light with an intensity of 1 mW/m^2 . \checkmark

(b) The energy, E_s , required for the electron to escape through the surface of the cathode is on the order of 10^{-19} J. Find how long it would take the electron to absorb this amount of energy, and explain why your result constitutes strong evidence that there is something wrong with the classical theory. \checkmark

7 As far as we know, the mass of the photon is zero. However, it's not possible to prove by experiments that anything is zero; all we can do is put an upper limit on the number. As of 2008, the best experimental upper limit on the mass of the photon is about 1×10^{-52} kg. Suppose that the photon's mass really isn't zero, and that the value is at the top of the range that is consistent with the present experimental evidence. In this case, the c occurring in relativity would no longer be interpreted as the speed of light. As with material particles, the speed v of a photon would depend on its energy, and could never be as great as c . Estimate the relative size $(c - v)/c$ of the discrepancy in speed, in the case of a photon of visible light.

▷ Answer, p. 459

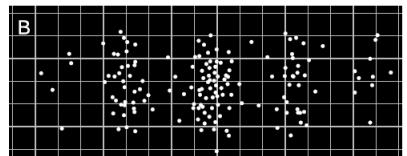
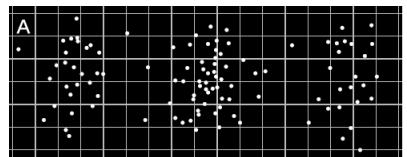
8 Give a numerical comparison of the number of photons per second emitted by a hundred-watt FM radio transmitter and a hundred-watt lightbulb. \checkmark

9 (a) A radio transmitter radiates power P in all directions, so that the energy spreads out spherically. Find the energy density at a distance r . \checkmark

(b) Let the wavelength be λ . As described in example 2 on p. 349, find the number of photons in a volume λ^3 at this distance r . \checkmark

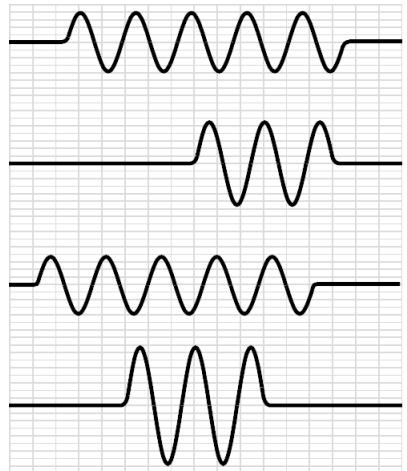
(c) For a 1000 kHz AM radio transmitting station, assuming reasonable values of P and r , verify, as claimed in the example, that the result from part b is very large.

- 10** The two diffraction patterns were made by sending a flash of light through the same double slit. Give a numerical comparison of the amounts of energy in the two flashes.



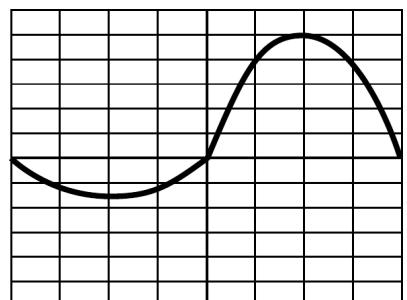
Problem 10.

- 11** Three of the four graphs are consistently normalized to represent a wave consisting of the same number of photons. Which one isn't? Explain.

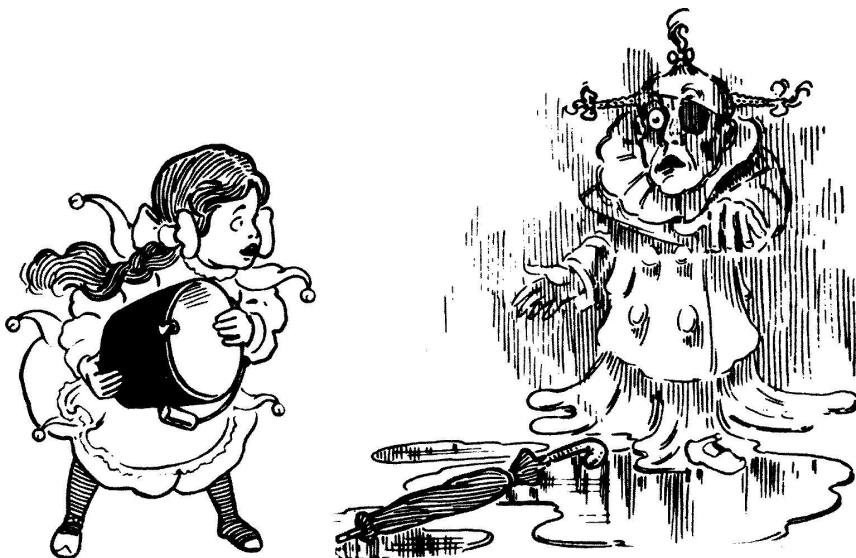


Problem 11.

- 12** Many radio antennas are designed so that they preferentially emit or receive electromagnetic waves in a certain direction. However, no antenna is perfectly directional. The wave shown in the figure represents a single photon being emitted by an antenna at the center. The antenna is directional, so there is a stronger wave on the right than on the left. What is the probability that the photon will be observed on the right?



Problem 12.



Dorothy melts the Wicked Witch of the West.

Chapter 16

Matter as a wave

[In] a few minutes I shall be all melted... I have been wicked in my day, but I never thought a little girl like you would ever be able to melt me and end my wicked deeds. Look out — here I go!

The Wicked Witch of the West

As the Wicked Witch learned the hard way, losing molecular cohesion can be unpleasant. That's why we should be very grateful that the concepts of quantum physics apply to matter as well as light. If matter obeyed the laws of classical physics, molecules wouldn't exist.

Consider, for example, the simplest atom, hydrogen. Why does one hydrogen atom form a chemical bond with another hydrogen atom? Roughly speaking, we'd expect a neighboring pair of hydrogen atoms, A and B, to exert no force on each other at all, attractive or repulsive: there are two repulsive interactions (proton A with proton B and electron A with electron B) and two attractive interactions (proton A with electron B and electron A with proton B). Thinking a little more precisely, we should even expect that once the two atoms got close enough, the interaction would be repulsive. For instance, if you squeezed them so close together that the two protons were almost on top of each other, there would be a tremendously strong repulsion between them due to the $1/r^2$ nature of the

electrical force. A more detailed calculation using classical physics gives an extremely weak binding, about 1/17 the strength of what we actually observe ([2388](#)), which is far too weak to make the bond hold together.

Quantum physics to the rescue! As we'll see shortly, the whole problem is solved by applying the same quantum concepts to electrons that we have already used for photons.

16.1 Electrons as waves

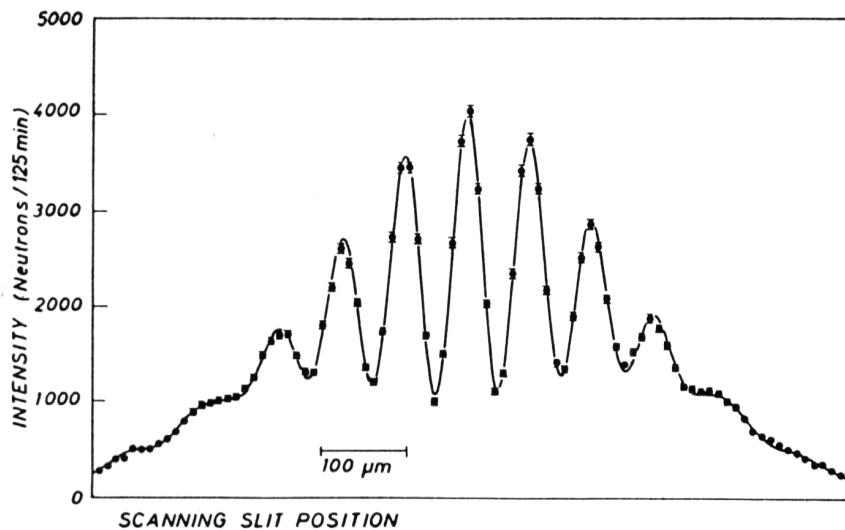
16.1.1 Wavelength related to momentum

We started our journey into quantum physics by studying the random behavior of *matter* in radioactive decay, and then asked how randomness could be linked to the basic laws of nature governing *light*. The probability interpretation of wave-particle duality was strange and hard to accept, but it provided such a link. It is now natural to ask whether the same explanation could be applied to matter. If the fundamental building block of light, the photon, is a particle as well as a wave, is it possible that the basic units of matter, such as electrons, are waves as well as particles?

A young French aristocrat studying physics, Louis de Broglie (pronounced “broylee”), made exactly this suggestion in his 1923 Ph.D. thesis. His idea had seemed so farfetched that there was serious doubt about whether to grant him the degree. Einstein was asked for his opinion, and with his strong support, de Broglie got his degree.

Only two years later, American physicists C.J. Davisson and L. Germer confirmed de Broglie's idea by accident. They had been studying the scattering of electrons from the surface of a sample of nickel, made of many small crystals. (One can often see such a crystalline pattern on a brass doorknob that has been polished by repeated handling.) An accidental explosion occurred, and when they put their apparatus back together they observed something entirely different: the scattered electrons were now creating an interference pattern! This dramatic proof of the wave nature of matter came about because the nickel sample had been melted by the explosion and then resolidified as a single crystal. The nickel atoms, now nicely arranged in the regular rows and columns of a crystalline lattice, were acting as the lines of a diffraction grating. The new crystal was analogous to the type of ordinary diffraction grating in which the lines are etched on the surface of a mirror (a reflection grating) rather than the kind in which the light passes through the transparent gaps between the lines (a transmission grating).

Although we will concentrate on the wave-particle duality of elec-



a / A double-slit interference pattern made with neutrons. (A. Zeilinger, R. Gähler, C.G. Shull, W. Treimer, and W. Mampe, *Reviews of Modern Physics*, Vol. 60, 1988.)

trons because it is important in chemistry and the physics of atoms, all the other “particles” of matter you’ve learned about show wave properties as well. Figure a, for instance, shows a wave interference pattern of neutrons.

It might seem as though all our work was already done for us, and there would be nothing new to understand about electrons: they have the same kind of funny wave-particle duality as photons. That’s almost true, but not quite. There are some important ways in which electrons differ significantly from photons:

1. Electrons have mass, and photons don’t.
2. Photons always move at the speed of light, but electrons can move at any speed less than c .
3. Photons don’t have electric charge, but electrons do, so electric forces can act on them. The most important example is the atom, in which the electrons are held by the electric force of the nucleus.
4. Electrons cannot be absorbed or emitted as photons are. Destroying an electron or creating one out of nothing would violate conservation of charge.

(In section 18.6 we will learn of one more fundamental way in which electrons differ from photons, for a total of five.)

Because electrons are different from photons, it is not immediately obvious which of the photon equations from chapter 15 can be

- The wavenumber $k = 2\pi/\lambda$ is inversely related to the wavelength λ .
- It has units of radians per meter.
- In three dimensions, k is the magnitude of a the wave-vector \mathbf{k} (p. 51).
- The wavevector k is to the wavelength as the frequency ω is to the period.

b / Review of the wavenumber k

applied to electrons as well. With hindsight, we know that there is a simple and consistent way of putting this all together. As a preliminary in order to make the notation simpler, we define a version of Planck's constant $\hbar = h/2\pi$, read as "h-bar." We also recall from sec. 2.4.6, p. 45, the definition of the wavenumber $k = 2\pi/\lambda$ along with a few of its properties, table b. With these definitions, our two fundamental quantum-mechanical relations are

$$E = \hbar\omega \quad [\text{the same as } E = hf]$$

$$p = \hbar k \quad [\text{the same as } p = h/\lambda]$$

It's only a slight exaggeration to say that these two equations summarize all of quantum mechanics. They are true for all the ordinary building blocks of light and matter in everyday life, and also for every weird creature in the particle-physics zoo. Each of these equations has a *particle* thing on the left and a *wave* thing on the right.

The second equation, although true for photons, takes on a greater importance for electrons. This is first of all because the momentum of matter is more likely to be significant than the momentum of light under ordinary conditions, and also because force is the transfer of momentum, and electrons are affected by electrical forces.

The wavelength of a cat

example 1

▷ What is the wavelength of a trotting cat?

▷ One may doubt whether the equation $p = \hbar k = h/\lambda$ should be applied to a cat, which is not just a single particle but a rather large collection of them. Throwing caution to the wind, however, we estimate the cat's mass at 1 kg and its trotting speed at 1 m/s. Its wavelength is therefore roughly

$$\begin{aligned} \lambda &= \frac{h}{p} \\ &= \frac{h}{mv} \\ &= \frac{6.63 \times 10^{-34} \text{ J}\cdot\text{s}}{(1 \text{ kg})(1 \text{ m/s})} \\ &\sim 10^{-33} \frac{(\text{kg}\cdot\text{m}^2/\text{s}^2)\cdot\text{s}}{\text{kg}\cdot\text{m/s}} \\ &= 10^{-33} \text{ m.} \end{aligned}$$

The wavelength found in this example is so fantastically small that we can be sure we will never observe any measurable wave phenomena with cats or any other human-scale objects. For example, if we sent the cat through a pair of double slits separated by 10 cm, then the angular spacing of the diffraction pattern would be 10^{-32} radians, which would make the fringes too close together

to be distinguished. This is in agreement with the correspondence principle.

Although a smaller mass in the equation $\lambda = h/mv$ does result in a longer wavelength, the wavelength is still quite short even for individual electrons under typical conditions, as shown in the following example.

The typical wavelength of an electron *example 2*

- ▷ Electrons in circuits and in atoms are typically moving through voltage differences on the order of 1 V, so that a typical energy is $(e)(1 \text{ V})$, which is on the order of 10^{-19} J . What is the wavelength of an electron with this amount of kinetic energy?
- ▷ This energy is nonrelativistic, since it is much less than mc^2 . Momentum and energy are therefore related by the nonrelativistic equation $K = p^2/2m$. Solving for p and substituting in to the equation for the wavelength, we find

$$\begin{aligned}\lambda &= \frac{h}{\sqrt{2mK}} \\ &= 1.6 \times 10^{-9} \text{ m.}\end{aligned}$$

This is on the same order of magnitude as the size of an atom, which is no accident: as we will discuss in the next chapter in more detail, an electron in an atom can be interpreted as a standing wave. The smallness of the wavelength of a typical electron also helps to explain why the wave nature of electrons wasn't discovered until a hundred years after the wave nature of light. To scale the usual wave-optics devices such as diffraction gratings down to the size needed to work with electrons at ordinary energies, we need to make them so small that their parts are comparable in size to individual atoms. This is essentially what Davisson and Germer did with their nickel crystal.

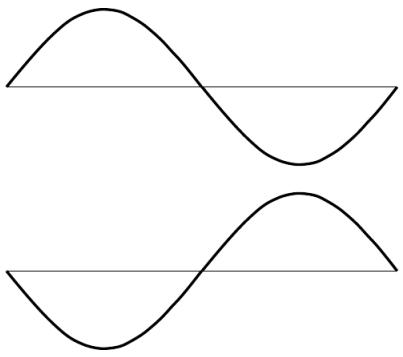
self-check A

These remarks about the inconvenient smallness of electron wavelengths apply only under the assumption that the electrons have typical energies. What kind of energy would an electron have to have in order to have a longer wavelength that might be more convenient to work with?

- ▷ Answer, p. 458

16.1.2 What kind of wave is it?

If a sound wave is a vibration of matter, and a photon is a vibration of electric and magnetic fields, what kind of a wave is an electron made of? The disconcerting answer is that there is no experimental “observable,” i.e., directly measurable quantity, to correspond to the electron wave itself. In other words, there are devices like microphones that detect the oscillations of air pressure in a sound wave, and devices such as radio receivers that measure



c / These two electron waves are not distinguishable by any measuring device.

the oscillation of the electric and magnetic fields in a light wave, but nobody has ever found any way to measure the electron wave directly.

We can of course detect the energy (or momentum) possessed by an electron just as we could detect the energy of a photon using a digital camera. (In fact I'd imagine that an unmodified digital camera chip placed in a vacuum chamber would detect electrons just as handily as photons.) But this only allows us to determine where the wave carries high probability and where it carries low probability. Probability is proportional to the square of the wave's amplitude, but measuring its square is not the same as measuring the wave itself. In particular, we get the same result by squaring either a positive number or its negative, so there is no way to determine the positive or negative sign of an electron wave. In general, the phase of the wavefunction is not an observable thing.

This meaninglessness of phase only applies to *absolute* phases. *Relative* phases do have real-world consequences. For example, figure a on p. 367 shows a double-slit interference pattern made by neutrons. There are places where we observe a lot of neutrons, and places where we don't. The places where we don't see many neutrons are the ones where waves are (at least partially) canceling, which means that they're out of phase — *relative* to each other. As a political metaphor, my wife and I would sometimes, before 2016, cancel each other out at the polls by voting for different political parties. It was a matter of opinion who was right, but it was an objective fact that we canceled.

Most physicists tend toward the school of philosophy known as operationalism, which says that a concept is only meaningful if we can define some set of operations for observing, measuring, or testing it. According to a strict operationalist, then, the electron wave itself is a meaningless concept, because we can't measure its absolute phase. Nevertheless, it turns out to be one of those concepts like love or humor that is impossible to measure and yet very useful to have around. We therefore give it a symbol, Ψ (the capital Greek letter psi), and a special name, the electron *wavefunction* (because it is a function of the coordinates x , y , and z that specify where you are in space). It would be impossible, for example, to calculate the shape of the electron wave in a hydrogen atom without having some symbol for the wave. But when the calculation produces a result that can be compared directly to experiment, the final algebraic result will turn out to involve only Ψ^2 , which is what is observable, not Ψ itself.

Since Ψ , unlike E and B , is not directly measurable, we are free to make the probability equations have a simple form: instead of having the probability density equal to some funny constant multiplied by Ψ^2 , we simply define Ψ so that the constant of proportion-

ality is one:

$$(\text{probability distribution}) = |\Psi|^2.$$

Since the probability distribution has units of m^{-3} , the units of Ψ must be $\text{m}^{-3/2}$. The square of a negative number is still positive, so the absolute value signs may seem unnecessary, but as we'll in sec. 17.5, p. 403, the wavefunction may in general be a complex number. In fact, only standing waves, not traveling waves, can really be represented by real numbers, although we will often cheat and draw pictures of traveling waves as if they were real-valued functions.

We have defined a wave as something that can superpose, and a particle as something that you can't have a fraction of (p. 351). Since you can't have a fraction of an electron, we conclude that if Ψ is a valid wavefunction for an electron, then $\Psi/2$ should be somehow illegal. We can see this from the requirement that the probability distribution be normalized (p. 326).

Normalization and phase of a particle in a box *example 3*

A certain electron confined to a box of length L has a wavefunction

$$\Psi(x) = \begin{cases} A\sin(2\pi x/L) & \text{if } 0 \leq x \leq L \\ 0 & \text{elsewhere.} \end{cases}$$

This quantity inside the sine function varies from 0 to 2π , so this is a standing wave with one wavelength fitting inside the box, \checkmark . If we pick positive or negative values of A we can have the two functions in figure c, p. 370, but this is a difference in phase, which we don't expect to be physically observable. Normalization requires that

$$\int_0^L |\Psi|^2 dx = 1$$

or

$$|A|^2 \int_0^L \sin^2(2\pi x/L) dx = 1.$$

This integral can be done using straightforward calculus, but a shortcut is to recognize that the sine oscillates symmetrically between 0 and 1, and therefore over any whole number of half-cycles its average value is 1/2. We can therefore find the same result for the definite integral by replacing the sine with 1/2, and this results in

$$|A|^2 = \frac{2}{L}.$$

Two possible values of A are $\sqrt{2/L}$ and $-\sqrt{2/L}$, but these represent physically indistinguishable states. We could also have a complex value of A , as long as $|A|^2$ had the correct value.

16.1.3 Quantum numbers and bra-ket notation

When we want to describe a wavefunction, it can be cumbersome to write out its equation as in example 3, the particle in a box. In that particular example, it's much easier just to draw a picture \curvearrowleft or to define some simple numerical label like $N = 2$ as we did in problem 12 on p. 59. A number like this is called a *quantum number*, and in this particular example the N quantum number is a wavelength or energy label. (It isn't a momentum label, because the standing wave is a superposition of traveling waves going in both directions.)

When we want to refer specifically to the wavefunction that is labeled in these ways, there is a handy notation involving angle brackets, called the “bra-ket” notation. In our example, we could write something like

$$|\curvearrowleft\rangle$$

or

$$|N = 2\rangle.$$

These are called kets. We can also have notation where the bar and angle bracket are flipped the other way, like $\langle \curvearrowleft |$ or $\langle N = 2 |$. These are called bras. For real-valued wavefunctions, which are good enough for standing waves, there is no important distinction to be made between bras and kets. When we sandwich together a bra and a ket, both of which represent the same state, this is a shorthand way of notating the type of integral that we did in example 3 for normalization. That is, if we want to express the requirement that the state be normalized, we can write something like

$$\langle \curvearrowleft | \curvearrowleft \rangle = 1,$$

which is a lot easier than writing out the whole integral as $\int_0^L |A|^2 \sin^2(2\pi x/L) dx = 1$. The bra-ket notation is generalized in sec. 17.7.1, p. 409.

16.1.4 “Same state” versus “same wavefunction”

We've already seen that phases (or at least absolute phases) are not physically meaningful. We saw this principle at work in example 3, where $|\curvearrowleft\rangle$ and $|\curvearrowleft\rangle$ would give identical predictions about probabilities. In general two different wavefunctions can actually represent the same state, if they only differ by a constant factor. It doesn't matter whether we multiply by -1 , or for that matter by $1/2$, forming $|\curvearrowleft\rangle$. The latter would not be normalized, but it wouldn't be a different state, just an unnormalized version of the same state. We could of course rule out unnormalized wavefunctions, but that's not a good idea, for the following reasons.

Suppose you do the Schrödinger's cat experiment (p. 358). At the end of the experiment, your brain is in a superposition of two states that really are different: a state in which your brain saw a

live cat, and a state in which it saw a dead one. Let's label these states as $|\text{brain L}\rangle$ and $|\text{brain D}\rangle$. The state of your brain is

$$c|\text{brain L}\rangle + c'|\text{brain D}\rangle.$$

Now as far as $|\text{brain L}\rangle$ is concerned, a live cat is the only possibility that has happened. That brain doesn't know about the alternate universe in which $|\text{brain D}\rangle$ saw a dead cat.

In a certain calculation, we might say that c was a negative number, but the brain doesn't say, "Huh, I feel negative today," or "The universe seems to have flipped its phase now."

Similarly, the brain has no way of knowing whether c is big or small. Suppose that the Quantum Police did decree that all wavefunctions had to be normalized. Then we would have $|c| < 1$. (If the L and D outcomes each have probability 1/2, then $|c| = |c'| = 1/\sqrt{2}$.) But in $|\text{brain L}\rangle$'s world, it's meaningless to talk about this. That brain doesn't say, "I feel weak, as though my normalization was sagging," or "The universe seems gray and boring today — I think its normalization has gotten lower."

For these reasons, we adopt the convention that if $|\dots\rangle$ is a wavefunction and c is any nonzero number (either real or complex), then any wavefunction $c|\dots\rangle$ represents the *same state*.

As a matter of mathematical elegance, waves are things that superpose, and this means that they inhabit a vector space (p. 155). One of the axioms defining a vector space is that we have well-defined operations of addition and multiplication by a scalar. If we ruled out unnormalized wavefunctions, we would be violating this requirement. And physically, it is sometimes OK to talk about an unnormalized state as representing more than one particle, or, in the case of a traveling wave, a steady flux of particles like the beam of a cathode ray tube. We don't actually want to stop talking completely about unnormalized wavefunctions, like someone from the 1950's who's afraid to use the word "toilet" in polite company.

None of this is meant to imply that the c 's are never important. Relative phases *are* meaningful. For example, $|\text{VV}\rangle + |\text{V}\text{D}\rangle$ will produce a different interference pattern than $|\text{VV}\rangle - |\text{V}\text{D}\rangle$. And we do need to keep track of normalization in order to say anything about probabilities. For example, it's not true that in the state described by the unnormalized wavefunction $5|\curvearrowleft\rangle + 7|\curvearrowright\rangle$, the probability of \curvearrowleft is 25; probabilities have to be between 0 and 1.

16.2 Dispersive waves

A colleague of mine who teaches chemistry loves to tell the story about an exceptionally bright student who, when told of the equation $p = h/\lambda$, protested, "But when I derived it, it had a factor of 2!" The issue that's involved is a real one, albeit one that could be

glossed over (and is, in most textbooks) without raising any alarms in the mind of the average student. The present optional section addresses this point; it is intended for the student who wishes to delve a little deeper.

Here's how the now-legendary student was presumably reasoning. We start with the equation $v = f\lambda$, which is valid for any sine wave, whether it's quantum or classical. Let's assume we already know $E = hf$, and are trying to derive the relationship between wavelength and momentum:

$$\begin{aligned}\lambda &= \frac{v}{f} \\ &= \frac{vh}{E} \\ &= \frac{vh}{\frac{1}{2}mv^2} \\ &= \frac{2h}{mv} \\ &= \frac{2h}{p}.\end{aligned}$$

The reasoning seems valid, but the result does contradict the accepted one, which is after all solidly based on experiment.

The mistaken assumption is that we can figure everything out in terms of pure sine waves. Mathematically, the only wave that has a perfectly well defined wavelength and frequency is a sine wave, and not just any sine wave but an infinitely long sine wave, d. The unphysical thing about such a wave is that it has no leading or trailing edge, so it can never be said to enter or leave any particular region of space. Our derivation made use of the velocity, v , and if velocity is to be a meaningful concept, it must tell us how quickly stuff (mass, energy, momentum, ...) is transported from one region of space to another. Since an infinitely long sine wave doesn't remove any stuff from one region and take it to another, the "velocity of its stuff" is not a well defined concept.

Of course the individual wave peaks do travel through space, and one might think that it would make sense to associate their speed with the "speed of stuff," but as we will see, the two velocities are in general unequal when a wave's velocity depends on wavelength. Such a wave is called a *dispersive* wave, because a wave pulse consisting of a superposition of waves of different wavelengths will separate (disperse) into its separate wavelengths as the waves move through space at different speeds. Nearly all the waves we have encountered have been nondispersive. For instance, sound waves and light waves (in a vacuum) have speeds independent of wavelength. A water wave is one good example of a dispersive wave. Long-wavelength water waves travel faster, so a ship at sea that encounters a storm typically sees the long-wavelength parts of the wave first. When dealing

with dispersive waves, we need symbols and words to distinguish the two speeds. The speed at which wave peaks move is called the phase velocity, v_p , and the speed at which “stuff” moves is called the group velocity, v_g .

An infinite sine wave can only tell us about the phase velocity, not the group velocity, which is really what we would be talking about when we refer to the speed of an electron. If an infinite sine wave is the simplest possible wave, what’s the next best thing? We might think the runner up in simplicity would be a wave train consisting of a chopped-off segment of a sine wave, e. However, this kind of wave has kinks in it at the end. A simple wave should be one that we can build by superposing a small number of infinite sine waves, but a kink can never be produced by superposing any number of infinitely long sine waves.

Actually the simplest wave that transports stuff from place to place is the pattern shown in figure f. Called a beat pattern, it is formed by superposing two sine waves whose wavelengths are similar but not quite the same. If you have ever heard the pulsating howling sound of musicians in the process of tuning their instruments to each other, you have heard a beat pattern. The beat pattern gets stronger and weaker as the two sine waves go in and out of phase with each other. The beat pattern has more “stuff” (energy, for example) in the areas where constructive interference occurs, and less in the regions of cancellation. As the whole pattern moves through space, stuff is transported from some regions and into other ones.

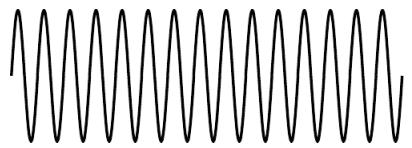
If the frequency of the two sine waves differs by 10%, for instance, then ten periods will occur between times when they are in phase. Another way of saying it is that the sinusoidal “envelope” (the dashed lines in figure f) has a frequency equal to the difference in frequency between the two waves. For instance, if the waves had frequencies of 100 Hz and 110 Hz, the frequency of the envelope would be 10 Hz.

Applying similar reasoning to the wavelength, the wavenumber k of the envelope equals the difference between the wavenumbers of the two sine waves.

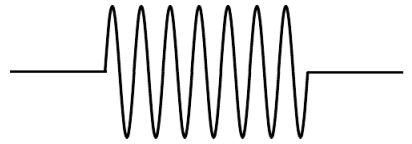
The group velocity is the speed at which the envelope moves through space. Let $\Delta\omega$ and Δk be the differences between the frequencies and wavenumbers of the two sine waves, which means that they equal the frequency and wavenumber of the envelope. The group velocity is $v_g = \omega_{\text{envelope}}/k_{\text{envelope}} = \Delta\omega/\Delta k$. If $\Delta\omega$ and Δk are small, we can approximate this expression as a derivative,

$$v_g = \frac{d\omega}{dk}.$$

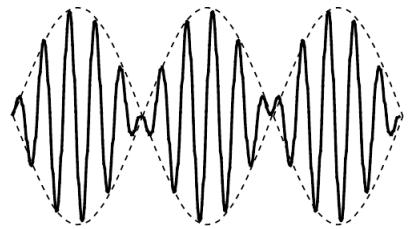
This expression is usually taken as the definition of the group velocity for wave patterns that consist of a superposition of sine waves



d / Part of an infinite sine wave.



e / A finite-length sine wave.



f / A beat pattern created by superimposing two sine waves with slightly different wavelengths.

having a narrow range of frequencies and wavelengths. In quantum mechanics, with $\omega = E/\hbar$ and $k = p/\hbar$, we have $v_g = dE/dp$. In the case of a nonrelativistic electron the relationship between energy and momentum is $E = p^2/2m$, so the group velocity is $dE/dp = p/m = v$, exactly what it should be. It is only the phase velocity that differs by a factor of two from what we would have expected, but the phase velocity is not the physically important thing.

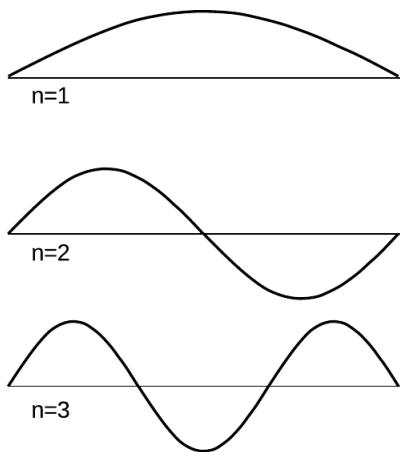
16.3 Bound states

Electrons are at their most interesting when they're in atoms, that is, when they are bound within a small region of space. We can understand a great deal about atoms and molecules based on simple arguments about such bound states, without going into any of the realistic details of atom. The simplest model of a bound state is known as the particle in a box: like a ball on a pool table, the electron feels zero force while in the interior, but when it reaches an edge it encounters a wall that pushes back inward on it with a large force. In particle language, we would describe the electron as bouncing off of the wall, but this incorrectly assumes that the electron has a certain path through space. It is more correct to describe the electron as a wave that undergoes 100% reflection at the boundaries of the box.

Like generations of physics students before me, I rolled my eyes when initially introduced to the unrealistic idea of putting a particle in a box. It seemed completely impractical, an artificial textbook invention. Today, however, it has become routine to study electrons in rectangular boxes in actual laboratory experiments. The “box” is actually just an empty cavity within a solid piece of silicon, amounting in volume to a few hundred atoms. The methods for creating these electron-in-a-box setups (known as “quantum dots”) were a by-product of the development of technologies for fabricating computer chips.

For simplicity let's imagine a one-dimensional electron in a box, i.e., we assume that the electron is only free to move along a line. The resulting standing wave patterns, of which the first three are shown in the figure, are just like some of the patterns we encountered with sound waves in musical instruments. The wave patterns must be zero at the ends of the box, because we are assuming the walls are impenetrable, and there should therefore be zero probability of finding the electron outside the box. Each wave pattern is labeled according to n , the number of peaks and valleys it has. In quantum physics, these wave patterns are referred to as “states” of the particle-in-the-box system.

The following seemingly innocuous observations about the particle in the box lead us directly to the solutions to some of the most



g / Three possible standing-wave patterns for a particle in a box.

vexing failures of classical physics:

The particle's energy is quantized (can only have certain values). Each wavelength corresponds to a certain momentum, and a given momentum implies a definite kinetic energy, $E = p^2/2m$. (This is the second type of energy quantization we have encountered. The type we studied previously had to do with restricting the number of particles to a whole number, while assuming some specific wavelength and energy for each particle. This type of quantization refers to the energies that a single particle can have. Both photons and matter particles demonstrate both types of quantization under the appropriate circumstances.)

The particle has a minimum kinetic energy. Long wavelengths correspond to low momenta and low energies. There can be no state with an energy lower than that of the $n = 1$ state, called the ground state.

The smaller the space in which the particle is confined, the higher its kinetic energy must be. Again, this is because long wavelengths give lower energies.

Spectra of thin gases

example 4

A fact that was inexplicable by classical physics was that thin gases absorb and emit light only at certain wavelengths. This was observed both in earthbound laboratories and in the spectra of stars. The figure on the left shows the example of the spectrum of the star Sirius, in which there are "gap teeth" at certain wavelengths. Taking this spectrum as an example, we can give a straightforward explanation using quantum physics.

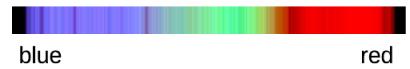
Energy is released in the dense interior of the star, but the outer layers of the star are thin, so the atoms are far apart and electrons are confined within individual atoms. Although their standing-wave patterns are not as simple as those of the particle in the box, their energies are quantized.

When a photon is on its way out through the outer layers, it can be absorbed by an electron in an atom, but only if the amount of energy it carries happens to be the right amount to kick the electron from one of the allowed energy levels to one of the higher levels. The photon energies that are missing from the spectrum are the ones that equal the difference in energy between two electron energy levels. (The most prominent of the absorption lines in Sirius's spectrum are absorption lines of the hydrogen atom.)

The stability of atoms

example 5

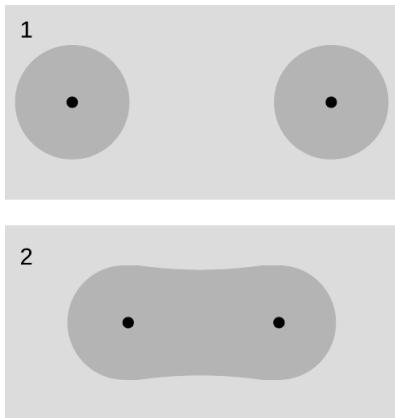
In many Star Trek episodes the Enterprise, in orbit around a planet, suddenly lost engine power and began spiraling down toward the planet's surface. This was utter nonsense, of course, due to conservation of energy: the ship had no way of getting rid of energy, so it did not need the engines to replenish it.



h / The spectrum of the light from the star Sirius.

Consider, however, the electron in an atom as it orbits the nucleus. The electron *does* have a way to release energy: it has an acceleration due to its continuously changing direction of motion, and according to classical physics, any accelerating charged particle emits electromagnetic waves. According to classical physics, atoms should collapse!

The solution lies in the observation that a bound state has a minimum energy. An electron in one of the higher-energy atomic states can and does emit photons and hop down step by step in energy. But once it is in the ground state, it cannot emit a photon because there is no lower-energy state for it to go to.



i / Two hydrogen atoms bond to form an H_2 molecule. In the molecule, the two electrons' wave patterns overlap, and are about twice as wide.

Chemical bonds

example 6

I began this section with a classical argument that chemical bonds, as in an H_2 molecule, should not exist. Quantum physics explains why this type of bonding does in fact occur. There are actually two effects going on, one due to kinetic energy and one due to electrical energy. We'll concentrate on the kinetic energy effect in this example. Example 6 on page 402 revisits the H_2 bond in more detail. (There is also a qualitatively different type of bonding called ionic bonding that occurs when an atom of one element steals the electron from an atom of another element.)

The kinetic energy effect is pretty simple. When the atoms are next to each other, the electrons are shared between them. The “box” is about twice as wide, and a larger box allows a smaller kinetic energy. Energy is required in order to separate the atoms.

Discussion questions

A Neutrons attract each other via the strong nuclear force, so according to classical physics it should be possible to form nuclei out of clusters of two or more neutrons, with no protons at all. Experimental searches, however, have failed to turn up evidence of a stable two-neutron system (dineutron) or larger stable clusters. These systems are apparently not just unstable in the sense of being able to beta decay but unstable in the sense that they don't hold together at all. Explain based on quantum physics why a dineutron might spontaneously fly apart.

B The following table shows the energy gap between the ground state and the first excited state for four nuclei, in units of picojoules. (The nuclei were chosen to be ones that have similar structures, e.g., they are all spherical in shape.)

nucleus	energy gap (picojoules)
4He	3.234
^{16}O	0.968
^{40}Ca	0.536
^{208}Pb	0.418

Explain the trend in the data.

16.4 The uncertainty principle

16.4.1 Eliminating randomness through measurement?

A common reaction to quantum physics, among both early-twentieth-century physicists and modern students, is that we should be able to get rid of randomness through accurate measurement. If I say, for example, that it is meaningless to discuss the path of a photon or an electron, one might suggest that we simply measure the particle's position and velocity many times in a row. This series of snapshots would amount to a description of its path.

A practical objection to this plan is that the process of measurement will have an effect on the thing we are trying to measure. This may not be of much concern, for example, when a traffic cop measures your car's motion with a radar gun, because the energy and momentum of the radar pulses are insufficient to change the car's motion significantly. But on the subatomic scale it is a very real problem. Making a videotape through a microscope of an electron orbiting a nucleus is not just difficult, it is theoretically impossible. The video camera makes pictures of things using light that has bounced off them and come into the camera. If even a single photon of visible light was to bounce off of the electron we were trying to study, the electron's recoil would be enough to change its behavior significantly.

16.4.2 The Heisenberg uncertainty principle

This insight, that measurement changes the thing being measured, is the kind of idea that clove-cigarette-smoking intellectuals outside of the physical sciences like to claim they knew all along. If only, they say, the physicists had made more of a habit of reading literary journals, they could have saved a lot of work. The anthropologist Margaret Mead has recently been accused of inadvertently encouraging her teenaged Samoan informants to exaggerate the freedom of youthful sexual experimentation in their society. If this is considered a damning critique of her work, it is because she could have done better: other anthropologists claim to have been able to eliminate the observer-as-participant problem and collect untainted data.

The German physicist Werner Heisenberg, however, showed that in quantum physics, *any* measuring technique runs into a brick wall when we try to improve its accuracy beyond a certain point. Heisenberg showed that the limitation is a question of *what there is to be known*, even in principle, about the system itself, not of the ability or inability of a specific measuring device to ferret out information that is knowable but not previously hidden.

Suppose, for example, that we have constructed an electron in a box (quantum dot) setup in our laboratory, and we are able to adjust



j / Werner Heisenberg (1901-1976). Heisenberg helped to develop the foundations of quantum mechanics, including the Heisenberg uncertainty principle. He was the scientific leader of the Nazi atomic-bomb program up until its cancellation in 1942, when the military decided that it was too ambitious a project to undertake in wartime, and too unlikely to produce results.

the length L of the box as desired. All the standing wave patterns pretty much fill the box, so our knowledge of the electron's position is of limited accuracy. If we write Δx for the range of uncertainty in our knowledge of its position, then Δx is roughly the same as the length of the box:

$$\Delta x \approx L$$

If we wish to know its position more accurately, we can certainly squeeze it into a smaller space by reducing L , but this has an unintended side-effect. A standing wave is really a superposition of two traveling waves going in opposite directions. The equation $p = h/\lambda$ really only gives the magnitude of the momentum vector, not its direction, so we should really interpret the wave as a 50/50 mixture of a right-going wave with momentum $p = h/\lambda$ and a left-going one with momentum $p = -h/\lambda$. The uncertainty in our knowledge of the electron's momentum is $\Delta p = 2h/\lambda$, covering the range between these two values. Even if we make sure the electron is in the ground state, whose wavelength $\lambda = 2L$ is the longest possible, we have an uncertainty in momentum of $\Delta p = h/L$. In general, we find

$$\Delta p \gtrsim h/L,$$

with equality for the ground state and inequality for the higher-energy states. Thus if we reduce L to improve our knowledge of the electron's position, we do so at the cost of knowing less about its momentum. This trade-off is neatly summarized by multiplying the two equations to give

$$\Delta p \Delta x \gtrsim h.$$

Although we have derived this in the special case of a particle in a box, it is an example of a principle of more general validity:

The Heisenberg uncertainty principle

It is not possible, even in principle, to know the momentum and the position of a particle simultaneously and with perfect accuracy. The uncertainties in these two quantities are always such that $\Delta p \Delta x \gtrsim h$.

(To make this approximate inequality into an exact one, we would need to pick a mathematical definition of our measure of uncertainty. If we use the standard deviation, p. 327, then it can be shown that $\Delta p \Delta x \geq h/4\pi$.)

Note that although I encouraged you to think of this derivation in terms of a specific real-world system, the quantum dot, no reference was ever made to any specific laboratory equipment or procedures. The argument is simply that we cannot *know* the particle's position very accurately unless it *has* a very well defined position, it cannot have a very well defined position unless its wave-pattern

covers only a very small amount of space, and its wave-pattern cannot be thus compressed without giving it a short wavelength and a correspondingly uncertain momentum. The uncertainty principle is therefore a restriction on how much there is to know about a particle, not just on what we can know about it with a certain technique.

An estimate for electrons in atoms *example 7*

- ▷ A typical energy for an electron in an atom is on the order of $(1 \text{ volt}) \cdot e$, which corresponds to a speed of about 1% of the speed of light. If a typical atom has a size on the order of 0.1 nm, how close are the electrons to the limit imposed by the uncertainty principle?
- ▷ If we assume the electron moves in all directions with equal probability, the uncertainty in its momentum is roughly twice its typical momentum. This is only an order-of-magnitude estimate, so we take Δp to be the same as a typical momentum:

$$\begin{aligned}\Delta p \Delta x &= p_{\text{typical}} \Delta x \\ &= (m_{\text{electron}})(0.01c)(0.1 \times 10^{-9} \text{ m}) \\ &= 3 \times 10^{-34} \text{ J}\cdot\text{s}\end{aligned}$$

This is on the same order of magnitude as Planck's constant, so evidently the electron is "right up against the wall." (The fact that it is somewhat less than h is of no concern since this was only an estimate, and we have not stated the uncertainty principle in its most exact form.)

self-check B

If we were to apply the uncertainty principle to human-scale objects, what would be the significance of the small numerical value of Planck's constant?

▷ Answer, p. 458

Discussion questions

- A** Compare Δp and Δx for the two lowest energy levels of the one-dimensional particle in a box, and discuss how this relates to the uncertainty principle.
- B** On a graph of Δp versus Δx , sketch the regions that are allowed and forbidden by the Heisenberg uncertainty principle. Interpret the graph: Where does an atom lie on it? An elephant? Can either p or x be measured with perfect accuracy if we don't care about the other?

16.5 Decoherence and quantum computing

16.5.1 Decoherence

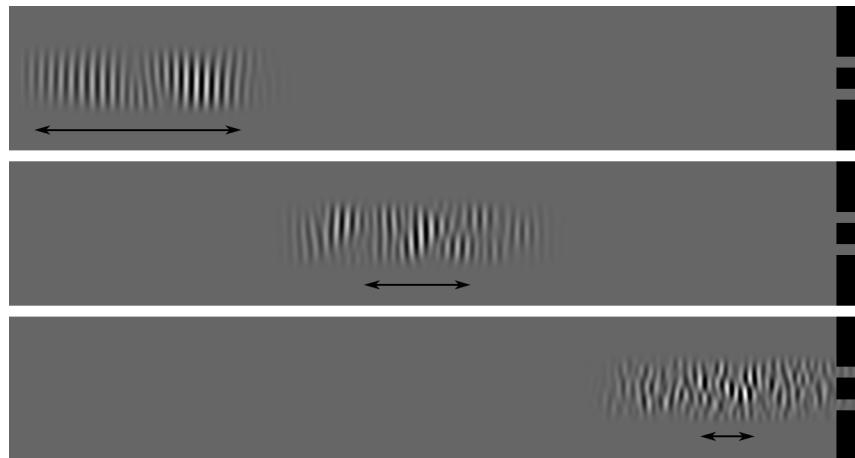
Starting around 1970, physicists began to realize that ideas involving a loss of coherence, or "decoherence," could help to explain some things about quantum mechanics that had previously seemed

mysterious. The classical notions of coherence and coherence length were described in sec. 12.8, p. 273, and quantum-mechanical decoherence was briefly introduced on p. 358.

One mystery was the fact that it is difficult to demonstrate wave interference effects with large objects. This is partly because the wavelength $\lambda = h/p = h/mv$ tends to be small for an object with a large mass (example 1, p. 368). But even taking this into account, we do not seem to have much luck observing, for example, double-slit diffraction of very large molecules, even when we use slits with appropriate dimensions and a detector with a good enough angular resolution.

A large molecule such as the one in the Eibenberger experiment is represented by its wavepacket. As the molecule starts out, its coherence length, shown by the arrows, is quite long. As it flies to the right, it is bombarded by infrared photons, which randomize its phase, causing its coherence length to shorten exponentially: by a factor of two in the second panel, and by a further factor of two in the final one. When the packet enters the double slit, its coherence length is on the same order of magnitude as the slits' spacing d , which will worsen but not entirely eliminate the observability of interference fringes. (This is only a schematic representation, with the wavepacket shown as being many orders of magnitude bigger than its actual size in relation to the vacuum chamber. Also, the real experiment used a reflecting grating, not a transmitting double slit.)

In the early days of quantum mechanics, people like Bohr and Heisenberg imagined that there was simply a clear division between the macroscopic and microscopic worlds. Big things and small things just had different rules: Newton's laws in one case, quantum mechanics in the other. But this is no longer a tenable position, because we now know that there is no limit on the distance scales over which quantum-mechanical behavior can occur. For example, a communication satellite carried out a demonstration in 2017 in which a coherence length of 1200 km was demonstrated using photons.¹



The insight about decoherence was the following. Consider the most massive material object that has so far been successfully diffracted through a grating, which was a molecule consisting of about 810 atoms in an experiment by Eibenberger *et al.* in 2013.² While this molecule was propagating through the apparatus as a wave, the experimenters needed to keep it from simply being stopped by a collision with an air molecule. For this reason, they had to do the experiment inside a vacuum chamber, with an extremely good vacuum. But even then, the molecule was being bombarded by

¹Yin *et al.*, arxiv.org/abs/1707.01339

²arxiv.org/abs/1310.8343

photons of infrared light emitted from the walls of the chamber. The effect of this bombardment is to disrupt the molecule's wavefunction and reduce its coherence length (p. 274).

This decoherence effect was the reason that the experiment was limited to molecules of the size they used. Even though the molecules took only about 400 nanoseconds to fly through the apparatus, there was a significant amount of decoherence. A larger molecule would have been a bigger target for photons and would have undergone decoherence more quickly, making interference unobservable.

16.5.2 Brains, classical computers, and quantum computers

Decoherence helps to explain why it seems as though measurements of quantum systems only produce one result, chosen at random. Suppose you do the Schrödinger's cat experiment (p. 358), after which, in the notation of sec. 16.1.4, p. 372, your brain is in the superposition of states

$$c|\text{Brain L}\rangle + c'|\text{Brain D}\rangle.$$

Although absolute phases and normalizations are not physically observable, relative phases are. So in principle, you could tell that your brain was in a superposed state, because, for example, the state

$$|\text{Brain L}\rangle + |\text{Brain D}\rangle$$

is different from

$$|\text{Brain L}\rangle - |\text{Brain D}\rangle,$$

which has a different relative phase. This really does work with electrons, since $|\psi\rangle + |\psi\rangle$ gives a different interference pattern than $|\psi\rangle - |\psi\rangle$. But because your brain is a macroscopic object exposed to lots of subatomic collisions, decoherence affects it on an incredibly short timescale. It randomizes the relative phases of the $|\text{Brain L}\rangle$ and $|\text{Brain D}\rangle$ parts, so that we can never observe any interference pattern between them.

For these reasons, the brain is a classical computer. But there is a completely different type of computing device, the quantum computer. Rather than bits, a quantum computer operates on qubits. A qubit, or quantum bit, can exist in a mixture of states like $c|0\rangle + c'|1\rangle$. You might think that a quantum computer would be worse than a classical one, since it would probably act like a classical computer with some randomness thrown in, making it unreliable. But the randomness of quantum mechanics is only apparent. It crops up only when decoherence happens, as often occurs in measurements. A quantum computer has to be isolated cleanly from its environment, so that there will be no decoherence. In such an environment, there really is an observable difference between $|0\rangle + |1\rangle$ and $|0\rangle - |1\rangle$. Furthermore, the different qubits in a quantum computer can become entangled in complicated ways. A chunk of memory in a quantum computer is therefore a much richer structure than a same-size

chunk of classical bits. This leads to a concept of quantum information, which is a higher grade of information than classical information. That is, we can losslessly convert from classical information to quantum information, but not the other way around.

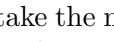
Quantum computers can efficiently solve certain types of problems that a classical computer finds intractable, such as factoring large numbers into their prime factors. For this reason, the cryptographic infrastructure that lets you safely transmit your credit card number to amazon.com will become completely broken if large-scale quantum computing ever becomes a reality. Luckily for our privacy, quantum computers are hard to build, and the biggest factoring problems they can currently solve are examples like $15 = 3 \times 5$.

16.6 A crude model of the hydrogen atom

16.6.1 Modeling

Probably the most important reason for wanting to understand quantum physics is in order to understand the atom. As tiny as they may be, atoms can be pretty complicated. Let's think about ways to simplify them — even at the risk of oversimplifying. Start with hydrogen, because it's the simplest of all atoms, consisting of only one electron and one proton. Because the proton is 1800 times more massive than the electron, it's a good approximation to imagine that it stays at rest while the electron moves around, influenced by the proton's static electric field. This is good: we've reduced our problem to the discussion of just one particle.

Classically, the electron moves around in the proton's electric field, speeding up, slowing down, and never straying farther than a certain distance r because it doesn't have enough energy. As it does all these things, its momentum changes, so if we mix in a little quantum mechanics, $p = \hbar k$ then tells us that its wavenumber and wavelength change. So its wavefunction can't be a nice, simple sine wave with a single well-defined wavelength. We expect it to look more like this:  When the electron is close to the proton, its electrical energy is low, so its kinetic energy is high, and its wavelength is short. When it's farther out, its wavelength is longer. In chapter 17 we'll talk in more detail about how to handle this sort of thing.

But for now we're going to do something much more crude, which is to *approximate* the electron's wavefunction as a sine wave like this one . We take the nodes and antinodes and make them nice and regular: all evenly spaced and equal in amplitude, and filling the whole space available to the electron based on the amount of energy it has.

Also, by drawing the wavefunctions on a piece of paper in this way, I've snuck yet another crude approximation past you. We're

treating the hydrogen atom as one-dimensional, $\Psi(x)$ rather than $\Psi(x, y, z)$. The wave fills in the line segment $-r \leq x \leq r$, where r is the maximum distance. The drawings also implicitly imply that Ψ is a function whose outputs are real rather than complex, but we can get away with this because this is a standing wave.

16.6.2 Estimation of the energy levels

Let n be the same quantum number defined in problem 12, p. 59, and in sec. 16.1.3, p. 372, i.e., it's the number of antinodes or the number of half-wavelengths. In the examples drawn above, $n = 9$. The ground state has $n = 1$. Because n is the number of half-wavelengths, the wavelength is $4r/n$, but because this whole calculation is so crude, we're going to throw away factors like the 4, and just say that

$$\lambda \sim r/n.$$

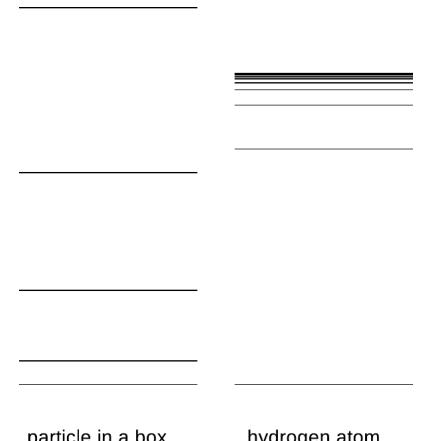
If r had the same value for every standing-wave pattern, then we'd essentially be solving the particle-in-a-box problem, but there is a fundamental difference here, which is that we don't have a box of a fixed size with impenetrable walls. If we did, then it would be impossible to separate the electron from the proton, whereas in real life we know that this can be done by putting in a certain amount of energy to ionize the atom. The force keeping the electron bound isn't an infinite force encountered when it bounces off of a wall, it's the attractive electrical force from the nucleus. If we put more energy into the electron, it's like throwing a ball upward with a higher energy — it will get farther out before coming back down.

Figure 1 shows how we expect this to turn out. In the hydrogen atom, we expect r to increase as we go to states of higher energy. This tends to keep the wavelengths of the high energy states from getting too short, reducing their kinetic energy. The closer and closer crowding of the energy levels in hydrogen also makes sense because we know that there is a certain energy that would be enough to make the electron escape completely, and therefore the sequence of bound states cannot extend above that energy.

Now let's crunch some numbers. When the electron is at the maximum classically allowed distance r from the proton, it has zero kinetic energy. Thus when the electron is at distance r , its energy is purely electrical,

$$[1] \quad E = -\frac{ke^2}{r},$$

where k is the Coulomb constant (not the wavenumber). The zero-level of the electrical energy scale is chosen to be the energy of an electron and a proton that are infinitely far apart. With this choice, negative energies correspond to bound states and positive energies to unbound ones.



particle in a box hydrogen atom

I / The energy levels of a particle in a box, contrasted with those of the hydrogen atom.

Finally we assume that the typical kinetic energy of the electron is on the same order of magnitude as the absolute value of its total energy. (This is true to within a factor of two for a typical classical system like a planet in a circular orbit around the sun.) We then have

$$\begin{aligned}
 [2] \quad & \text{absolute value of total energy} \\
 & = \frac{ke^2}{r} \\
 & \sim K \\
 & = p^2/2m \\
 & = (h/\lambda)^2/2m \\
 & \sim h^2 n^2 / 2mr^2
 \end{aligned}$$

We now solve the equation $ke^2/r \sim h^2 n^2 / 2mr^2$ for r and throw away numerical factors we can't hope to have gotten right, yielding

$$[3] \quad r \sim \frac{h^2 n^2}{m k e^2}.$$

Plugging $n = 1$ into this equation gives $r = 2$ nm, which is indeed on the right order of magnitude compared to the observed sizes of atoms. Finally we combine equations [3] and [1] to find

$$E \sim -\frac{mk^2e^4}{h^2} \cdot \frac{1}{n^2} \quad [\text{result of crude treatment}]$$

which turns out to be correct except for the numerical factors we never aimed to find. The exact result is

$$E = -\frac{mk^2e^4}{2\hbar^2} \cdot \frac{1}{n^2}, \quad [\text{exact result}]$$

which is different by a factor of $2\pi^2$. It might seem incredible that such a crude approximation would provide a result so close to the correct one. However, the factor in front is constrained by the requirement that the result have units of energy. All of the exponents in this factor have to be as they are just based on units. For the ground state, $n = 1$, we will show in example 5, p. 401, that the exact result is correct.

16.6.3 Comparison with experiment

The experimental technique for measuring the energy levels of an atom accurately is spectroscopy: the study of the spectrum of light emitted (or absorbed) by the atom. Only photons with certain energies can be emitted or absorbed by a hydrogen atom, for example, since the amount of energy gained or lost by the atom must equal the difference in energy between the atom's initial and final states. Spectroscopy had become a highly developed art several

decades before Einstein even proposed the photon, and the Swiss spectroscopist Johann Balmer determined in 1885 that there was a simple equation that gave all the wavelengths emitted by hydrogen. In modern terms, we think of the photon wavelengths merely as indirect evidence about the underlying energy levels of the atom, and we rework Balmer's result into an equation for these atomic energy levels:

$$E_n = -\frac{A}{n^2},$$

where $A = 2.2 \times 10^{-18}$ J. About 30 years later, the constant A was explained to be the constant factor in the equation at the end of the previous section.

Discussion questions

- A** Sketch graphs of r and E versus n for the hydrogen atom, and compare with analogous graphs for the one-dimensional particle in a box.
- B** States of hydrogen with n greater than about 10 are never observed in the sun. Why might this be?

Notes for chapter 16

2366 Failure of the classical H₂ molecule

We use classical physics to make an estimate of the electrical binding energy of the H₂ molecule. It is much too small compared to reality, and much too small to make the molecule hold together.

To give classical physics a fair chance, we should not use any classical model with moving parts, such as the planetary model, because such models are always unstable due to their ability to radiate away energy as electromagnetic waves. We also want a model that is fairly easy to calculate with. Let's use a model in which the proton is a point charge fixed at the center of a rigid electron cloud whose density of charge is e^{-r} , where r is in units of some basic distance scale (about 0.05 nm in SI units), and the charge is in units of the fundamental charge. This is actually somewhat realistic as a time-averaged picture of the charge of a hydrogen atom, and it turns out that the electrical potential energy we need has been calculated by A. Mukherji, arxiv.org/abs/0903.3304. We assume that the two clouds simply interpenetrate one another. The resulting electrical energy, including electron-electron, proton-proton, and electron-proton interactions, turns out to be

$$U = e^{-r} \left(\frac{1}{r} + \frac{5}{16} - \frac{3}{16}r - \frac{1}{48}r^2 \right).$$

This function has a very shallow minimum at $r \approx 3.8$ (which is about triple the real-world value for the bond length), attaining a minimum value of about -0.01 in our energy units, which converts to -0.3 eV. The negative sign does indicate an equilibrium, but the binding energy is far too small compared to the real-world value of -4.7 eV. Not only is the calculated binding energy an order of magnitude too small compared to experiment, but it is of a size that would probably cause the molecule to break apart spontaneously due to the Heisenberg uncertainty principle.

Problems

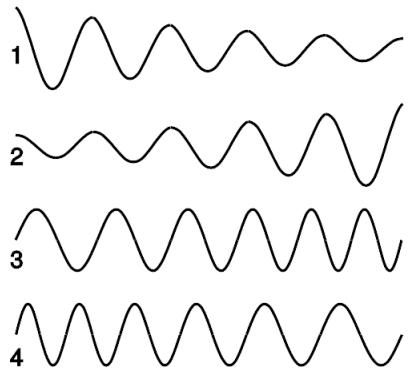
Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 In a television, suppose the electrons are accelerated from rest through a voltage difference of 10^4 V. What is their final wavelength?

✓

2 The figures show the wavefunction of an electron as a function of position. Which one could represent an electron speeding up as it moves to the right? Explain.



3 Use the Heisenberg uncertainty principle to estimate the minimum velocity of a proton or neutron in a ^{208}Pb nucleus, which has a diameter of about 13 fm ($1 \text{ fm} = 10^{-15} \text{ m}$). Assume that the speed is nonrelativistic, and then check at the end whether this assumption was warranted.

✓

4 Suppose that an electron, in one dimension, is confined to a certain region of space so that its wavefunction is given by

$$\Psi = \begin{cases} 0 & \text{if } x < 0 \\ A \sin(2\pi x/L) & \text{if } 0 \leq x \leq L \\ 0 & \text{if } x > L \end{cases}$$

Determine the constant A from normalization.

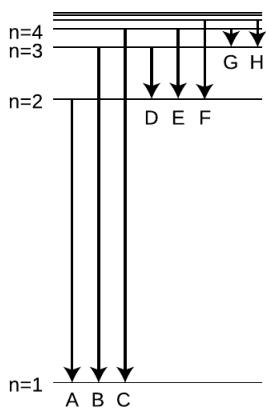
✓

5 Find the energy of a nonrelativistic particle in a one-dimensional box of length L , expressing your result in terms of L , the particle's mass m , the number of peaks and valleys n in the wavefunction, and fundamental constants.

✓

6 A free electron that contributes to the current in an ohmic material typically has a speed of 10^5 m/s (much greater than the drift velocity).

- (a) Estimate its de Broglie wavelength, in nm. ✓
- (b) If a computer memory chip contains 10^8 electric circuits in a 1 cm^2 area, estimate the linear size, in nm, of one such circuit. ✓
- (c) Based on your answers from parts a and b, does an electrical engineer designing such a chip need to worry about wave effects such as diffraction?
- (d) Estimate the maximum number of electric circuits that can fit on a 1 cm^2 computer chip before quantum-mechanical effects become important.



Problem 7.

7 The figure shows eight of the possible ways in which an electron in a hydrogen atom could drop from a higher energy state to a state of lower energy, releasing the difference in energy as a photon. Of these eight transitions, only D, E, and F produce photons with wavelengths in the visible spectrum.

(a) Which of the visible transitions would be closest to the violet end of the spectrum, and which would be closest to the red end? Explain.

(b) In what part of the electromagnetic spectrum would the photons from transitions A, B, and C lie? What about G and H? Explain.

(c) Is there an upper limit to the wavelengths that could be emitted by a hydrogen atom going from one bound state to another bound state? Is there a lower limit? Explain.

8 Find an equation for the wavelength of the photon emitted when the electron in a hydrogen atom makes a transition from energy level n_1 to level n_2 . ✓

9 Hydrogen is the only element whose energy levels can be expressed exactly in an equation. Calculate the ratio λ_E/λ_F of the wavelengths of the transitions labeled E and F in problem 7 on p. 390. Express your answer as an exact fraction, not a decimal approximation. In an experiment in which atomic wavelengths are being measured, this ratio provides a natural, stringent check on the precision of the results. ✓

10 Assume that the kinetic energy of an electron in the $n = 1$ state of a hydrogen atom is on the same order of magnitude as the absolute value of its total energy, and estimate a typical speed at which it would be moving. (It cannot really have a single, definite speed, because its kinetic and interaction energy trade off at different distances from the proton, but this is just a rough estimate of a typical speed.) Based on this speed, were we justified in assuming that the electron could be described nonrelativistically?

11 Use physical reasoning to explain how the equation for the energy levels of hydrogen,

$$E_n = -\frac{mk^2e^4}{2\hbar^2} \cdot \frac{1}{n^2},$$

should be generalized to the case of an atom with atomic number Z that has had all its electrons removed except for one.

12 A muon is a subatomic particle that acts exactly like an electron except that its mass is 207 times greater. Muons can be created by cosmic rays, and it can happen that one of an atom's electrons is displaced by a muon, forming a muonic atom. If this happens to a hydrogen atom, the resulting system consists simply of a proton plus a muon.

- (a) How would the size of a muonic hydrogen atom in its ground state compare with the size of the normal atom?
(b) If you were searching for muonic atoms in the sun or in the earth's atmosphere by spectroscopy, in what part of the electromagnetic spectrum would you expect to find the absorption lines?

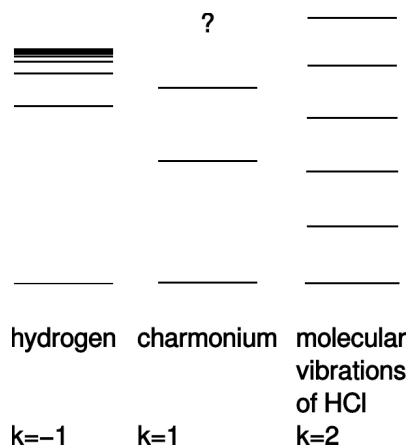
13 Before the quantum theory, experimentalists noted that in many cases, they would find three lines in the spectrum of the same atom that satisfied the following mysterious rule: $1/\lambda_1 = 1/\lambda_2 + 1/\lambda_3$. Explain why this would occur. Do not use reasoning that only works for hydrogen — such combinations occur in the spectra of all elements. [Hint: Restate the equation in terms of the energies of photons.]

14 In sec. 16.6.2, p. 385, we derived an approximate expression for the energies of states in hydrogen. As input to the calculation, we used the the proportionality $U \propto r^{-1}$, which is a characteristic of the electrical interaction. The result for the energy of the n th standing wave pattern was $E_n \propto n^{-2}$.

There are other systems of physical interest in which we have $U \propto r^k$ for values of k besides -1 . Problem 4 discusses the ground state of the harmonic oscillator, with $k = 2$ (and a positive constant of proportionality). In particle physics, systems called charmonium and bottomonium are made out of pairs of subatomic particles called quarks, which interact according to $k = 1$, i.e., a force that is independent of distance. (Here we have a positive constant of proportionality, and $r > 0$ by definition. The motion turns out not to be too relativistic, so the Schrödinger equation is a reasonable approximation.) The figure shows actual energy levels for these three systems, drawn with different energy scales so that they can all be shown side by side. The sequence of energies in hydrogen approaches a limit, which is the energy required to ionize the atom. In charmonium, only the first three levels are known.³

Generalize the method used for $k = -1$ to any value of k , and find the exponent j in the resulting proportionality $E_n \propto n^j$. Compare the theoretical calculation with the behavior of the actual energies shown in the figure. Comment on the limit $k \rightarrow \infty$. \checkmark

³See Barnes et al., “The XYZs of Charmonium at BES,” arxiv.org/abs/hep-ph/0608103. To avoid complication, the levels shown are only those in the group known for historical reasons as the Ψ and J/Ψ .



Problem 14.

- 15** The electron, proton, and neutron were discovered, respectively, in 1897, 1919, and 1932. The neutron was late to the party, and some physicists felt that it was unnecessary to consider it as fundamental. Maybe it could be explained as simply a proton with an electron trapped inside it. The charges would cancel out, giving the composite particle the correct neutral charge, and the masses at least approximately made sense (a neutron is heavier than a proton). (a) Given that the diameter of a proton is on the order of 10^{-15} m, use the Heisenberg uncertainty principle to estimate the trapped electron's minimum momentum. ✓
(b) Find the electron's minimum kinetic energy. ✓
(c) Show via $E = mc^2$ that the proposed explanation may have a problem, because the contribution to the neutron's mass from the electron's kinetic energy would be comparable to the neutron's entire mass.

Chapter 17

The Schrödinger equation

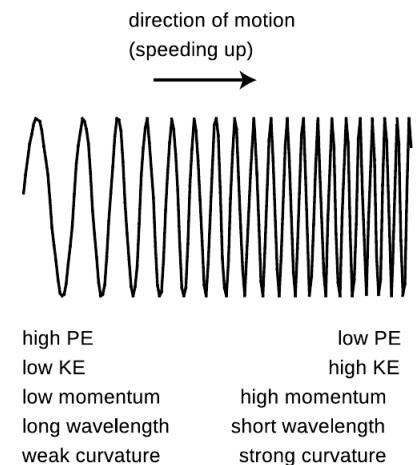
17.1 Electrons in electric fields

17.1.1 Defining a wavelength when the wavelength is varying

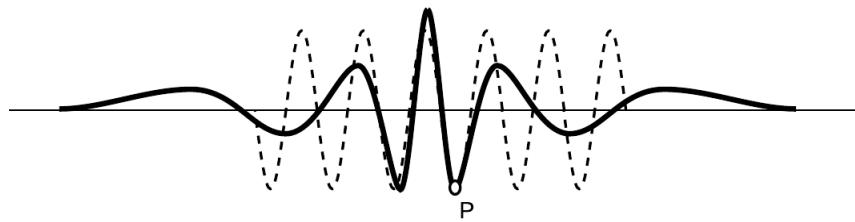
So far the only electron wave patterns we've considered have been simple sine waves, but whenever an electron finds itself in an electric field, it must have a more complicated wave pattern. Let's consider the example of an electron being accelerated by the electron gun at the back of a TV tube. Newton's laws are not useful, because they implicitly assume that the path taken by the particle is a meaningful concept. Conservation of energy is still valid in quantum physics, however. In terms of energy, the electron is moving from a region of low voltage into a region of higher voltage. Since its charge is negative, it loses electrical energy by moving to a higher voltage, so its kinetic energy increases. As its electrical energy goes down, its kinetic energy goes up by an equal amount, keeping the total energy constant. Increasing kinetic energy implies a growing momentum, and therefore a shortening wavelength, a .

The wavefunction as a whole does not have a single well-defined wavelength, but the wave changes so gradually that if you only look at a small part of it you can still pick out a wavelength and relate it to the momentum and energy. (The picture actually exaggerates by many orders of magnitude the rate at which the wavelength changes.)

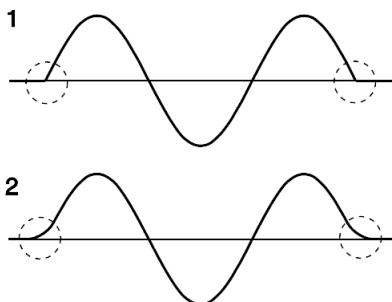
But what if the electric field was stronger? The electric field in an old-fashioned vacuum tube TV screen is only $\sim 10^5$ N/C, but the electric field within an atom is more like 10^{12} N/C. In figure b, the wavelength changes so rapidly that there is nothing that looks like a sine wave at all. We could get a rough idea of the wavelength in a given region by measuring the distance between two peaks, but that would only be a rough approximation. Suppose we want to know the wavelength at point P . The trick is to construct a sine wave, like the one shown with the dashed line, which matches the curvature of the actual wavefunction as closely as possible near P . The sine wave that matches as well as possible is called the “osculating” curve, from a Latin word meaning “to kiss.” The wavelength of the osculating curve is the wavelength that will relate correctly to conservation of energy.



a / An electron in a gentle electric field gradually shortens its wavelength as it gains energy. (As discussed on p. 371, it is actually not quite correct to graph the wavefunction of an electron as a real number unless it is a standing wave, which isn't the case here.)



b / A typical wavefunction of an electron in an atom (heavy curve) and the osculating sine wave (dashed curve) that matches its curvature at point P.



c / The wavefunction's tails go where classical physics says they shouldn't.

17.1.2 Tunneling

We implicitly assumed that the particle-in-a-box wavefunction would cut off abruptly at the sides of the box, c/1, but that would be unphysical. A kink has infinite curvature, and curvature is related to energy, so it can't be infinite. A physically realistic wavefunction must always "tail off" gradually, c/2. In classical physics, a particle can never enter a region in which its interaction energy U would be greater than the amount of energy it has available. But in quantum physics the wavefunction will always have a tail that reaches into the classically forbidden region. If it was not for this effect, called tunneling, the fusion reactions that power the sun would not occur due to the high electrical energy nuclei need in order to get close together! Tunneling is discussed in more detail on p. 397.

17.2 The Schrödinger equation

In section 17.1 we were able to apply conservation of energy to an electron's wavefunction, but only by using the clumsy graphical technique of osculating sine waves as a measure of the wave's curvature. You have learned a more convenient measure of curvature in calculus: the second derivative. To relate the two approaches, we take the second derivative of a sine wave:

$$\begin{aligned}\frac{d^2}{dx^2} \sin kx &= \frac{d}{dx} (k \cos kx) \\ &= -k^2 \sin kx.\end{aligned}$$

Taking the second derivative gives us back the same function, but with a minus sign and a constant out in front that is related to the wavelength. We can thus relate the second derivative to the osculating wavelength:

$$[1] \quad \frac{d^2 \Psi}{dx^2} = -k^2 \Psi,$$

where $k = 2\pi/\lambda$. This could be solved for k or λ in terms of Ψ , but it will turn out below to be more convenient to leave it in this form.

Applying this to conservation of energy, we have

$$\begin{aligned} E &= K + U \\ [2] \quad &= \frac{p^2}{2m} + U \\ &= \frac{(\hbar k)^2}{2m} + U \end{aligned}$$

We can simplify our algebra by multiplying both sides of equation [2] by Ψ to make it look more like equation [1]:

$$E \cdot \Psi = \frac{(\hbar k)^2}{2m} \Psi + U \cdot \Psi,$$

which leads to the important result known as the **Schrödinger equation**:

$$E \cdot \Psi = -\frac{\hbar^2}{2m} \frac{d^2 \Psi}{dx^2} + U \cdot \Psi$$

(Actually this is a simplified version of the Schrödinger equation, applying only to standing waves in one dimension.) Physically it is a statement of conservation of energy. The total energy E must be constant, so the equation tells us that a change in interaction energy U must be accompanied by a change in the curvature of the wavefunction. This change in curvature relates to a change in wavelength, which corresponds to a change in momentum and kinetic energy.

self-check A

Considering the assumptions that were made in deriving the Schrödinger equation, would it be correct to apply it to a photon? To an electron moving at relativistic speeds?

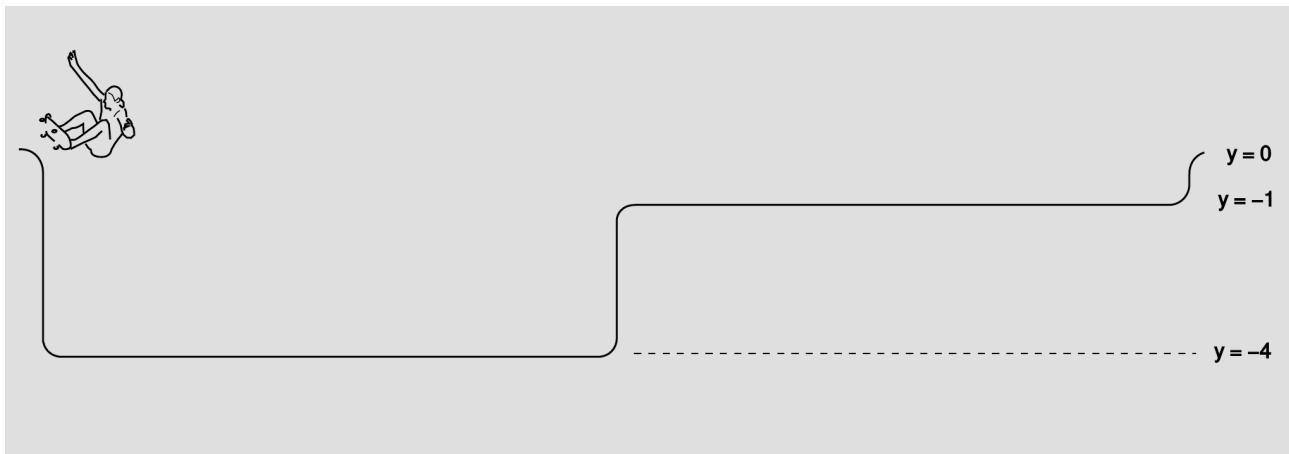
▷ Answer, p.

458

Usually we know right off the bat how U depends on x , so the basic mathematical problem of quantum physics is to find a function $\Psi(x)$ that satisfies the Schrödinger equation for a given interaction-energy function $U(x)$. An equation, such as the Schrödinger equation, that specifies a relationship between a function and its derivatives is known as a differential equation.

Discussion question

A The figure shows a skateboarder tipping over into a swimming pool with zero initial kinetic energy. There is no friction, the corners are smooth enough to allow the skater to pass over the smoothly, and the vertical distances are small enough so that negligible time is required for the vertical parts of the motion. The pool is divided into a deep end and a shallow end. Their widths are equal. The deep end is four times deeper. (1) Classically, compare the skater's velocity in the left and right regions, and infer the probability of finding the skater in either of the two halves if an observer peeks at a random moment. (2) Quantum-mechanically, this could be a one-dimensional model of an electron shared between two atoms in a diatomic molecule. Compare the electron's kinetic energies, momenta, and wavelengths in the two sides. For simplicity, let's assume that there is no tunneling into the classically forbidden regions. What is the simplest standing-wave pattern that you can draw, and what are the probabilities of finding the electron in one side or the other? Does this obey the correspondence principle?



17.3 Solutions when U is constant; tunneling

The detailed study of the solution of the Schrödinger equation is beyond the scope of this book, but we can gain some important insights by considering the easiest version of the Schrödinger equation, in which the interaction energy U is constant. We can then rearrange the Schrödinger equation as follows:

$$\frac{d^2 \Psi}{dx^2} = \frac{2m(U - E)}{\hbar^2} \Psi,$$

which boils down to

$$\frac{d^2 \Psi}{dx^2} = a \Psi,$$

where, according to our assumptions, a is independent of x . We need to find a function whose second derivative is the same as the original function except for a multiplicative constant. The only functions with this property are sine waves and exponentials:

$$\begin{aligned}\frac{d^2}{dx^2} [q \sin(rx + s)] &= -qr^2 \sin(rx + s) \\ \frac{d^2}{dx^2} [qe^{rx+s}] &= qr^2 e^{rx+s}\end{aligned}$$

The sine wave gives negative values of a , $a = -r^2$, and the exponential gives positive ones, $a = r^2$. The former applies to the classically allowed region with $U < E$.

This leads us to a quantitative calculation of the tunneling effect discussed briefly in sec. 17.1.2, p. 394. The wavefunction evidently tails off exponentially in the classically forbidden region. Suppose, as shown in figure d, a wave-particle traveling to the right encounters a barrier that it is classically forbidden to enter. Although the form of the Schrödinger equation we're using technically does not apply to traveling waves (because it makes no reference to time), it turns out that we can still use it to make a reasonable calculation of the probability that the particle will make it through the barrier. If we let the barrier's width be w , then the ratio of the wavefunction on the left side of the barrier to the wavefunction on the right is

$$\frac{qe^{rx+s}}{qe^{r(x+w)+s}} = e^{-rw}.$$

Probabilities are proportional to the squares of wavefunctions, so the probability of making it through the barrier is

$$\begin{aligned}P &= e^{-2rw} \\ &= \exp\left(-\frac{2w}{\hbar} \sqrt{2m(U-E)}\right).\end{aligned}$$

self-check B

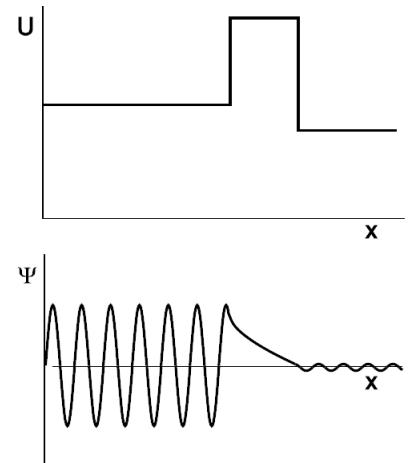
If we were to apply this equation to find the probability that a person can walk through a wall, what would the small value of Planck's constant imply?

▷ Answer, p. 458

Tunneling in alpha decay

example 1

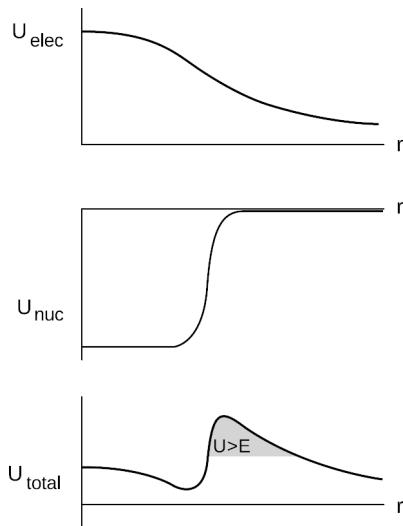
Naively, we would expect alpha decay to be a very fast process. The typical speeds of neutrons and protons inside a nucleus are extremely high (see problem 3). If we imagine an alpha particle coalescing out of neutrons and protons inside the nucleus, then at the typical speeds we're talking about, it takes a ridiculously small amount of time for them to reach the surface and try to escape. Clattering back and forth inside the nucleus, we could



d / Tunneling through a barrier. (As discussed on p. 371, it is actually not quite correct to graph the wavefunction of an electron as a real number unless it is a standing wave, which isn't the case here.)

imagine them making a vast number of these “escape attempts” every second.

Consider figure e, however, which shows the interaction energy for an alpha particle escaping from a nucleus. The electrical energy is kq_1q_2/r when the alpha is outside the nucleus, while its variation inside the nucleus has the shape of a parabola, as a consequence of the shell theorem. The nuclear energy is constant when the alpha is inside the nucleus, because the forces from all the neighboring neutrons and protons cancel out; it rises sharply near the surface, and flattens out to zero over a distance of ~ 1 fm, which is the maximum distance scale at which the strong force can operate. There is a classically forbidden region immediately outside the nucleus, so the alpha particle can only escape by quantum mechanical tunneling. (It’s true, but somewhat counterintuitive, that a *repulsive* electrical force can make it more difficult for the alpha to get out.)



e / The electrical, nuclear, and total interaction energies for an alpha particle escaping from a nucleus.

In reality, alpha-decay half-lives are often extremely long — sometimes billions of years — because the tunneling probability is so small. Although the shape of the barrier is not a rectangle, the equation for the tunneling probability on page 397 can still be used as a rough guide to our thinking. Essentially the tunneling probability is so small because $U - E$ is fairly big, typically about 30 MeV at the peak of the barrier.

A marble tunneling out of a box

example 2

On p. 334, I introduced Gell-Mann’s whimsically named totalitarian principle, that any process not forbidden by a conservation law will happen with some nonzero probability or rate. One of the examples I used there was a marble locked in a box. As a silly example, let’s make a crude numerical estimate of the marble’s probability of tunneling out through the box. As in self-check B on p. 397, we expect based on the correspondence principle that this probability will be very small.

A typical marble has a radius on the order of a centimeter, and let’s say the box has about that thickness. We’ll say that the marble has a mass of 5 g and contains something like $n = 10^{23}$ atoms. The atomic energy scale is roughly 1 eV, so that if the marble is to be *inside* the wall of the box, its electrical energy U probably has to be something on the order of n multiplied by 1 eV, giving $U \sim 10^4$ J, or about an order of magnitude greater than the kinetic energy of a bullet. We could of course give the marble that much energy, and then it would blast through the box easily, but the point here is to estimate its probability of just tunneling out, so we’ll say that the kinetic energy is negligible, and therefore $U - E$ is basically the same as U . Plugging in numbers, we find $P \sim e^{-10^{33}}$. When you have to express a small number using a stack of exponents like this, you know that it’s very small. For

example, this number is many, many orders of magnitude smaller than 10^{-1000} .

Beta decay: a push or pull on the way out the door example 3

The nucleus ^{64}Cu undergoes β^+ and β^- decay with similar probabilities and energies. Each of these decays releases a fixed amount of energy Q due to the difference in mass between the parent nucleus and the decay products. This energy is shared randomly between the beta and the neutrino. In experiments, the beta's energy is easily measured, while the neutrino flies off without interacting. Figure f shows the energy spectrum of the β^+ and β^- in these decays.¹ There is a relatively high probability for the beta and neutrino each to carry off roughly half the kinetic energy, the reason being identical to the kind of phase-space argument discussed in sec. 9.4, p. 205. Therefore in each case we get a bell-shaped curve stretching from 0 up to the energy Q , with Q being slightly different in the two cases.

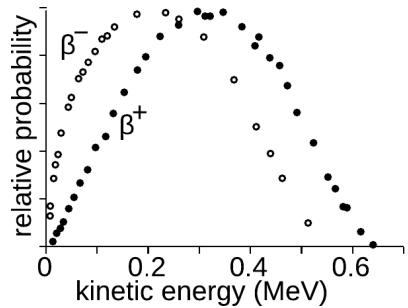
So we expect the two bell curves to look almost the same except for a slight rescaling of the horizontal axis. Yes — but we also see markedly different behavior at low energies. At very low energies, there is almost no chance to see a β^+ with very low energy, but quite a high probability for a β^- .

We could try to explain this difference in terms of the release of electrical energy. The β^+ is repelled by the nucleus, so it gets an extra push on the way out the door. A β^- should be held back as it exits, and so should lose some energy. The bell curves should be shifted up and down in energy relative to one another, as observed.

But if we try to estimate this energy shift using classical physics, we come out with a wildly incorrect answer. This would be a process in which the beta and neutrino are released in a point-like event inside the nucleus. The radius r of the ^{64}Cu nucleus is on the order of 4 fm ($1 \text{ fm} = 10^{-15} \text{ m}$). Therefore the energy lost or gained by the β^+ or β^- on the way out would be $U \sim kZ\epsilon^2/r \sim 10 \text{ MeV}$. The actual shift is much smaller.

To understand what's really going on, we need quantum mechanics. A beta in the observed energy range has a wavelength of about 2000 fm, which is hundreds of times greater than the size of the nucleus. Therefore the beta cannot be much better localized than that when it is emitted. This means that we should really use something more like $r \sim 500 \text{ fm}$ (a quarter of a wavelength) in our calculation of the electrical energy. This gives $U \sim 0.08 \text{ MeV}$, which is about the right order of magnitude compared to observation.

A byproduct of this analysis is that a β^+ is always emitted within



f / β^+ and β^- spectra of ^{64}Cu .

¹Redrawn from Cook and Langer, 1948.

the classically forbidden region, and then has to tunnel out through the barrier. As in example 1, we have the counterintuitive fact about quantum mechanics that a repulsive force can *hinder* the escape of a particle.

17.4 Three dimensions

For simplicity, we've been considering the Schrödinger equation in one dimension, so that Ψ is only a function of x , and has units of $\text{m}^{-1/2}$ rather than $\text{m}^{-3/2}$. Since the Schrödinger equation is a statement of conservation of energy, and energy is a scalar, the generalization to three dimensions isn't particularly complicated. The total energy term $E \cdot \Psi$ and the interaction energy term $U \cdot \Psi$ involve nothing but scalars, and don't need to be changed at all. In the kinetic energy term, however, we're essentially basing our computation of the kinetic energy on the squared magnitude of the momentum, p_x^2 , and in three dimensions this would clearly have to be generalized to $p_x^2 + p_y^2 + p_z^2$. The obvious way to achieve this is to replace the second derivative $d^2 \Psi / dx^2$ with the sum $\partial^2 \Psi / \partial x^2 + \partial^2 \Psi / \partial y^2 + \partial^2 \Psi / \partial z^2$. In other words, we replace the second derivative with the Laplacian (sec. 2.7.4, p. 53),

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2},$$

just as we did for waves such as sound. We recall from sec. 2.7.4 that:

- The partial derivative symbol ∂ , introduced on page 40, indicates that when differentiating with respect to a particular variable, the other variables are to be considered as constants.
- Like the second derivative, the Laplacian is essentially a measure of curvature.
- As shown in figure ag, p. 55, we can also think of the Laplacian as a measure of how much the value of a function at a certain point differs from the average of its value on nearby points.

A classically allowed region with constant U *example 4*
In a classically allowed region with constant U , we expect the solutions to the Schrödinger equation to be sine waves. A sine wave in three dimensions has the form

$$\Psi = \sin(k_x x + k_y y + k_z z).$$

When we compute $\partial^2 \Psi / \partial x^2$, double differentiation of sin gives $-\sin$, and the chain rule brings out a factor of k_x^2 . Applying all three second derivative operators, we get

$$\begin{aligned}\nabla^2 \Psi &= \left(-k_x^2 - k_y^2 - k_z^2\right) \sin(k_x x + k_y y + k_z z) \\ &= -\left(k_x^2 + k_y^2 + k_z^2\right) \Psi.\end{aligned}$$

The Schrödinger equation gives

$$\begin{aligned} E \cdot \Psi &= -\frac{\hbar^2}{2m} \nabla^2 \Psi + U \cdot \Psi \\ &= -\frac{\hbar^2}{2m} \cdot -\left(k_x^2 + k_y^2 + k_z^2\right) \Psi + U \cdot \Psi \\ E - U &= \frac{\hbar^2}{2m} \left(k_x^2 + k_y^2 + k_z^2\right), \end{aligned}$$

which can be satisfied since we're in a classically allowed region with $E - U > 0$, and the right-hand side is manifestly positive.

Exact treatment of the ground state of hydrogen example 5

The general treatment of the hydrogen atom for all values of n is beyond the mathematical scope of this book, but it's fairly straightforward to verify it for a particular n , especially given a lucky guess as to what functional form to try for the wavefunction. The form that works for the ground state is

$$\Psi = ue^{-r/a},$$

where $r = \sqrt{x^2 + y^2 + z^2}$ is the electron's distance from the proton, and u provides for normalization. We showed in example 9, p. 55 that the Laplacian of this function is

$$\nabla^2 \Psi = \left(-\frac{2}{ar} + \frac{1}{a^2}\right) \Psi.$$

The Schrödinger equation gives

$$\begin{aligned} E \cdot \Psi &= -\frac{\hbar^2}{2m} \nabla^2 \Psi + U \cdot \Psi \\ &= \frac{\hbar^2}{2m} \left(\frac{2}{ar} - \frac{1}{a^2}\right) \Psi - \frac{ke^2}{r} \cdot \Psi \end{aligned}$$

If we require this equation to hold for all r , then we must have equality for both the terms of the form (constant) $\times \Psi$ and for those of the form (constant/ r) $\times \Psi$. That means

$$E = -\frac{\hbar^2}{2ma^2}$$

and

$$0 = \frac{\hbar^2}{mar} - \frac{ke^2}{r}.$$

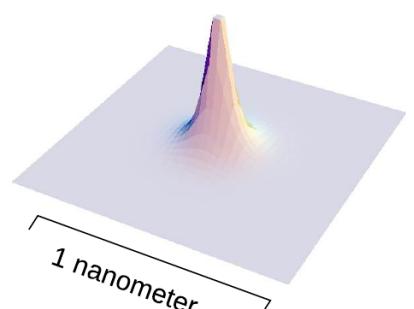
These two equations can be solved for the unknowns a and E , giving

$$a = \frac{\hbar^2}{mke^2}$$

and

$$E = -\frac{mk^2e^4}{2\hbar^2},$$

where the result for the energy agrees with the equation given on p. 386. The calculation of the normalization constant u is relegated to homework problem 3.



g / The ground-state wavefunction of the hydrogen atom, example 5, is graphed in the x - y plane as $\Psi(x, y, 0)$, omitting the third spatial dimension.

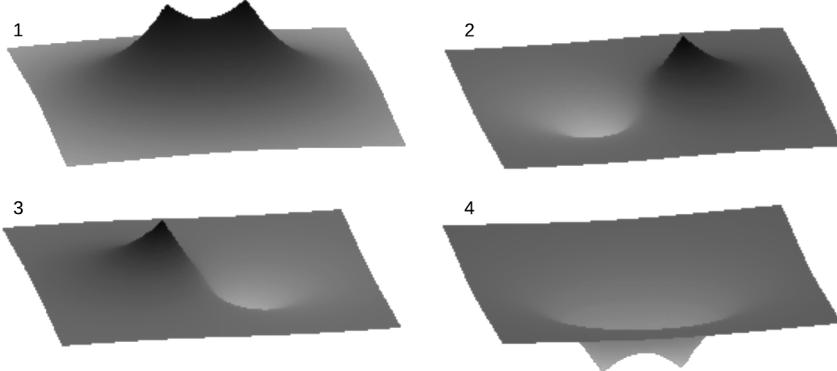
Wave phases in the hydrogen molecule

example 6

In example 6 on page 378, I argued that the existence of the H_2 molecule could essentially be explained by a particle-in-a-box argument: the molecule is a bigger box than an individual atom, so each electron's wavelength can be longer, its kinetic energy lower. Now that we're in possession of a mathematical expression for the wavefunction of the hydrogen atom in its ground state, we can make this argument a little more rigorous and detailed. Suppose that two hydrogen atoms are in a relatively cool sample of monoatomic hydrogen gas. Because the gas is cool, we can assume that the atoms are in their ground states. Now suppose that the two atoms approach one another. Making use again of the assumption that the gas is cool, it is reasonable to imagine that the atoms approach one another slowly. Now the atoms come a little closer, but still far enough apart that the region between them is classically forbidden. Each electron can tunnel through this classically forbidden region, but the tunneling probability is small. Each one is now found with, say, 99% probability in its original home, but with 1% probability in the other nucleus. Each electron is now in a state consisting of a superposition of the ground state of its own atom with the ground state of the other atom. There are two peaks in the superposed wavefunction, but one is a much bigger peak than the other.

An interesting question now arises. What are the relative phases of the two electrons? As discussed on page 370, the *absolute* phase of an electron's wavefunction is not really a meaningful concept. Suppose atom A contains electron Alice, and B electron Bob. Just before the collision, Alice may have wondered, "Is my phase positive right now, or is it negative? But of course I shouldn't ask myself such silly questions," she adds sheepishly.

h / Example 6.



But *relative* phases are well defined. As the two atoms draw closer and closer together, the tunneling probability rises, and eventually gets so high that each electron is spending essentially

50% of its time in each atom. It's now reasonable to imagine that either one of two possibilities could obtain. Alice's wavefunction could either look like $h/1$, with the two peaks in phase with one another, or it could look like $h/2$, with opposite phases. Because relative phases of wavefunctions are well defined, states 1 and 2 are physically distinguishable.² In particular, the kinetic energy of state 2 is much higher (by about 5 eV); roughly speaking, it is like the two-hump wave pattern of the particle in a box, as opposed to 1, which looks roughly like the one-hump pattern with a much longer wavelength. Not only that, but an electron in state 1 has a large probability of being found in the central region, where it has a large negative electrical energy due to its interaction with both protons. State 2, on the other hand, has a low probability of existing in that region. (This effect also equals about 5 eV.) Thus state 1 represents the true ground-state wavefunction of the H_2 molecule, and putting both Alice and Bob in that state results in a lower energy than their total energy when separated, so the molecule is bound, and will not fly apart spontaneously.

State $h/3$, on the other hand, is not physically distinguishable from $h/2$, nor is $h/4$ from $h/1$. Alice may say to Bob, "Isn't it wonderful that we're in state 1 or 4? I love being stable like this." But she knows it's not meaningful to ask herself at a given moment which state she's in, 1 or 4.

17.5 Use of complex numbers

In a classically forbidden region, a particle's total energy, $U + K$, is less than its U , so its K must be negative. If we want to keep believing in the equation $K = p^2/2m$, then apparently the momentum of the particle is the square root of a negative number. This is a symptom of the fact that the Schrödinger equation fails to describe all of nature unless the wavefunction and various other quantities are allowed to be complex numbers. In particular it is not possible to describe traveling waves correctly without using complex wavefunctions. Complex numbers were reviewed in subsection 5.7.2, p. 125.

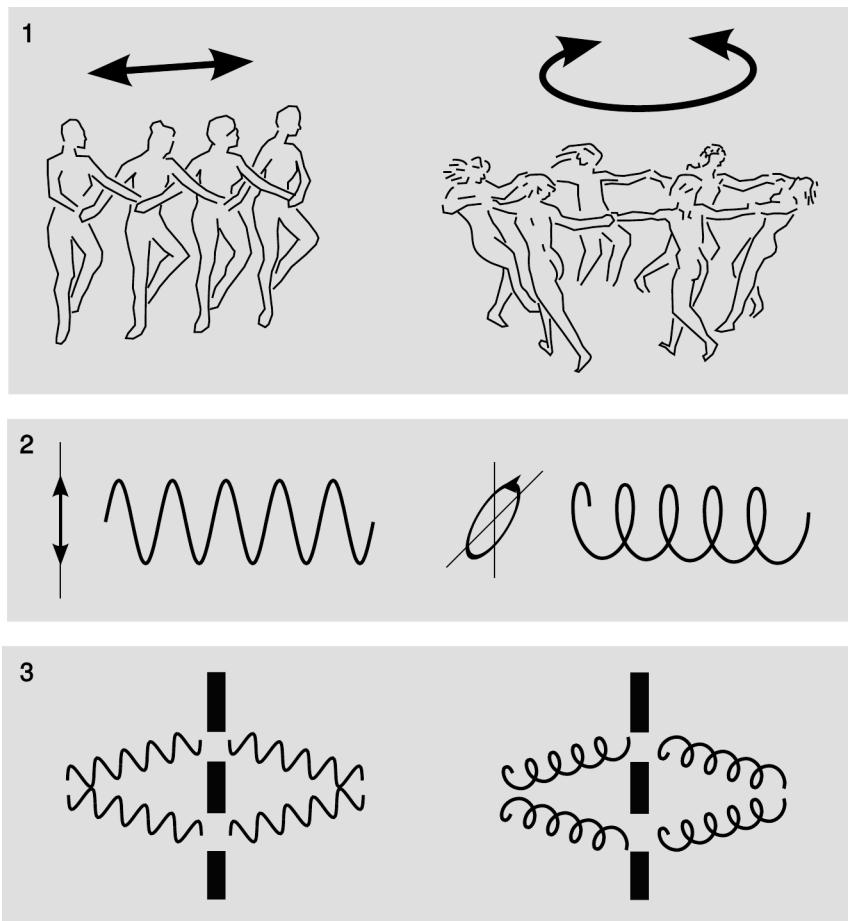
This may seem like nonsense, since real numbers are the only ones that are, well, real! Quantum mechanics can always be related to the real world, however, because its structure is such that the results of measurements always come out to be real numbers.

²The reader who has studied chemistry may find it helpful to make contact with the terminology and notation used by chemists. The state represented by pictures 1 and 4 is known as a σ orbital, which is a type of "bonding orbital." The state in 2 and 3 is a σ^* , a kind of "antibonding orbital." Note that although we will not discuss electron spin or the Pauli exclusion principle until sec. 18.5, p. 432, those considerations have no effect on this example, since the two electrons can have opposite spins.

i / 1. Oscillations can go back and forth, but it's also possible for them to move along a path that bites its own tail, like a circle. Photons act like one, electrons like the other.

2. Back-and-forth oscillations can naturally be described by a segment taken from the real number line, and we visualize the corresponding type of wave as a sine wave. Oscillations around a closed path relate more naturally to the complex number system. The complex number system has rotation built into its structure, e.g., the sequence $1, i, i^2, i^3, \dots$ rotates around the unit circle in 90-degree increments.

3. The double slit experiment embodies the one and only mystery of quantum physics. Either type of wave can undergo double-slit interference.



For example, we may describe an electron as having non-real momentum in classically forbidden regions, but its average momentum will always come out to be real (the imaginary parts average out to zero), and it can never transfer a non-real quantity of momentum to another particle.

A complete investigation of these issues is beyond the scope of this book, and this is why we have normally limited ourselves to standing waves, which can be described with real-valued wavefunctions. Figure i gives a visual depiction of the difference between real and complex wavefunctions. The following remarks may also be helpful.

Neither of the graphs in i/2 should be interpreted as a path traveled by something. This isn't anything mystical about quantum physics. It's just an ordinary fact about waves, which we first encountered in sec. 2.1, p. 32, where we saw the distinction between the motion of a wave and the motion of a wave pattern. In *both* examples in i/2, the wave pattern is moving in a straight line to the right.

The helical graph in i/2 shows a complex wavefunction whose value rotates around a circle in the complex plane with a frequency f

related to its energy by $E = hf$. As it does so, its squared magnitude $|\Psi|^2$ stays the same, so the corresponding probability stays constant. Which direction does it rotate? This direction is purely a matter of convention, since the distinction between the symbols i and $-i$ is arbitrary — both are equally valid as square roots of -1 . We can, for example, arbitrarily say that electrons with positive energies have wavefunctions whose phases rotate counterclockwise, and as long as we follow that rule consistently within a given calculation, everything will work. Note that it is not possible to define anything like a right-hand rule here, because the complex plane shown in the right-hand side of i/2 doesn't represent two dimensions of physical space; unlike a screw going into a piece of wood, an electron doesn't have a direction of rotation that depends on its direction of travel.

Superposition of complex wavefunctions *example 7*

- ▷ The right side of figure i/3 is a cartoonish representation of double-slit interference; it depicts the situation at the center, where symmetry guarantees that the interference is constructive. Suppose that at some off-center point, the two wavefunctions being superposed are $\Psi_1 = b$ and $\Psi_2 = bi$, where b is a real number with units. Compare the probability of finding the electron at this position with what it would have been if the superposition had been purely constructive, $b + b = 2b$.
- ▷ The probability per unit volume is proportional to the square of the magnitude of the total wavefunction, so we have

$$\frac{P_{\text{off center}}}{P_{\text{center}}} = \frac{|b + bi|^2}{|b + b|^2} = \frac{1^2 + 1^2}{2^2 + 0^2} = \frac{1}{2}.$$

Figure j shows a method for visualizing complex wavefunctions. The idea is to use colors to represent complex numbers, according to the arbitrary convention defined in figure j/1. Brightness indicates magnitude, and the rainbow hue shows the argument. Because this representation can't be understood in a black and white printed book, the figure is also reproduced on the back cover of printed copies. To avoid any confusion, note that the use of rainbow colors does not mean that we are representing actual visible light. In fact, we will be using these visual conventions to represent the wavefunctions of a material particle such as an electron. It is arbitrary that we use red for positive real numbers and blue-green for negative numbers, and that we pick a handedness for the diagram such that going from red toward yellow means going counterclockwise. Although physically the rainbow is a linear spectrum, we are not representing physical colors here, and we are exploiting the fact that the human brain tends to perceive color as a circle rather than a line, with violet and red being perceptually similar. One of the limitations of this representation is that brightness is limited, so we can't represent complex numbers with arbitrarily large magnitudes.

j / 1. A representation of complex numbers using color and brightness. 2. A wave traveling toward the right. 3. A wave traveling toward the left. 4. A standing wave formed by superposition of waves 2 and 3. 5. A two-dimensional standing wave. 6. A double-slit diffraction pattern.

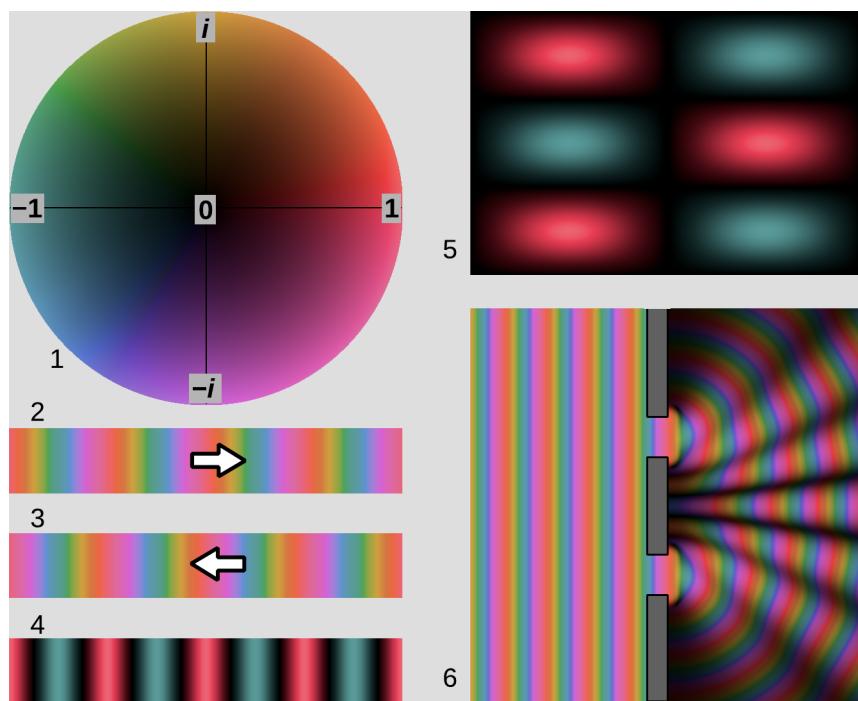


Figure j/2 shows a traveling wave as it propagates to the right. The standard convention in physics is that for a wave moving in a certain direction, the phase in the forward direction is farther counterclockwise in the complex plane, and you can verify for yourself that this is the case by comparing with the convention defined by j/1. The function being plotted here is $\Psi = e^{ikx}$, where $k = 2\pi/\lambda$ is the wavenumber. For the use of the complex exponential, see sec. 5.7.3, p .129; it simply represents a point on the unit circle in the complex plane. The wavelength λ is a constant and can be measured, for example, from one yellow point to the next. The wavelength is *not* different at different points on the figure, because we are using the colors merely as a visual encoding of the complex numbers — so, for example, a red point on the figure is not a point where the wave has a longer wavelength than it does at a blue point.

Figure j/3 represents a wave traveling to the left.

Figure j/4 shows a standing wave created by superimposing the traveling waves from j/2 and j/3, $\Psi_4 = (\Psi_2 + \Psi_3)/2$. (The reason for the factor of 2 is simply that otherwise some portions of Ψ_4 would have magnitudes too great to be represented using the available range of brightness.) All points on this wave have real values, represented by red and blue-green. We made the superposition real by an appropriate choice of the phases of Ψ_2 and Ψ_3 . This is always possible to do when we have a standing wave, but it is *only* possible for a standing wave, and this is the reason for all of the disclaimers in the captions of previous figures in which I took the liberty of

representing a traveling wave as a sine-wave graph.

Figure j/5 shows a two-dimensional standing wave of a particle in a box, and j/6 shows a double-slit interference pattern. (In the latter, I've cheated by making the amplitude of the wave on the right-hand half of the picture much greater than it would actually be.)

A paradox resolved

example 8

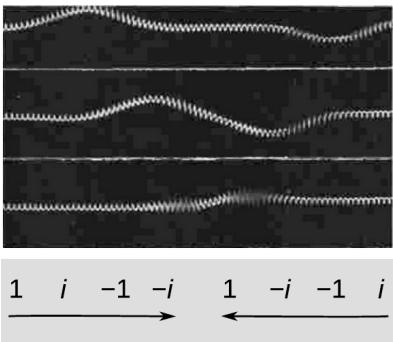
Consider the following paradox. Suppose we have an electron that is traveling wave, and its wavefunction looks like a wave-train consisting of 5 cycles of a sine wave. Call the distance between the leading and trailing edges of the wave-train L , so that $\lambda = L/5$. By sketching the wave, you can easily check that there are 11 points where its value equals zero. Therefore at a particular moment in time, there are 11 points where a detector has zero probability of detecting the electron.

But now consider how this would look in a frame of reference where the electron is moving more slowly, at one fifth of the speed we saw in the original frame. In this frame, L is the same, but λ is five times greater, because $\lambda = h/p$. Therefore in this frame we see only one cycle in the wave-train. Now there are only 3 points where the probability of detection is zero. But how can this be? All observers, regardless of their frames of reference, should agree on whether a particular detector detects the electron.

The resolution to this paradox is that it starts from the assumption that we can depict a traveling wave as a real-valued sine wave, which is zero in certain places. Actually, we can't. It has to be a complex number with a rotating phase angle in the complex plane, as in figure j/2, and a *constant* magnitude.

Discussion question

A The zero level of interaction energy U is arbitrary, e.g., it's equally valid to pick the zero of gravitational energy to be on the floor of your lab or at the ceiling. Suppose we're doing the double-slit experiment, i/3, with electrons. We define the zero-level of U so that the total energy $E = U + K$ of each electron is positive. and we observe a certain interference pattern like the one in figure i on p. 351. What happens if we then redefine the zero-level of U so that the electrons have $E < 0$?



B The top panel of the figure shows a series of snapshots in the motion of two pulses on a coil spring, one negative and one positive, as they move toward one another and superpose. The final image is very close to the moment at which the two pulses cancel completely. The following discussion is simpler if we consider infinite sine waves rather than pulses. How can the cancellation of two such mechanical waves be reconciled with conservation of energy? What about the case of colliding electromagnetic waves?

Quantum-mechanically, the issue isn't conservation of energy, it's conservation of probability, i.e., if there's initially a 100% probability that a particle exists somewhere, we don't want the probability to be more than or less than 100% at some later time. What happens when the colliding waves have real-valued wavefunctions Ψ ? Now consider the sketches of complex-valued wave pulses shown in the bottom panel of the figure as they are getting ready to collide.

17.6 Linearity of the Schrödinger equation

Some mathematical relationships and operations are *linear*, and some are not. For example, $2 \times (3+2)$ is the same as $2 \times 3 + 2 \times 2$, but $\sqrt{1+1} \neq \sqrt{1} + \sqrt{1}$. Differentiation is a linear operation, $(f+g)' = f' + g'$. The Schrödinger equation is built out of derivatives, so it is linear as well. That is, if Ψ_1 and Ψ_2 are both solutions of the Schrödinger equation, then so is $\Psi_1 + \Psi_2$. Linearity normally implies linearity with respect both to addition and to multiplication by a scalar. For example, if Ψ is a solution, then so is $\Psi + \Psi + \Psi$, which is the same as 3Ψ .

Linearity guarantees that the phase of a wavefunction makes no difference as to its validity as a solution to the Schrödinger equation. If $\sin kx$ is a solution, then so is the sine wave $-\sin kx$ with the opposite phase. This fact is logically interdependent with the fact that, as discussed on p. 370, the phase of a wavefunction is unobservable. For measuring devices and humans are material objects that can be described by wavefunctions. So suppose, for example, that we flip the phase of all the particles inside the entire laboratory. By linearity, the evolution of this measurement process is still a valid solution of the Schrödinger equation.

The Schrödinger equation is a wave equation, and its linearity implies that the waves obey the principle of superposition. In most cases in nature, we find that the principle of superposition for waves is at best an approximation. For example, if the amplitude of a tsunami is so huge that the trough of the wave reaches all the way down to the ocean floor, exposing the rocks and sand as it passes overhead, then clearly there is no way to double the amplitude of the wave and still get something that obeys the laws of physics. Even at less extreme amplitudes, superposition is only an approximation for water waves, and so for example it is only approximately true that when two sets of ripples intersect on the surface of a pond, they

pass through without “seeing” each other.

It is therefore natural to ask whether the apparent linearity of the Schrödinger equation is only an approximation to some more precise, nonlinear theory. This is not currently believed to be the case. If we are to make sense of Schrödinger’s cat (sec. 15.4, p. 358), then the experimenter who sees a live cat and the one who sees a dead cat must remain oblivious to their other selves, like the ripples on the pond that intersect without “seeing” each other. Attempts to create slightly nonlinear versions of standard quantum mechanics have been shown to have implausible physical properties, such as allowing the propagation of signals faster than c . (This is known as Gisin’s theorem. The original paper, “Weinberg’s non-linear quantum mechanics and supraluminal communications,” is surprisingly readable and nonmathematical.)

The linearity of the Schrödinger equation is what allows us to talk about its solutions as vectors in a vector space (p. 155). For example, if Ψ_1 represents an unstable nucleus that has not yet gamma decayed, and Ψ_2 is its state after the decay, then any superposition $\alpha\Psi_1 + \beta\Psi_2$, with real or complex coefficients α and β , is a possible wavefunction, and we can notate this as a vector, $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, in a two-dimensional vector space.

People wrestling with the weirdness of Schrödinger’s cat sometimes say that it’s impossible to have a superposition of the live cat and the dead cat. Not true. The linearity of the Schrödinger equation guarantees that we can take any wavefunctions Ψ_1 and Ψ_2 and superpose them, and this is exactly what happens, according to quantum physics, when we do the experiment. What *is* impossible, for all practical purposes, is to observe any type of interaction between the live and dead cats, such as wave interference, which we estimated on p. 369 and found to be undetectable, due to the extremely small wavelength.

17.7 The inner product and observables

17.7.1 The inner product

In sec. 6.6, p. 156, we discussed the idea of an inner product, which is an operation on a vector space that acts like the dot product, i.e., it takes two vectors as inputs and gives back a scalar as an output. We also saw that any time we had a way of measuring magnitudes, we automatically got an inner product for free. In quantum mechanics, we do have a way of measuring magnitudes, which is the total probability of a given wavefunction. We even have a convenient notation for this, the bra-ket notation introduced in sec. 16.1.3, p. 372. In example 3, we had a wavefunction $|\curvearrowleft\rangle$, and the fact of its being normalized was written as $\langle\curvearrowleft|\curvearrowleft\rangle = 1$.

$$\begin{aligned}\langle u|v\rangle &= \langle v|u\rangle^* \\ \langle u|\alpha v + \beta w\rangle &= \alpha\langle u|v\rangle + \beta\langle u|w\rangle \\ \langle \alpha u + \beta v|w\rangle &= \alpha^*\langle u|w\rangle + \beta^*\langle v|w\rangle.\end{aligned}$$

k / Properties of the inner product.
Except for the complex conjugates, these are the same as the properties of the dot product from Euclidean geometry.

This is a statement that the total probability for *something* to happen must be 1, but we can also think of it as a statement that the “magnitude” of $\sqrt{\cdot}$ has a certain value. Because we have a way of defining magnitudes of wavefunctions, we automatically get an inner product.

If we’re dealing with wavefunctions that are expressed as functions of position, then it’s pretty clear how to define an appropriate inner product: $\langle u|v\rangle = \int u^*v dx$. We need to use the complex conjugate u^* rather than just u , for the following reason. When we take the inner product of a wavefunction with itself, it has a probability interpretation. Probabilities are always real and positive. So we need to set things up in such a way that $\langle u|u\rangle$ is always real and positive. When we take $\langle u|u\rangle$, the thing inside the integral is u^*u , which is the same as $|u|^2$, and this is indeed real and positive, so when we integrate it we get an answer that is real and positive.

Note the similarity between the expression $\int u^*v dx$ and the expression $u_x v_x + u_y v_y + u_z v_z$ for a dot product: the integral is a continuous sum, and the dot product is a discrete sum.

We recall from our previous discussion (p. 156) that an inner product can generally be described as a measure of how similar two vectors are. For example, in the Euclidean plane, vectors that are perpendicular to each other have a dot product of zero, which tells you that they lie along lines that are completely different. When two vectors have an inner product of zero, we say that they are *orthogonal*.

Physically, two wavefunctions have a zero inner product if and only if they are completely distinguishable from each other by the measurement of some observable. By analogy with vectors in Euclidean space, we say that the two wavefunctions are orthogonal. For example, $\langle \curvearrowleft | \curvearrowright \rangle = 0$, as can be verified from the integral $\int_0^\pi \sin x \sin 2x dx = 0$. These states are also distinguishable by measuring either their momentum or their energy.

Suppose that u and v are both properly normalized wavefunctions. If $|\langle u|v\rangle| = 1$, then the states are identical.³ If $\langle u|v\rangle = 0$, then u and v are completely distinguishable from one another. There is also the intermediate case where $\langle u|v\rangle$ has a magnitude greater than 0 but less than 1. In this case, we could say that u is a mixture of v plus some other state w that is distinguishable from v , i.e., that

$$|u\rangle = \alpha|v\rangle + \beta|w\rangle.$$

where $\langle v|w\rangle = 0$. We then have

$$\langle u|v\rangle = (\alpha\langle v| + \beta\langle w|)|v\rangle = \alpha.$$

³If the inner product is, for example, -1 , then the wavefunctions differ only by an unobservable difference in phase, so they really describe the same state.

$$\langle \curvearrowleft | \curvearrowright \rangle = 1$$

The wave \curvearrowright is properly normalized.

$$\langle \curvearrowleft | \curvearrowright \rangle = 0$$

The waves \curvearrowleft and \curvearrowright are perfectly distinguishable.

$$\langle \curvearrowleft | \curvearrowleft \rangle = -0.81$$

The wave \curvearrowleft can be expressed as $-0.81\curvearrowright$ plus distinguishable waves. Or: measurements have probability $(-0.81)^2 \approx 0.66$ of saying one of these waves is the same as the other.

Now suppose that we make measurements capable of determining whether or not the system is in the state v . If the system is prepared in state u , and we make these measurements on it, then by the linearity of the Schrödinger equation, the result is that the measuring apparatus or observer ends up in a Schrödinger's-cat state that looks like

$$\alpha|\text{observed } v\rangle + \beta|\text{observed } w\rangle.$$

We interpret squares of amplitudes as probabilities, so

$$P = |\alpha|^2 = |\langle u|v\rangle|^2$$

gives us the probability that we will have observed the state to be v . This final leap in the logic, to a probability interpretation, has felt mysterious to several generations of physicists, but recent work has clarified the situation somewhat.

17.7.2 Observables

When I was in college, my stepmother attempted to make me more hip, more artistic, and less uncool. I would go over for a visit sometimes on the weekend, and she would sit me down, pour us each a glass of wine, and draw me out. In one of these conversations, I insisted that love didn't exist, because there was no way to measure it. The physicist's way of thinking, which by that time I had already started to adopt, is that we should distinguish carefully between things that are *observables* and things that aren't. One of the basic principles of quantum mechanics is that $\langle u|v\rangle = 0$ if and only if there is some observable that perfectly distinguishes state u from state v . (We may actually need more than one observable.)

This makes it important to be clear on what really is an observable and what isn't. Informally, what we mean by an observable is that:

1. we can tell by looking at the state what value it has, or at least assign probabilities to values;

2. it has a single value; and
3. (optionally) it has some counterpart in classical physics.

The following are all observables:

- position
- momentum
- energy
- angular momentum (ch. 18).

The following are not:

- phase and normalization (sec. 16.1.4)
- time (because many systems, such as an atom, are too simple to act as clocks).

A couple of interesting borderline cases are wavelength and angle. Classically, an electron doesn't have a wavelength, so we would usually choose to talk about its momentum as an observable, not its wavelength (although they relate via $p = h/\lambda$, so it really doesn't make much difference). For an electron in an atom, thinking in a simplified picture in two dimensions, we could talk about its angle in the plane, but this causes problems because angles are not really single-valued, e.g., 0 is the same as 360°.

The fact that position is an observable but time is not is one of the things that makes it difficult to reconcile quantum mechanics with relativity, which considers time to be just another dimension. This difficulty has been reconciled for special relativity, but not for general relativity.

For those with some background in linear algebra, it may be helpful to connect this to the idea of eigenvalues and eigenvectors. In quantum mechanics, we associate with an observable such as momentum p some linear operator \mathcal{O}_p . In this example, \mathcal{O}_p is basically the derivative operator (multiplied by some constant factors). If a particle is in a state of definite momentum, then it is an eigenvector of \mathcal{O}_p with eigenvalue p . The basic motivation for using this style of definition is that if Ψ is an eigenvector, then $c\Psi$ is also an eigenvector, with the same eigenvalue; therefore this mathematical setup automatically keeps phase and normalization from being observables.

17.8 Time evolution and unitarity

17.8.1 The simplest cases of time evolution

So far we have not said too much about how a wavefunction changes with time, except that a state of definite energy E has some frequency given by $E = \hbar\omega$. As a simple example, consider a particle in a box. Let's say that it's initially in the ground state, which is a standing wave like \curvearrowleft . It has some energy, so we can determine its frequency and period. Based on experience with standing waves on a string, we expect that after half a period it will look like \curvearrowright . The problem is how it gets from the frown to the smile. If it behaved like a wave on a string, it would go from \curvearrowleft to — to \curvearrowright . But passing through the flat intermediate state won't work, because of the fundamental structure of quantum mechanics, which includes the principle of state fundamentalism (p. 334). The state, represented by our picture of the wavefunction, is supposed to be all there is to know about the system. But if someone presented us with the zero wavefunction — , we would have *no* information at all about the state.

So the wavefunction's amplitude can't oscillate back and forth along the real number line between positive and negative values, as it would for a wave on a string. What it actually does is to spin around in a circle in the complex plane. That is, instead of going like

$$1 \times \curvearrowleft \quad 0 \times \curvearrowleft \quad -1 \times \curvearrowleft, \quad [\text{wrong}]$$

it goes like

$$1 \times \curvearrowleft \quad -i \times \curvearrowleft \quad -1 \times \curvearrowleft. \quad [\text{right}]$$

(We have $-i$ here rather than i because the convention is to have it spin clockwise.) This is enough to define the time-dependence of any state that has a definite energy: the energy tells us a frequency ω , and then we know that the time-evolution of the state is simply that it spins its phase like $e^{-i\omega t}$, as in Euler's formula (sec. 5.7.3, p. 129).

Another simple example is a traveling wave with a definite energy, figure m. Frozen at one moment in time, such a wave looks like e^{ikx} , which we visualize as a repeating rainbow. If we let this wave travel, then as we saw in ch. 2, its rigidly gliding motion should be described by a function whose input is of the form $kx - \omega t$. In other words, we have $\Psi = e^{i(kx - \omega t)}$. An observer who is watching the wave go by will say that a certain point with a fixed phase, say $\Psi = 1$, which is red in the figure, is moving to the right at a certain speed. But an observer who just stays in one place as the wave washes over her will say that Ψ is spinning around in a circle in the complex plane. To see this more explicitly, we can break up the exponential into separate factors, $\Psi = e^{ikx}e^{-i\omega t}$. The time-dependent factor of

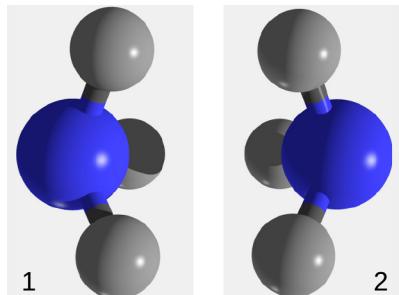


m / A quantum-mechanical traveling wave with a definite energy, and therefore a definite frequency. Complex values are represented as colors, according to the conventions introduced in figure j, p. 406. The wave should actually be thought of as extending infinitely far in both directions, but I've drawn a finite piece of it in order to make it more visually obvious that it's traveling to the right.

$e^{-i\omega t}$ is just what we described a moment ago: a phase spinning clockwise in the complex plane.

17.8.2 The two-state system

As the next step up in sophistication we could consider a system that has two possible states. A simple example is the ammonia molecule in figure n. If the molecule is isolated and has no angular momentum, then it can never get from one orientation to the other by rotating. It can, however, flip from one to the other by turning itself inside out like an umbrella in a strong wind. Classically, if we initially put the molecule in one orientation, then it wouldn't have enough energy to get through the intermediate flat configuration. But quantum-mechanically it can get through by tunneling. Therefore if we prepare the system in a pure $|1\rangle$ state at some time, we expect that in the future it will be in some mixture $a_1|1\rangle + a_2|2\rangle$, and if we check it there will be some probability $|a_2|^2$ of finding that it's switched.



n / The ammonia molecule, in states that are inverted relative to one another.

Now this implies that $|1\rangle$ and $|2\rangle$ cannot be states of definite energy, because if they were, then after we prepared the system in an initial state with $|a_1| = 1$ and $|a_2| = 0$, all that would happen would be that the complex number a_1 would spin its phase clockwise around the unit circle at some frequency given by $E = \hbar\omega$. The probability $|a_1|^2$ would stay the same, and indeed every observable of the system would stay the same, because absolute phases don't mean anything. This is not the case — we expect by the totalitarian principle (p. 334) that the system *can* tunnel from 1 to 2. Therefore $|1\rangle$ and $|2\rangle$ must *not* be states of definite energy. The actual states of definite energy are standing wave patterns that look like $\sim\!\sim$ (the ground state) and $\sim\!\sim$ (the excited state). Because these states have slightly different energies, they oscillate at slightly different frequencies. If we initially put the system in the state $|\sim\!\sim\rangle + |\sim\!\sim\rangle$, then there is constructive interference on the left and destructive interference on the right. This is what it would mean to prepare the molecule in the state $|1\rangle$. State $|1\rangle$ is this superposition of the two energy states. Because the frequencies are unequal, eventually the two waves will flip their relative phases, and we will have state $|2\rangle$. This is in fact the most general possible behavior for *any* quantum-mechanical system with two states: it can at most just oscillate between them. The oscillation can be complete (like the ammonia's $1 \leftrightarrow 2$), nonexistent (if the system is in a state of definite energy), or partial (if the system is in an unequal mixture of two energy states).

17.8.3 The time-dependent Schrödinger equation

When a system has not just one or two but many possible states, the time-evolution is given by an equation known as the time-dependent Schrödinger equation, which we will not write down explicitly here. For example, a nonrelativistic electron in free space

has infinitely many possible states, since there are no boundary conditions constraining its wave pattern. In this example the time-dependent Schrödinger equation becomes a certain wave equation, which is dispersive (sec. 16.2, p. 373), i.e., different wavelengths move at different speeds. It's not hard to see why this is, since $p = h/\lambda$, and $p = mv$ for nonrelativistic motion.⁴

For standing waves, the time-dependent Schrödinger equation is equivalent to the simpler time-independent version. Both versions are perfectly linear (sec. 17.6, p. 408).

17.8.4 Unitarity

In discussion question B on p. 408, we considered two traveling waves that collided head-on and superposed, and we convinced ourselves that probability would be conserved. It's possible to prove that the time-evolution of the wavefunction always results in conservation of probability.

To put this in real-world terms, suppose that your history teacher calls on you in class and asks you what happened on December 7, 1941. With a straight face, you answer, “Professor, I can guarantee that *something* happened on that day.” “That's correct. And what about August 6, 1945?” “Yes, something also happened on that day.” In other words, if we have a properly normalized wavefunction at a certain time, then we expect it to remain properly normalized at all later times.

The time evolution is also completely deterministic, so that if we know Ψ initially, we can always predict it in the future. We can also “predict” backward in time, so that the system's history can always be recovered from knowledge of its present state. Thus there is never any loss of information over time.

Summarizing, we have the following important principle:

Unitary evolution of the wavefunction

The wavefunction evolves over time in a deterministic and *unitary* manner, meaning that probability is conserved and information is never lost.

The word “unitary” is defined more precisely in linear algebra [Z420](#).

Since we think of quantum mechanics as being all about randomness, this determinism may seem surprising. But determinism in the time-evolution of the wavefunction isn't the same as determinism in the results of experiments as perceived and recorded by a human brain. Suppose that you prepare a uranium atom in its ground state, then wait one half-life and observe whether or not

⁴But as discussed in sec. 16.2, it is necessary to distinguish phase velocity from group velocity.

it has decayed, as in the thought experiment of Schrödinger's cat (p. 358). There is no uncertainty or randomness about the wavefunction of the whole system (atom plus you) at the end. We know for sure what it looks like. It consists of an equal superposition of two states, one in which the atom has decayed and your brain has observed that fact, and one in which the atom has not yet decayed and that fact is instead recorded in your brain.

As a possible example of a violation of unitarity, in an exotic context, consider the disappearance of matter into a black hole. If I throw my secret teenage diary into a black hole, then it contributes a little bit to the black hole's mass, but the embarrassing information on the pages is lost forever. This loss of information seems to imply nonunitarity. This is one of several arguments suggesting that quantum mechanics cannot fully handle the gravitational force. Thus although physicists currently seem to possess a completely successful theory of gravity (Einstein's theory of general relativity) and a completely successful theory of the microscopic world (quantum mechanics), the two theories are irreconcilable, and we can only make educated guesses, for example, about the behavior of a hypothetical microscopic black hole.

A standing wave

example 9

As discussed on p. 413, a standing wave \curvearrowleft will evolve, after a quarter of a period, into $-i \times \curvearrowleft$. The probability density depends only on the squared *magnitude* of the wavefunction, and is therefore unchanged over this evolution. Thus if the initial wave was normalized, then so is the final one, since multiplying a complex number by $-i$ doesn't change its magnitude, only its argument (phase). This is consistent with unitarity.

Unitarity also requires that information is never lost, and we can verify this. Suppose that we see the wavefunction $-i \times \curvearrowleft$ at the later time. We can then infer what the wavefunction must have been a quarter-period earlier: it must have been \curvearrowleft .

Any state of definite energy

example 10

The reasoning of example 9 immediately carries over to any state of definite energy and any time interval, since the time evolution just amounts to spinning the phase in the complex plane at frequency ω , but this has no effect on any probabilities.

A mixture of energies

example 11

Suppose we have a particle in a box with the initial wavefunction

$$\curvearrowleft + \curvearrowright.$$

The momentum of the short-wavelength part is twice the momentum of the long-wavelength one, so assuming the motion is non-relativistic, the two energies differ by a factor of four. Therefore this is not a state of definite energy.

The energies of the two parts differ by a factor of four, so by $E =$

$\hbar\omega$ so do their frequencies. As an example, let's consider the evolution of this state over a time equal to half the period of the low-energy part, which is two full periods of the high-energy part. After this time, we have

$$\curvearrowleft + \curvearrowright.$$

A brute-force verification of the conservation of probability is possible, but involves a somewhat messy and tedious calculation that $\int_0^\pi [A \sin x + \sin 2x]^2 dx$ is the same as $\int_0^\pi [-A \sin x + \sin 2x]^2 dx$. (If you graph the two functions in a utility such as the online app desmos, you should be able to see easily that this holds visually.)

A much quicker and easier method is to use the linearity of the inner product. For the total initial probability, we have

$$\langle \curvearrowleft + \curvearrowright | \curvearrowleft + \curvearrowright \rangle = \langle \curvearrowleft | \curvearrowleft \rangle + \langle \curvearrowright | \curvearrowright \rangle + \langle \curvearrowleft | \curvearrowright \rangle + \langle \curvearrowright | \curvearrowleft \rangle.$$

But since the states \curvearrowleft and \curvearrowright are perfectly distinguishable by measurement of an observable (energy), it follows that $\langle \curvearrowleft | \curvearrowright \rangle = 0$ and $\langle \curvearrowright | \curvearrowleft \rangle = 0$. Therefore the total initial probability simplifies to the expression

$$\langle \curvearrowleft | \curvearrowleft \rangle + \langle \curvearrowright | \curvearrowright \rangle.$$

This has a simple interpretation: if we measure the energy, then the first term is the probability that we'll find the lower energy, and the second one is the probability of the higher energy. If we assume that the state initially had the right normalization, then the two terms in the expression add up to 1.

If we now go through similar reasoning for the wavefunction at the later time, we get

$$\langle \curvearrowleft | \curvearrowleft \rangle + \langle \curvearrowright | \curvearrowright \rangle.$$

But by linearity the first term can be massaged into the form $\langle -\curvearrowleft | -\curvearrowleft \rangle = (-1)^2 \langle \curvearrowleft | \curvearrowleft \rangle = \langle \curvearrowleft | \curvearrowleft \rangle$, as before. Probability has been conserved, as required by unitarity.

Because example 11 didn't appeal to any specific properties of the states \curvearrowleft and \curvearrowright , it can easily be extended to a proof of unitarity for a very large class of states: anything that we can form by mixing a bunch of states of definite energy.

Radioactive decay

example 12

In sec. 14.5, p. 333, I argued that the exponential decay law for radioactivity followed from very general ideas about quantum mechanics. The argument was that an unstable nucleus would be likely to be found in its ground state, and since the ground state is a single state, it has no way of retaining any memory. Therefore, if it's survived up until a certain time t , then its probability of surviving for some additional time interval dt can't depend on t . We

now have enough of a picture of quantum mechanics to allow a more detailed understanding of this point.

An initial stab at the problem shows that there is more subtlety to it than might have been imagined. It seems natural to describe the system in terms of two states: an undecayed state and a decayed one. But as we saw in sec. 17.8.2, p. 414, exponential decay is not even a possibility for a two-state system in quantum mechanics: the most general possible behavior for such a system is an oscillation back and forth between the two states.

Of course when an atom emits a photon or a nucleus undergoes radioactive decay, there are not just two states but infinitely many. For example, if a nucleus alpha decays, the energy of the alpha is fixed by conservation of energy, but its momentum vector \mathbf{p} can point in any one of an infinite number of directions.

Let's assume that we can describe the initial, undecayed nucleus as a *single state* $|0\rangle$, its ground state. We label each possible final state, which describes both the state of the daughter nucleus and the state of the alpha, by the alpha particle's momentum, $|\mathbf{p}\rangle$. If the nucleus is initially in the state

$$|0\rangle,$$

then because alpha decay *is* possible (the totalitarian principle again), after some time interval Δt the system will be in a mixture of decayed and undecayed states,

$$a|0\rangle + \dots,$$

where a is some number and \dots denotes an infinite sum over the various $|\mathbf{p}\rangle$ states. The probability of decay after this amount of time is $|a|^2$, and unitarity requires that $|a| < 1$. What will happen after a second time interval Δt ? Quantum mechanics is perfectly linear, so the rule for carrying the state forward in time by Δt must be a linear function. Let's call this function L , so that $L(|0\rangle) = a|0\rangle + \dots$ Now suppose as well that alpha decay is *irreversible*. After all, it seems unlikely that the alpha would bounce off of something, come back, and rejoin the daughter nucleus to rebuild the original parent. Let's see what happens when we do a second time step by computing $L(L(|0\rangle))$. By linearity, we get $L(a|0\rangle + \dots) = aL(a|0\rangle) + L(\dots)$, or

$$a^2|0\rangle + \dots$$

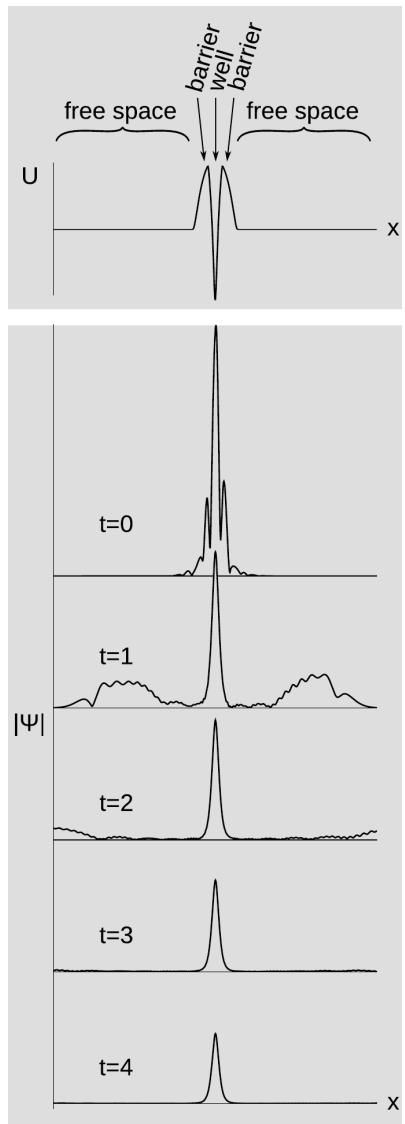
Here the new \dots is different from the old one, and we've made use of the assumption of irreversibility by assuming that $L(\dots)$ doesn't contribute to the amplitude of $|0\rangle$. We have exponential decay, with the decay probability going like 1, $|a|^2$, $|a|^4$, \dots This result depended on two assumptions: (1) that it was valid to measure

the undecayed state as a single state, and (2) that the decay was irreversible.

There is a further interesting application of unitarity here, which is that exponential decay *cannot* be exact at all times. For if it were, then we could extrapolate *back* in time to before the start, and we would find that the decay probability at $t = -\Delta t$, one clock tick before the start, was $|a|^{-2}$. This would be greater than 1, which would violate unitarity. What's going on here? In reality, the original nucleus had to be formed somehow, probably in nuclear reactions inside a supernova billions of years ago. That process of formation was not clean and tidy, and the nucleus only later settled down into the behavior described above.

Figure o shows an example of what happens when we start off with this kind of messy state. This is a computer simulation of the time-dependent Schrödinger equation for a particle in one dimension, which we take as a toy model. The particle is like the alpha particle, and the potential U shown in the graph represents the effect of the attractive and repulsive forces of the rest of the nucleus.

The initial state at $t = 0$ is not a state of definite energy — a state of definite energy is one that doesn't actually physically evolve, it just spins its phase. We can see that this state looks like a noisy mixture of a bunch of different wavelengths. The shorter wavelengths have higher energies. As soon as we start the simulation, a bunch of these waves fly off in both directions. We see them escaping at $t = 1$. But the lower-energy waves are too low in energy to be classically allowed to escape over the barriers. They can only escape by tunneling, which is a slower process. By $t = 1$, the shape of the part of the wave remaining inside the well has turned into a bell shape. This bell shape is what we were talking about before when we wrote $|0\rangle$. At $t = 2$ the rate of decay sensed by a detector near the edge of the graph is still fluctuating a lot. But by $t = 3$ the system has settled down to a steady process of decay in which the system acts as we assumed in our simplified model: the undecayed part is well modeled by a single state, and the decay is irreversible. (Beyond this time, the outgoing waves get too small to see easily on the graphs, but they still exist.)



o / A numerical simulation giving the escape of a particle out of a well surrounded by a barrier.

Notes for chapter 17

[2415](#) Precise mathematical definition of unitarity

Linear algebra defines unitarity as follows. A unitary transformation is one that preserves inner products. That is, \mathcal{O} is unitary if $\langle \mathcal{O}u | \mathcal{O}v \rangle = \langle u | v \rangle$. This is similar to the way in which rotations preserve dot products in Euclidean geometry. You may have heard about the idea of an orthogonal matrix in a math course, with the classic example being a rotation matrix in two or three dimensions. A unitary matrix is the same concept, but generalized to linear algebra over the complex numbers. The linearity of the Schrödinger equation guarantees that the evolution of the wavefunction from one time to another time can be represented by a linear operator \mathcal{O} . We require this to be unitary, and it is unitary for the Schrödinger equation.

Problems

Key

- ✓ A computerized answer check is available online.
- * A difficult problem.

1 (a) A distance scale is shown below the wavefunction illustrated in figure g on page 401. Compare this with the order-of-magnitude estimate derived in section 16.6.2, p. 385, for the radius r at which the wavefunction begins tailing off. Was the estimate on the right order of magnitude?

(b) Although we normally say the moon orbits the earth, actually they both orbit around their common center of mass, which is below the earth's surface but not at its center. The same is true of the hydrogen atom. Does the center of mass lie inside the proton, or outside it?

2 The wavefunction of the electron in the ground state of a hydrogen atom, shown in the top left of figure j on p. 436, is

$$\Psi = \pi^{-1/2} a^{-3/2} e^{-r/a},$$

where r is the distance from the proton, and $a = \hbar^2/kme^2 = 5.3 \times 10^{-11}$ m is a constant that sets the size of the wave. The figure doesn't show the proton; let's take the proton to be a sphere with a radius of $b = 0.5$ fm.

- (a) Reproduce figure j in a rough sketch, and indicate, relative to the size of your sketch, some idea of how big a and b are.
- (b) Calculate symbolically, without plugging in numbers, the probability that at any moment, the electron is inside the proton. [Hint: Does it matter if you plug in $r = 0$ or $r = b$ in the equation for the wavefunction?] ✓
- (c) Calculate the probability numerically. ✓
- (d) Based on the equation for the wavefunction, is it valid to think of a hydrogen atom as having a finite size? Can a be interpreted as the size of the atom, beyond which there is nothing? Or is there any limit on how far the electron can be from the proton?

3 Show that the wavefunction given in problem 2 is properly normalized.

4 In classical mechanics, an interaction energy of the form $U(x) = \frac{1}{2}kx^2$ gives a harmonic oscillator: the particle moves back and forth at a frequency $\omega = \sqrt{k/m}$. This form for $U(x)$ is often a good approximation for an individual atom in a solid, which can vibrate around its equilibrium position at $x = 0$. (For simplicity, we restrict our treatment to one dimension, and we treat the atom as a single particle rather than as a nucleus surrounded by electrons). The atom, however, should be treated quantum-mechanically, not classically. It will have a wave function. We expect this wave function to have one or more peaks in the classically allowed region, and we expect it to tail off in the classically forbidden regions to the right and left. Since the shape of $U(x)$ is a parabola, not a series of flat steps as in figure d on page 397, the wavy part in the middle will not be a sine wave, and the tails will not be exponentials.

(a) Show that there is a solution to the Schrödinger equation of the form

$$\Psi(x) = e^{-bx^2},$$

and relate b to k , m , and \hbar . To do this, calculate the second derivative, plug the result into the Schrödinger equation, and then find what value of b would make the equation valid for *all* values of x . This wavefunction turns out to be the ground state. Note that this wavefunction is not properly normalized — don't worry about that.

(b) Sketch a graph showing what this wavefunction looks like.

(c) Let's interpret b . If you changed b , how would the wavefunction look different? Demonstrate by sketching two graphs, one for a smaller value of b , and one for a larger value.

(d) Making k greater means making the atom more tightly bound. Mathematically, what happens to the value of b in your result from part a if you make k greater? Does this make sense physically when you compare with part c? ✓

5 Show that a wavefunction of the form $\Psi = e^{by} \sin ax$ is a possible solution of the Schrödinger equation in two dimensions, with a constant potential U . Can we tell whether it would apply to a classically allowed region, or a classically forbidden one?

6 *This problem generalizes the one-dimensional result from problem 16-5, p. 389.*

Find the energy levels of a particle in a three-dimensional rectangular box with sides of length a , b , and c . ✓

7 Americium-241 is an artificial isotope used in smoke detectors. It undergoes alpha decay, with a half-life of 432 years. As discussed in example 1 on page 397, alpha decay can be understood as a tunneling process, and although the barrier is not rectangular in shape, the equation for the tunneling probability on page 397 can still be used as a rough guide to our thinking. For americium-241, the tunneling probability is about 1×10^{-29} . Suppose that this nucleus were to decay by emitting a helium-3 nucleus instead of an alpha particle (helium-4). Estimate the relevant tunneling probability, assuming that the total energy E remains the same. This higher probability is contrary to the empirical observation that this nucleus is not observed to decay by ${}^3\text{He}$ emission with any significant probability, and in general ${}^3\text{He}$ emission is almost unknown in nature; this is mainly because the ${}^3\text{He}$ nucleus is far less stable than the helium-4 nucleus, and the difference in binding energy reduces the energy available for the decay.

8 The wavefunction Ψ of an electron is a complex number. Make up an example of a value for the wavefunction that is not a real number, and consider the following expressions: Ψ^2 , $|\Psi|^2$, $|\Psi^2|$. Which of these would it make sense to interpret as a probability density? All of them? Some? Only one? \triangleright Solution, p. 454

9 In problem 4 on p. 422, you showed that a wavefunction of the form

$$\Psi_0(x) = e^{-x^2/2}$$

was a solution of the Schrödinger equation for the quantum harmonic oscillator in one dimensions. (We ignore units, and the factor of 1/2 in the exponent is just a convention.) It represents the ground state. The wavefunction of the first excited state is

$$\Psi_1(x) = xe^{-x^2/2},$$

with the same value of b .

- Show that these states are orthogonal in the sense defined on p. 410.
- What is an observable that would distinguish them?

10 Consider the wavefunctions $\Psi_1 = \nearrow$ and $\Psi_2 = \nwarrow$ for a particle in a one-dimensional box. Suppose we have the superposition $\Psi = A(2\Psi_1 + \Psi_2)$.

- (a) If Ψ is to be properly normalized, what is $|A|?$ ✓
- (b) Sketch the wavefunction.
- (c) Suppose you can measure the position of the particle very accurately. What is the probability that the particle will be found in the left half of the box? ✓
- (d) Instead of measuring position, suppose you measure the energy of the state. What is the probability that you'll measure the ground state energy? ✓
- (e) Suppose that the wavefunction had been $\Phi = A(2\Psi_1 - \Psi_2)$. Which of your answers to parts a-d would remain the same, and which would change? (You need not redo the work for the ones that would change. Just give your reasoning as to whether they would or would not.) [Problem by B. Shotwell.]

Chapter 18

Quantization of angular momentum

18.1 Quantization of angular momentum

Angular momentum is quantized in quantum physics. As an example, consider a quantum wave-particle confined to a circle, like a wave in a circular moat surrounding a castle. A sine wave in such a “quantum moat” cannot have any old wavelength, because an integer number of wavelengths must fit around the circumference, C , of the moat. The larger this integer is, the shorter the wavelength, and a shorter wavelength relates to greater momentum and angular momentum. Since this integer is related to angular momentum, we use the symbol ℓ for it:

$$\lambda = C/\ell.$$

The angular momentum is

$$L = rp.$$

Here, $r = C/2\pi$, and $p = h/\lambda = h\ell/C$, so

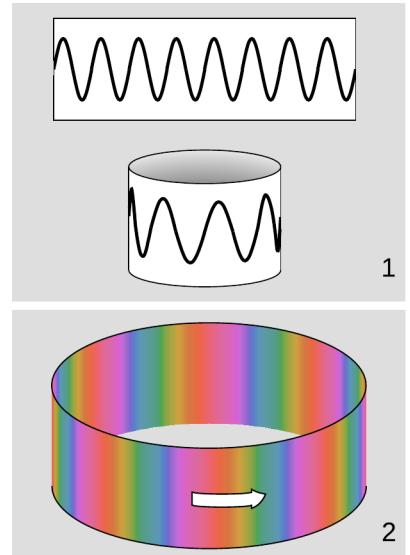
$$\begin{aligned} L &= \frac{C}{2\pi} \cdot \frac{h\ell}{C} \\ &= \frac{h}{2\pi}\ell. \end{aligned}$$

In the example of the quantum moat, angular momentum is quantized in units of $h/2\pi$. This makes $h/2\pi$ a pretty important number, so we define the abbreviation $\hbar = h/2\pi$. This symbol is read “h-bar.”

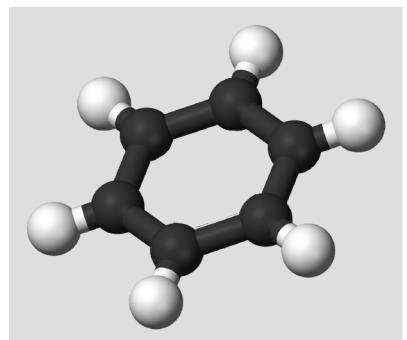
In fact, this is a completely general fact in quantum physics, not just a fact about the quantum moat:

Quantization of angular momentum

The angular momentum of a particle due to its motion through space is quantized in units of \hbar .

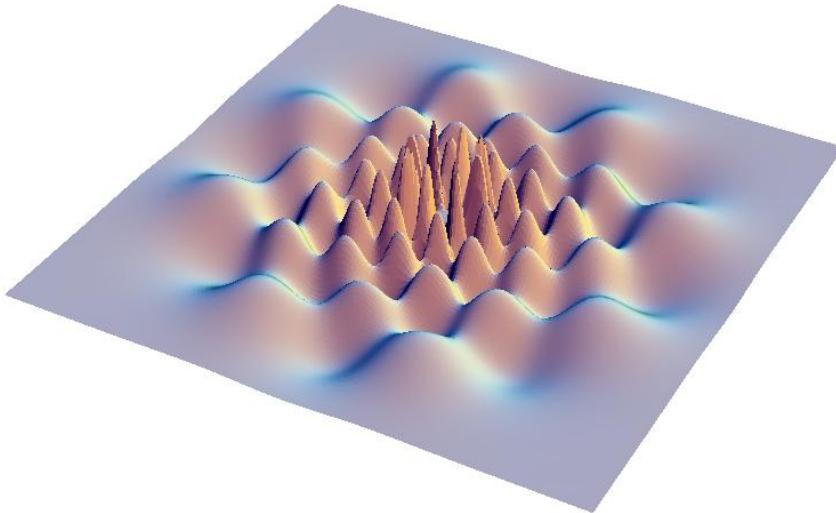


a / 1. Eight wavelengths fit around this circle ($\ell = 8$). This is a standing wave. 2. A traveling wave with $\ell = 8$, depicted according to the color conventions defined in figure j, p. 406.



b / In the benzene molecule, the valence electrons of the carbon atoms occupy quantum states similar to the one in figure a (with smaller values of ℓ , not $\ell = 8$).

c / A depiction of the wavefunction of a certain state in the hydrogen atom. The square is a plane slicing through the center of the atom, so that only two of the three dimensions are shown, say x and y , but not z . The up-down direction in the picture doesn't represent z , it represents Ψ .



self-check A

What is the angular momentum of the wavefunction shown in figure c?

▷ Answer, p. 458

18.2 Three dimensions

Our discussion of quantum-mechanical angular momentum has so far been limited to rotation in a plane, for which we can simply use positive and negative signs to indicate clockwise and counter-clockwise directions of rotation. An atom, however, is unavoidably three-dimensional. We recall from the classical treatment of angular momentum in three-dimensions that the angular momentum of a particle is defined as the vector cross product $\mathbf{r} \times \mathbf{p}$. For an object like a spinning wheel, this vector points along the axis, in the direction given by a right-hand rule.

There is a basic problem here: the angular momentum of the electron in an atom depends on both its distance \mathbf{r} from the proton and its momentum \mathbf{p} , so in order to know its angular momentum precisely it would seem we would need to know both its position and its momentum simultaneously with good accuracy. This, however, seems forbidden by the Heisenberg uncertainty principle.

Actually the uncertainty principle does place limits on what can be known about a particle's angular momentum vector, but it does not prevent us from knowing its magnitude as an exact integer multiple of \hbar . The reason is that in three dimensions, there are really three separate uncertainty principles:

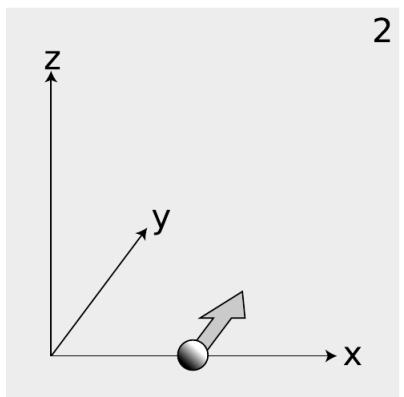
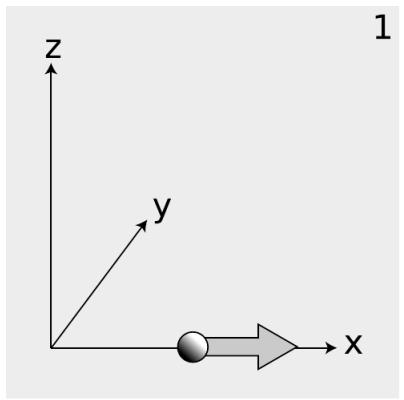
$$\begin{aligned}\Delta p_x \Delta x &\gtrsim \hbar \\ \Delta p_y \Delta y &\gtrsim \hbar \\ \Delta p_z \Delta z &\gtrsim \hbar\end{aligned}$$

Now consider a particle, d/1, that is moving along the x axis at

position x and with momentum p_x . We may not be able to know both x and p_x with unlimited accuracy, but we can still know the particle's angular momentum about the origin exactly: it is zero, because the particle is moving directly away from the origin.

Suppose, on the other hand, a particle finds itself, $d/2$, at a position x along the x axis, and it is moving parallel to the y axis with momentum p_y . It has angular momentum $x p_y$ about the z axis, and again we can know its angular momentum with unlimited accuracy, because the uncertainty principle only relates x to p_x and y to p_y . It does not relate x to p_y .

As shown by these examples, the uncertainty principle does not restrict the accuracy of our knowledge of angular momenta as severely as might be imagined. However, it does prevent us from knowing all three components of an angular momentum vector simultaneously. The most general statement about this is the following theorem:



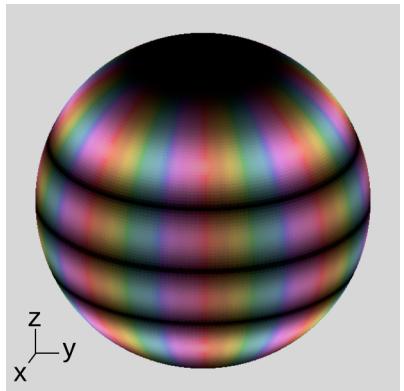
The angular momentum vector in quantum physics

The most that can be known about a (nonzero) orbital angular momentum vector is its magnitude and one of its three vector components. Both are quantized in units of \hbar .

d / Reconciling the uncertainty principle with the definition of angular momentum.

To see why this is true, consider the example wavefunction shown in figure e. This is the like the quantum moat of figure a, p. 425, but extended to one more dimension. If we slice the sphere in any plane perpendicular to the z axis, we get an 8-cycle circular rainbow exactly like figure a. This is required because $L_z = 8\hbar$. But if we take a slice perpendicular to some other axis, such as the y axis, we don't get a circular rainbow as we would for a state with a definite value of L_y . It is obviously not possible to get circular rainbows for slices perpendicular to more than one axis. For those with a taste for rigor, a more careful mathematical argument is given in note [≥439](#).

As a consequence of this fact, we find that when the magnitude of the angular momentum is $L = \ell\hbar$, the total number of states with that value of L is $2\ell + 1$. For example, when $L = 2\hbar$, we have 5 possible values of L_z : $-2, -1, 0, 1$, and 2 . In the language of linear algebra (p. 156), these five states could be used as a basis for the five-dimensional space of states with $L = 2\hbar$. We could just as easily have chosen some other axis besides z . This would have just been a different choice of basis, which is arbitrary.

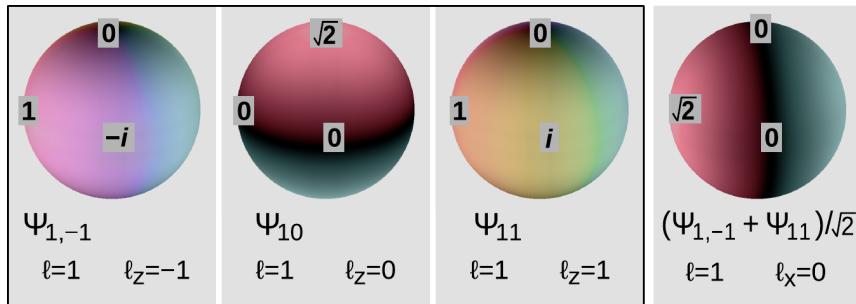


e / A wavefunction on the sphere with $|L| = 11\hbar$ and $L_z = 8\hbar$, shown using the color conventions defined in figure j, p. 406.

18.3 Quantum numbers

18.3.1 Completeness

f / The three states inside the box are a complete set of quantum numbers for $\ell = 1$. Other states with $\ell = 1$, such as the one on the right, are not really new: they can be expressed as superpositions of the original three we chose.



For a given ℓ , consider the set of states with all the possible values of the angular momentum's component along some fixed axis. This set of states is *complete*, meaning that they encompass all the possible states with this ℓ .

For example, figure f shows wavefunctions with $\ell = 1$ that are solutions of the Schrödinger equation for a particle that is confined to the surface of a sphere. Although the formulae for these wavefunctions are not particularly complicated,¹ they are not our main focus here, so to help with getting a feel for the idea of completeness, I have simply selected three points on the sphere at which to give numerical samples of the value of the wavefunction. These are the top (where the sphere is intersected by the positive z axis), left (x), and front (y). (Although the wavefunctions are shown using the color conventions defined in figure j, p. 406, these numerical samples should make the example understandable if you're looking at a black and white copy of the book.)

Suppose we arbitrarily choose the z axis as the one along which to quantize the component of the angular momentum. With this choice, we have three possible values for ℓ_z : -1 , 0 , and 1 . These three states are shown in the three boxes surrounded by the black rectangle. This set of three states is complete.

Consider, for example, the fourth state, shown on the right outside the box. This state is clearly identifiable as a copy of the $\ell_z = 0$ state, rotated by 90 degrees counterclockwise, so it is the $\ell_x = 0$ state. We might imagine that this would be an entirely new prize to be added to our stamp collection. But it is actually not a state that we didn't possess before. We can obtain it as the sum of the $\ell_z = -1$ and $\ell_z = 1$ states, divided by an appropriate normalization factor. Although I'm avoiding making this example an exercise in

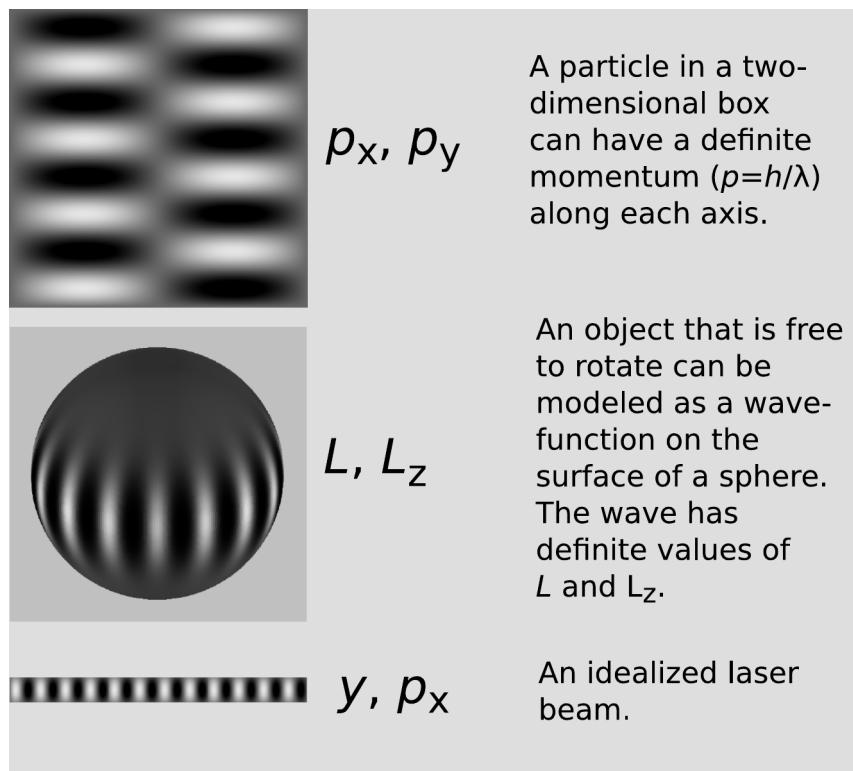
¹They are $\Psi_{1,-1} = \sin \theta e^{-i\phi}$, $\Psi_{10} = \sqrt{2} \cos \theta$, and $\Psi_{11} = \sin \theta e^{i\phi}$, where θ is the angle measured down from the z axis, and ϕ is the angle running counterclockwise around the z axis. These functions are called spherical harmonics.

manipulating formulae, it is easy to check that the sum does work out properly at the three sample points.

18.3.2 Sets of compatible quantum numbers

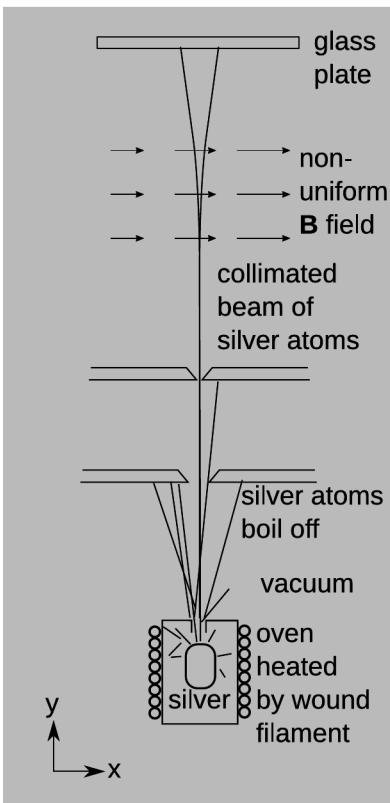
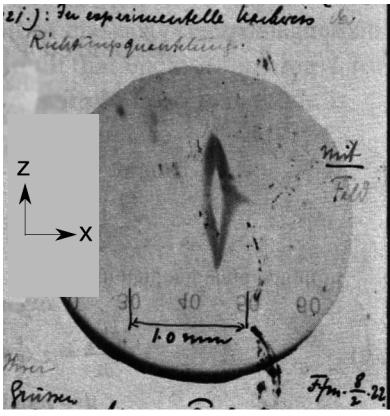
In sec. 16.1.3, p. 372, we discussed the idea of a quantum number, which is simply a label for a state. For example, a particle in a box with a wavefunction like \sqrt{N} can be labeled with the quantum number $N = 2$, which is an energy label. Sometimes a single quantum number isn't enough to give a complete label for a state. For example, if our state was an infinite plane wave in free space, rather than a standing wave in a box, then giving its energy wouldn't be enough information to completely describe the state. Different states could have the same energy but be propagating in different directions. For this reason, we will often want to use *sets* of quantum numbers.

Figure g shows some examples in which we can completely describe a wavefunction by giving a set of quantum numbers. It is important that the quantum numbers we use in describing a state be compatible. By analogy, “Bond, James, 007” would be a clear and consistent definition of the famous fictional spy, but in general this identification scheme would not work, because although almost everyone has a first and last name, most people do not have a license to kill with a corresponding double-oh number.



g / Three examples of sets of compatible quantum numbers.

The laser beam in the figure is a state described according to



h / Bottom: A schematic diagram of the Stern-Gerlach experiment. The z direction is out of the page. The entire apparatus is about 10 cm long. Top: A portion of Gerlach's celebratory 1922 postcard to Niels Bohr, with a photo showing the results. A coordinate system is superimposed. The orientation is flipped downward by 90 degrees compared to the schematic. The photo was taken through a microscope, and Gerlach drew the 1.0 mm scale on after the magnified photo had been printed.

its definite values p_x and y , so we have the vanishing uncertainties $\Delta p_x = 0$ and $\Delta y = 0$. Since the Heisenberg uncertainty principle doesn't talk about an x momentum in relation to a y position, this is OK. If we had been in doubt about whether this violated the uncertainty principle, we would have been reassured by our ability to draw the picture.

It is also possible to have *incompatible* quantum numbers. The combination of p_x with x would be an incompatible set of quantum numbers, because a state can't have a definite p_x and also a definite x . If we try to draw such a wave, we fail. L_x and L_z would also be an incompatible set.

18.3.3 Complete and compatible sets of quantum numbers

Let's summarize. Just as we expect everyone to have a first and last name, we expect there to be a complete and compatible set of quantum numbers for any given quantum-mechanical system. Completeness means that we have enough quantum numbers to uniquely describe every possible state of the system, although we may need to describe a state as a superposition, as with the state $\ell_x = 0$ in figure f on p. 428. Compatibility means that when we specify a set of quantum numbers, we aren't making a set of demands that can't be met.

18.4 The Stern-Gerlach experiment

In 1921, Otto Stern proposed an experiment about angular momentum, shown in figure h on p. 430, that his boss at the University of Frankfurt and many of his colleagues were certain wouldn't work. At this time, quantization of angular momentum had been proposed by Niels Bohr, but most physicists, if they had heard of it at all, thought of the idea as a philosophical metaphor or a mathematical trick that just happened to give correct results. World War I was over, hyperinflation was getting under way in Germany (a paper mark was worth a few percent of its prewar value), and the Nazi coup was still in the future, so that Stern, a Jew, had not yet been forced to flee to America. Because of the difficult economic situation, Stern and his colleague Walther Gerlach scraped up some of the funds to carry out the experiment from US banker Henry Goldman, cofounder of the investment house Goldman-Sachs.

The entire apparatus was sealed inside a vacuum chamber with the best vacuum obtainable at the time. A sample of silver was heated to 1000°C, evaporating it. The atoms leaving the oven encountered two narrow slits, so that what emerged was a beam with a width of only 0.03 mm, or about a third of the width of a human hair. The atoms then encountered a magnetic field. Because the atoms were electrically neutral, we would normally expect them to be unaffected by a magnetic field. But in the planetary model of

the atom, we imagine the electrons as orbiting in circles like little current loops, which would give the atom a magnetic dipole moment \mathbf{m} . Even if we are sophisticated enough about quantum mechanics not to believe in the circular orbits, it is reasonable to imagine that such a dipole moment would exist. When a dipole encounters a *nonuniform* field, it experiences a force.² The rapidly varying magnetic field for this experiment was provided by a pair of specially shaped magnet poles (not shown in the figure).

Because electrons have charge, we expect the motion of an electron to give it a magnetic dipole moment \mathbf{m} . But they also have mass, so for exactly the same reasons, we expect there to be some angular momentum \mathbf{L} as well. The analogy is in fact mathematically exact, so that $\mathbf{m} \propto \mathbf{L}$. Therefore this experiment with dipoles and magnetic fields is actually a probe of the behavior of angular momentum at the atomic level. Luckily for Stern and Gerlach, who had no modern knowledge of atomic structure, the silver atoms that they chose to use do happen to have nonzero total \mathbf{L} , and therefore nonzero \mathbf{m} . The atoms come out of the oven with random orientations.

The details of the electromagnetism are a little complicated ([2439](#)), but the result is that the beam should be deflected in the x direction by an amount proportional to L_x , the x component of its angular momentum. Classically, we would expect the following. Because the orientations of the atoms are random as they enter the magnetic field, they will have every possible value of L_x ranging from $-|\mathbf{L}|$ to $+|\mathbf{L}|$, and therefore we expect that when the magnetic field is turned on, the effect should be to smear out the image on the glass plate from a vertical line to a somewhat wider oval. The atoms are dispersed from left to right along a certain scale of measurement according to their random value of L_x . The spectrometer is a device for determining L_x , a continuously varying number.

But that's all the classical theory. Quantum mechanically, L_x is quantized, so that only certain very specific values of the deflection occur. Therefore we expect to see well separated vertical bands on the glass plate corresponding to the quantized values of L_x . This is approximately what is seen in figure h, although the field rapidly weakens outside the x - y plane, so we get the slightly more complicated pattern like a sideways lipstick kiss. The spin of the silver atom is clearly quantized, and it apparently has two possible values.

Discussion questions

A Could the Stern-Gerlach experiment be carried out with a beam of electrons?

²This is easier to see in the case of an *electric* dipole in a nonuniform *electric* field. If the dipole consists of charges $+q$ and $-q$ at opposite ends of a stick, then the nonuniform field will make unequal forces on them, and the total force will be nonzero.

B A few weeks after the Stern-Gerlach experiment's results became public, Einstein and Ehrenfest carried out the following reasoning, which seemed to them to make the results inexplicable. Before a particular silver atom enters the magnetic field, its magnetic moment \mathbf{m} is randomly oriented. Once it enters the magnetic field, it has an energy $\mathbf{m} \cdot \mathbf{B}$. Unless there is a mechanism for the transfer of energy in or out of the atom, this energy can't change, and therefore the magnetic moment can only precess about the \mathbf{B} vector, but the angle between \mathbf{m} and \mathbf{B} must remain the same. Therefore the atom cannot align itself with the field. (They considered various mechanisms of energy loss, such as collisions and radiation, and concluded that all of them were too slow by orders of magnitude to have an effect during the atom's time of flight.) It seemed to them that as soon as the atom left the oven, it was somehow required to have anticipated the direction of the field and picked one of two orientations with respect to it. How can this paradox be resolved?

C Suppose we send a beam of oxygen molecules, with $L = \hbar$, through a Stern-Gerlach spectrometer, throwing away the emerging parts with $\ell_x = -1$ and $+1$ to make a beam of the pure $\ell_x = 0$ state. Now we let this beam pass through a second spectrometer that is identical but oriented along the z axis. Can we produce a beam in which every molecule has both $\ell_x = 0$ and $\ell_z = +1$? Hint: See the example in fig. f, p. 428.

18.5 Intrinsic spin

18.5.1 Experimental evidence

We observe two values of ℓ_x (the two “lips”) in the Stern-Gerlach experiment. Why two? For a fixed value of ℓ , we have seen that the number of values of ℓ_x is $2\ell + 1$. If we set $2\ell + 1$ equal to 2, we get $\ell = 1/2$. If so, then we would have to conclude from these results that a silver atom has spin $1/2$, so that L_x takes on the two values $-\hbar/2$ and $+\hbar/2$. Although it took about five years for the experiment to be interpreted completely correctly, we now think of this “spin one half” as being the angular momentum of one of the electrons in the silver atom — we can think of it as the outermost electron (see example 2, p. 433).

This may seem paradoxical because the quantum moat (p. 425), for instance, gave only angular momenta that were integer multiples of \hbar , not half-units, and I claimed that angular momentum was always quantized in units of \hbar , not just in the case of the quantum moat. That whole discussion, however, assumed that the angular momentum would come from the motion of a particle through space. The $\hbar/2$ angular momentum of the electron is simply a property of the particle, like its charge or its mass. It has nothing to do with whether the electron is moving or not, and it does not come from any internal motion within the electron. Nobody has ever succeeded in finding any internal structure inside the electron, and even if there was internal structure, it would be mathematically impossible for it to result in a half-unit of angular momentum.

We simply have to accept this $\hbar/2$ angular momentum, called the “spin” of the electron — Mother Nature rubs our noses in it as an observed fact. Protons and neutrons have the same $\hbar/2$ spin, while photons have an intrinsic spin of \hbar . In general, half-integer spins are typical of material particles. Integral values are found for the particles that carry forces: photons, which embody the electric and magnetic fields of force, as well as the more exotic messengers of the nuclear and gravitational forces. The photon is particularly important: it has spin 1.

As was the case with ordinary angular momentum, we can describe spin angular momentum in terms of its magnitude, and its component along a given axis. We write s and s_z for these quantities, expressed in units of \hbar , so an electron has $s = 1/2$ and $s_z = +1/2$ or $-1/2$.

18.5.2 Odds and evens, and how they add up

From grade-school arithmetic, we have the rules

$$\begin{aligned} \text{even} + \text{even} &= \text{even} \\ \text{odd} + \text{even} &= \text{odd} \\ \text{odd} + \text{odd} &= \text{even}. \end{aligned}$$

Thus we know that $123456789 + 987654321$ is even, without having to actually compute the result. Dividing by two gives similar relationships for integer and half-integer angular momenta. For example, a half-integer plus an integer gives a half-integer, and therefore when we add the intrinsic spin $1/2$ of an electron to any additional, integer spin that the electron has from its motion through space, we get a half-integer angular momentum. That is, the *total* angular momentum of an electron will always be a half-integer. Similarly, when we add the intrinsic spin 1 of a photon to its angular momentum due to its integral motion through space, we will always get an integer. Thus the integer or half-integer character of any particle’s *total* angular momentum (spin + motion) is determined entirely by the particle’s spin.

These relationships tell us things about the spins we can make by putting together different particles to make bigger particles, and they also tell us things about decay processes.

Spin of the helium atom

example 1

A helium-4 atom consists of two protons, two neutrons, and two electrons. A proton, a neutron, and an electron each have spin $1/2$. Since the atom is a composite of six particles, each of which has half-integer spin, the atom as a whole has an integer angular momentum.

Silver atoms in the Stern-Gerlach experiment

example 2

The silver atoms used in the Stern-Gerlach experiment had an odd number of protons (47), an even number of neutrons (two

isotopes), and an odd number of electrons (47). The result is that the atom as a whole has an integer spin. However, only the electrons contribute significantly to the magnetic dipole moment of an atom ([2439](#)), so the experiment only probed their angular momentum, which was a half-integer value because 47 is odd. In principle this could be as high as $47/2$, but in the ground state it turns out to be only $1/2$, which can be interpreted as the intrinsic spin of one of the electrons. The other 46 electrons' orbital and intrinsic angular momentum end up canceling out.

Emission of a photon from an atom

example 3

An atom can emit light,

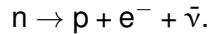


This works in terms of angular momentum because the photon's spin 1 is an integer. Thus, regardless of whether the atom's angular momentum is an integer or a half-integer, the process is allowed by conservation of angular momentum. If the atom's angular momentum is an integer, then we have integer = integer + 1, and if it's a half-integer, half-integer = half-integer + 1; either of these is possible. If not for this logic, it would be impossible for matter to emit light. In general, if we want a particle such as a photon to pop into existence like this, it must have an integer spin.

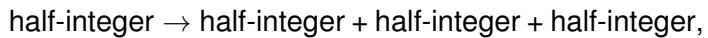
Beta decay

example 4

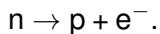
When a free neutron undergoes beta decay, we have



All four of these particles have spin $1/2$, so the angular momenta go like



which is possible, e.g., $1/2 = 3/2 - 5/2 + 3/2$. Because the neutrino has almost no interaction with normal matter, it normally flies off undetected, and the reaction was originally thought to be



With hindsight, this is impossible, because we can never have



The reasoning holds not just for the beta decay of a free neutron, but for any beta decay: a neutrino or antineutrino must be emitted in order to conserve angular momentum. But historically, this was not understood at first, and when Enrico Fermi proposed the existence of the neutrino in 1934, the journal to which he first submitted his paper rejected it as "too remote from reality."

18.5.3 Inner product

We've been thinking of an inner product like $\langle \Psi_1 | \Psi_2 \rangle$ as a measure of the overlap or similarity between two wave patterns, calculated using an integral like $\int \dots dx$. This is fine for integer angular momenta that particles have because they're moving through space, so that, e.g., for states in the "quantum moat," $\langle \ell = 0 | \ell = 1 \rangle = 0$ (ex. 18, p. 442). This makes sense because states are supposed to have an inner product of zero if they are perfectly distinguishable by measuring some observable like ℓ . We expect the same familiar behavior for a spin 1/2 that can exist in states \uparrow and \downarrow , so that $\langle \uparrow | \downarrow \rangle = 0$. Inner products for half-integer spins just won't be interpretable visually as overlaps of wave patterns, and that we won't need to calculate them as integrals.

18.5.4 Classification of states in hydrogen

In sec. 2.7.2, p. 52, we discussed the idea of degeneracy in the context of a classical wave. For standing waves on a square membrane, we have patterns like  and , which have the same frequency f . We say that these patterns are two-fold degenerate. In a quantum-mechanical context, $E = hf$, so degeneracy also implies that states have the same *energy*.

Because the energy of states in the hydrogen atom only depends on n , we have degeneracies, and these degeneracies get doubled because the electron's spin can have two values (example, fig. i). Getting the count right has big implications: in our example, the 8-fold degeneracy is the reason that the second row of the period table has eight chemical elements.

The degeneracy of the different ℓ_z and s_z states follows from symmetry, as in our original example of degeneracy on p. 52, and is therefore exact. The degeneracy with respect to different values of ℓ for the same n is not at all obvious, and is in fact not exact when effects such as relativity are taken into account. We refer to this as an "accidental" degeneracy. The very high level of degeneracy in the hydrogen atom means that when you observe it in the hydrogen spectrum in your lab course, there is a great deal of structure that is effectively hidden from you. Historically, physicists were fooled by the apparent simplicity of the spectrum, and more than 70 years passed between the measurement of the spectrum and the time when the degeneracies were fully recognized and understood.

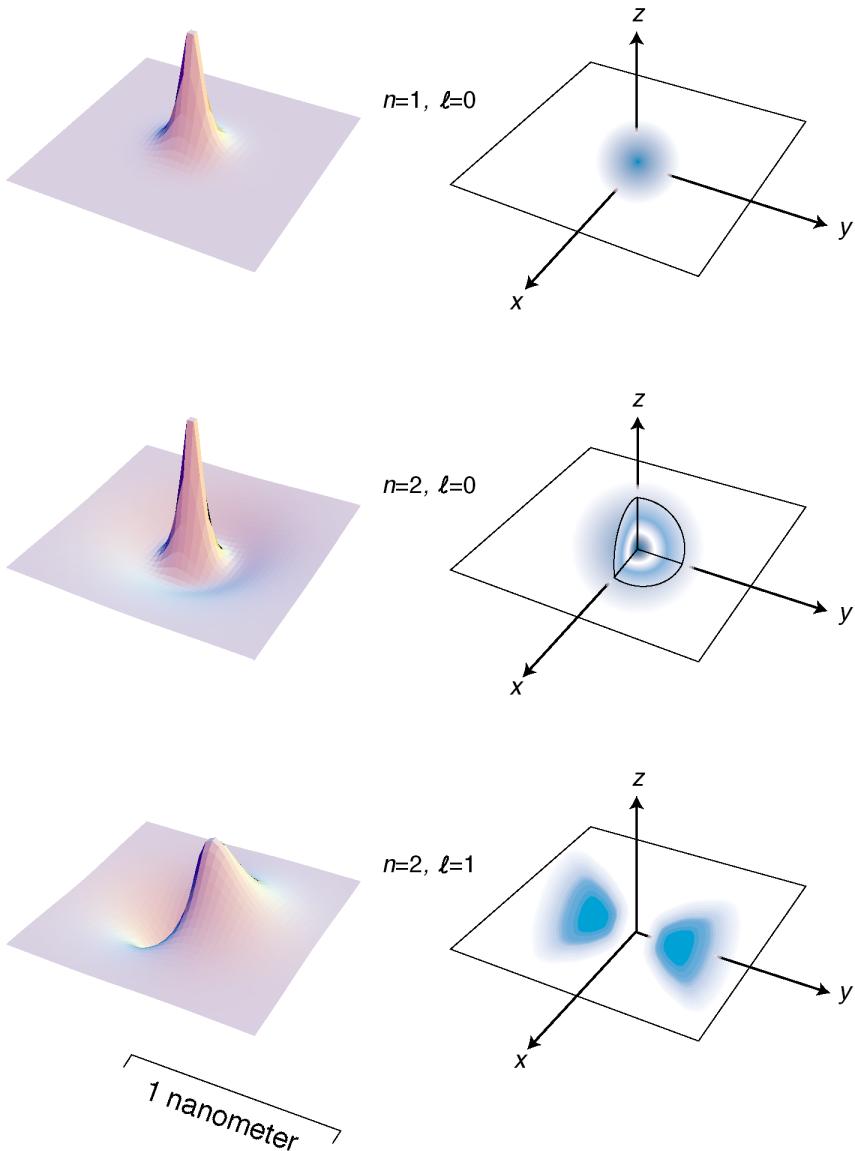
Figure j on page 436 shows the lowest-energy states of the hydrogen atom. The left-hand column of graphs displays the wavefunctions in the $x - y$ plane, and the right-hand column shows the probability distribution in a three-dimensional representation. The ground-state wavefunction, with $n = 1$, $\ell = 0$, was calculated in example 5, p. 401.

Example: the degeneracy of the $n = 2$ state in hydrogen.

The $n = 2$ energy level would be 4-fold degenerate if we didn't know about spin (one $\ell = 0$ state plus three $\ell = 1$ states), but the electron's spin makes it 8-fold degenerate.

i / Counting up a degeneracy.

j / The three states of the hydrogen atom having the lowest energies. Taking into account electron spin, the number of states is actually doubled.



Discussion questions

- A** The quantum number n is defined as the number of radii at which the wavefunction is zero, including $r = \infty$. Relate this to the features of figure j. Based on the definition, why can't there be an $n = 0$ state?
- B** Relate the features of the wavefunction plots in figure j to the corresponding features of the probability distribution pictures.
- C** How can you tell from the wavefunction plots in figure j which ones have which angular momenta?
- D** Criticize the following incorrect statement: "The $\ell = 8$ wavefunction in figure c has a shorter wavelength in the center because in the center the electron is in a higher energy level."
- E** Discuss the implications of the fact that the probability cloud in of the $n = 2, \ell = 1$ state is split into two parts.

18.6 The Pauli exclusion principle

What about other atoms besides hydrogen? It would seem that things would get much more complex with the addition of a second electron. A hydrogen atom only has one particle that moves around much, since the nucleus is so heavy and nearly immobile. Helium, with two, would be a mess. Instead of a wavefunction whose square tells us the probability of finding a single electron at any given location in space, a helium atom would need to have a wavefunction whose square would tell us the probability of finding two electrons at any given combination of points. Ouch! In addition, we would have the extra complication of the electrical interaction between the two electrons, rather than being able to imagine everything in terms of an electron moving in a static field of force created by the nucleus alone.

Despite all this, it turns out that we can get a surprisingly good description of many-electron atoms simply by assuming the electrons can occupy the same standing-wave patterns that exist in a hydrogen atom. The ground state of helium, for example, would have both electrons in states that are very similar to the $n = 1$ states of hydrogen. The second-lowest-energy state of helium would have one electron in an $n = 1$ state, and the other in an $n = 2$ state. The relatively complex spectra of elements heavier than hydrogen can be understood as arising from the great number of possible combinations of states for the electrons.

A surprising thing happens, however, with lithium, the three-electron atom. We would expect the ground state of this atom to be one in which all three electrons settle down into $n = 1$ states. What really happens is that two electrons go into $n = 1$ states, but the third stays up in an $n = 2$ state. This is a consequence of a new principle of physics:

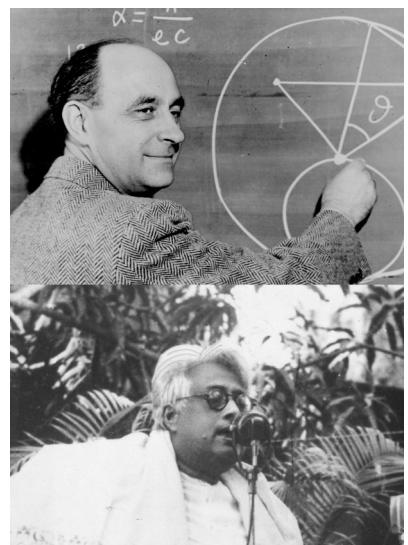
The Pauli Exclusion Principle

Two electrons can never occupy the same state. If one electron has wavefunction u , and another electron v , then $\langle u|v \rangle = 0$.

There are two $n = 1$ states, one with $s_z = +1/2$ and one with $s_z = -1/2$, but there is no third $n = 1$ state for lithium's third electron to occupy, so it is forced to go into an $n = 2$ state.

It can be proved mathematically that the Pauli exclusion principle applies to any type of particle that has half-integer spin. Thus two neutrons can never occupy the same state, and likewise for two protons. Such particles are referred to as fermions, after Enrico Fermi, and, broadly speaking, fundamental fermions are the particles that matter is made of.

Photons, however, are immune to the exclusion principle be-



k / Top: Enrico Fermi (1901-1954). Bottom: Satyendra Nath Bose (1894-1974).

cause their spin is an integer. Particles with integer spins are called bosons, after Satyendra Nath Bose. Bosons are generally the force-carriers in physics. Classically, we can say that radio signals work because electrical and magnetic forces propagate from the transmitter to the receiver, but quantum-mechanically, we would say that photons are what is being transmitted. Similarly, the strong nuclear force is transmitted by a type of bosons known as gluons, and the weak nuclear force by particles called the W and Z bosons.

This is a brief and incomplete sketch of what is known as the standard model of particle physics, which describes 17 types of fundamental particles, their properties and interactions.

To make the picture neat and tidy, we would like gravity to be like the other forces. Gravitational forces should be carried by some boson, which we would call the graviton. We can even tell some properties of the graviton: it should be massless, electrically neutral, and have spin $2\hbar$. However, there are fundamental reasons why our theories of quantum mechanics break down when we try to apply them to gravity. It is almost certainly true that gravitons exist, but we can't calculate much of anything about their detailed behavior. It is probably impractical as well to detect them directly using any foreseeable human technology — people designing hypothetical detectors end up talking about devices the size of an entire solar system. It is possible, however, that very sensitive and clever experiments could detect some effects that would give us a clue about how to reconcile gravity with quantum mechanics.

Notes for chapter 18

≥427 L_x incompatible with L_z

Theorem: On the sphere, if a wavefunction has definite values of both L_z and L_x , then it is a wavefunction that is constant everywhere, so $\mathbf{L} = 0$.

Lemma 1: If the component of ℓ_A along a certain axis A has a definite value and is nonzero, then (a) $\Psi = 0$ at the poles, and (b) Ψ is of the form $Ae^{i\ell_A\phi}$ on any circle in a plane perpendicular to the axis. Part a holds because $\mathbf{L} = 0$ if $r_\perp = 0$. For b, see p. 425.

Lemma 2: If the component of \mathbf{L} along a certain axis has a definite value and is zero, then Ψ is constant in any plane perpendicular to that axis. This follows from lemma 1 in the case where $\ell_A = 0$.

Case I: ℓ_z and ℓ_x are both nonzero. We have $\Psi = 0$ at the poles along both the x axis and the z axis. The z -axis pole is a point on the great circle perpendicular to the x axis, and vice versa, so applying 1b, $A = 0$ and Ψ vanishes on both of these great circles. But now if we apply 1b along any slice perpendicular to either axis, we get $\Psi = 0$ everywhere on that slice, so $\Psi = 0$ everywhere.

Case II: ℓ_z and ℓ_x are both zero. By lemma 2, Ψ is a constant everywhere.

Case III: One component is zero and the other nonzero. Let ℓ_z be the one that is zero. By 1a, $\Psi = 0$ at the x -axis pole, so by 2, $\Psi = 0$ on the great circle perpendicular to z . But then 1b tells us that $\Psi = 0$ everywhere.

≥431 Details of the electromagnetism involved in the Stern-Gerlach experiment

In this example, the forces in the x and z directions would be $F_x = \mathbf{m} \cdot (\partial \mathbf{B} / \partial x)$ and $F_z = \mathbf{m} \cdot (\partial \mathbf{B} / \partial z)$. (Because of Gauss's law for magnetism, these two derivatives are not independent — we have $\partial B_x / \partial x + \partial B_z / \partial z = 0$.)

Classically, we would expect the following. Each atom has an energy $\mathbf{m} \cdot \mathbf{B}$ due to its interaction with the magnetic field, and this energy

is conserved, so that the component m_x stays constant. However, there is a torque $\mathbf{m} \times \mathbf{B}$, and this causes the direction of the atom's angular momentum to precess, i.e., wobble like a top, with its angular momentum forming a cone centered on the x axis. This precession is extremely fast, carrying out about 10^{10} wobbles per second, so that the atom precesses about 10^6 times while traveling the 3.5 cm length of the spectrometer. So even though the forces F_x and F_z are typically about the same size, the rapid precession causes F_z to average out to nearly zero, and only a deflection in the x direction is expected.

Although the discussion of precession above is really classical rather than quantum-mechanical, the result of F_z averaging out to zero turns out to be approximately right if the field is strong.

As a side issue (example 2, p. 433), there is the question of whether the Stern-Gerlach experiment should also be sensitive to the angular momentum of the nucleus. It isn't very sensitive to this, because we have the analogy that mass is to angular momentum as charge is to the magnetic dipole moment, so that at a crude level of analysis, ignoring things like factors of two, we expect that for a fixed angular momentum, the dipole moment is proportional to the charge-to-mass ratio of the particle. For this reason, the magnetic dipole moment contributed by one of the protons is thousands of times smaller than that created by one of the electrons. The neutrons are electrically neutral, so we would not expect them to contribute at all. They actually contain electrically charged particles — quarks — but their contributions to the magnetic dipole moment are still small, for the same reason as in the case of the protons.

Problems

Key

- ✓ A computerized answer check is available online.
- ★ A difficult problem.

1 Estimate the angular momentum of a spinning basketball, in units of \hbar . Explain how this result relates to the correspondence principle.

2 The ground state of the nucleus boron-10 (${}^{10}\text{B}$) has spin $s = 3$, and is therefore degenerate.

(a) List the s_z states and check that the number of states you get is $2s + 1$.

(b) Based on the spin, is the ground state of ${}^{10}\text{B}$ a fermion, or is it a boson?

(c) ${}^{10}\text{B}$ has four excited states that are bound, and these have various spins. Could some of these states be fermionic and some bosonic?

3 Hydrogen-2 (${}^2\text{H}$) is referred to as deuterium. It contains one proton and one neutron. Its ground state is its only bound state, and in this state the neutron and proton have the following quantum numbers:

neutron: $\ell = 0, \ell_z = 0, s = 1/2, s_z = 1/2$

proton: $\ell = 0, \ell_z = 0, s = 1/2, s_z = 1/2$.

Here the z axis has been chosen parallel to the total angular momentum, and the total angular momentum is 1. Another state in which we could put the system is this one:

neutron: $\ell = 0, \ell_z = 0, s = 1/2, s_z = 1/2$

proton: $\ell = 0, \ell_z = 0, s = 1/2, s_z = -1/2$.

This state's total angular momentum is 0. (This state is observed to be unbound, but we're not concerned in this problem with whether states are bound or unbound.)

Suppose that our system instead consisted of two neutrons and no protons at all. Could you put them in the spin-1 state? In the spin-0 state? (Don't worry about whether these states are bound.)

4 This problem builds on the results of problems 16-5 (p. 389) and 17-6 (p. 422).

Suppose we have a three-dimensional box of dimensions $L \times L \times L/2$. Let the box be oriented so that the shorter dimension is along the z direction. For convenience, define the quantity $\epsilon = h^2/8mL^2$, which has units of energy.

(a) What are the five lowest energies allowed in this box, expressed in terms of ϵ ? Give the quantum numbers for each energy, and find the degeneracy (p. 435) of each.

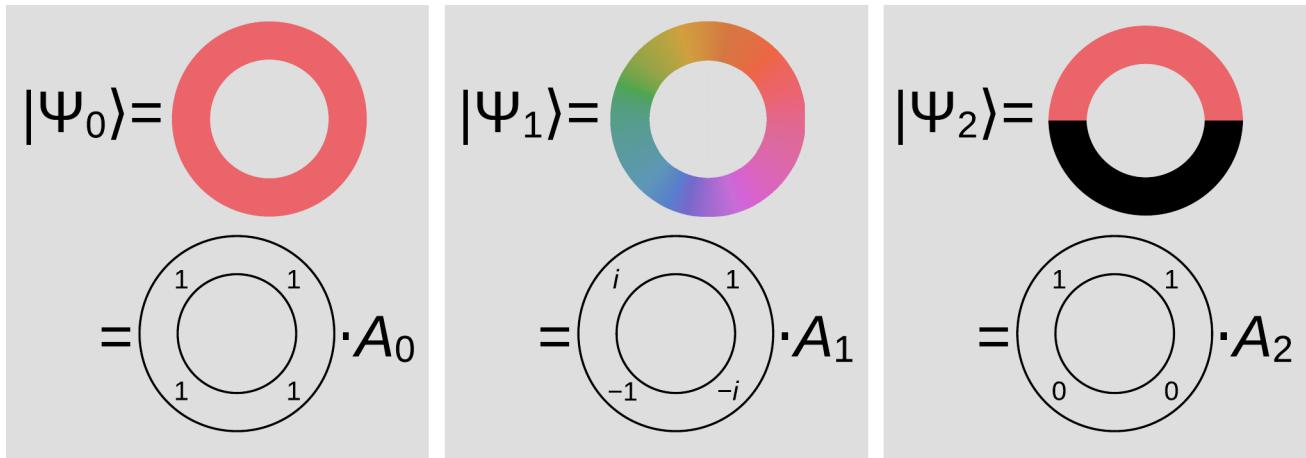
(b) Suppose we put five electrons in this box such that they have the lowest possible total energy. (Keep in mind that there is a limit to how many electrons can have the same spatial wavefunction.) What is the total energy of this state? \checkmark

(c) What are the two lowest-energy photons that can excite one of the five electrons (from the situation described in part b) to an excited state? \checkmark

[Problem by B. Shotwell.]

Exercise 18: The quantum moat

Consider the following three states for a particle confined to a circle, like a wave in the moat surrounding a castle.



We show each wavefunction first using the color conventions introduced on p. 406, and then using numbers. The numerical description is discretized, i.e., we only consider the values of the wavefunction at found points, evenly spaced around the circle. In this discretized representation, we define the inner product using a sum rather than an integral. We can refer to the wavefunctions by notations such as $|\Psi_0\rangle$ or simply $|0\rangle$. The A 's are normalization constants.

1. In this representation, how would a wavefunction like $\langle\Psi_1|$ differ from $|\Psi_1\rangle$?
2. Which of these are states of definite angular momentum?
3. The instructor will demonstrate the determination of the magnitude $|A_0|$ of the normalization constant for $|0\rangle$.
4. The students determine the magnitudes of A_1 and A_2 .
5. Suppose that we first prepare the particle in state 2, and then we measure its angular momentum. The instructor will use the discretized wavefunction to approximate $|\langle 0|2\rangle|^2$, which can be interpreted as the probability that the measurement results in $\ell = 0$.
6. The students compute $|\langle 1|2\rangle|^2$ and give a similar interpretation.³
7. How would $\langle 1|2\rangle$ compare with $\langle 2|1\rangle$?
8. Compute $|\langle 0|1\rangle|^2$ and interpret the result.

³This is the first of our results in which the discretization actually introduces a nonzero error. For comparison, you should find that your result is reasonably close to the exact $|A_1 A_2 \int_0^2 \exp(-i\pi x/2) dx|^2 = 2/\pi^2 \approx 0.20$.

Hints

Hints for chapter 2

Page 58, problem 6:

How could you change the values of x and t so that the value of y would remain the same? What would this represent physically?

Page 61, problem 19:

(a) The most straightforward approach is to apply the equation $\partial^2y/\partial t^2 = (T/\mu)\partial^2y/\partial x^2$. Although this equation was developed in the main text in the context of a straight string with a curvy wave on it, it works just as well for a circular loop; the left-hand side is simply the inward acceleration of any point on the rope. Note, however, that we've been assuming the string was (at least approximately) parallel to the x axis, which will only be true if you choose a specific value of x . You need to get an equation for y in terms of x in order to evaluate the right-hand side.

Hints for chapter 5

Page 140, problem 1:

The answers to the two parts are not the same.

Page 144, problem 18:

There are various ways of doing this, but one easy and natural approach is to change the base of the exponent to e using the same method that we would use for real numbers.

Hints for chapter 11

Page 255, problem 26:

Expand $\sin \theta$ in a Taylor series around $\theta = 90^\circ$.

Solutions to selected problems

Solutions for chapter 2

Page 58, problem 7:

We have $\omega = 2\pi f = 630$ Hz and $k = 2\pi/\lambda = 2\pi f/v = 0.66$ m $^{-1}$. The wave's equation is

$$u = \sin(kx - \omega t + \delta),$$

which, if we like, we can make into an explicit numerical equation,

$$u = \sin [(0.66 \text{ m}^{-1})x - (630 \text{ s}^{-1})t + \delta].$$

Page 60, problem 15:

(a) The quantity $x - y$ vanishes along the line $y = x$ lying in the first quadrant at a 45-degree angle between the axes. Squaring produces a trough parallel to this line, with a parabolic cross-section. Geometrically, the Laplacian can be interpreted as a measure of how much the value of f at a point differs from its average value on a small circle centered on that point. The trough is concave up, so we can predict that the Laplacian will be positive everywhere.

(b) The zero result is clearly wrong because it disagrees with our conclusion from part a that the Laplacian is positive. A correct calculation gives $\partial^2(x - y)^2/\partial x^2 + \partial^2(x - y)^2/\partial y^2 = 4$.

Remark: The mistake described in the question is a common one, and is apparently based on

the idea that the notation ∇^2 must mean applying an operator ∇ twice. For those with some exposure to vector calculus, it may be of interest to note that the Laplacian *is* equivalent to the divergence of the gradient, which can be notated either $\operatorname{div}(\operatorname{grad} f)$ or $\nabla \cdot (\nabla f)$. The important thing to recognize is that the gradient, notated $\operatorname{grad} f$ or ∇f , outputs a *vector*, not a scalar like the quantity Q defined in this problem.

Solutions for chapter 3

Page 83, problem 6:

- (a) The Poynting vectors cancel.
- (b) The electric fields cancel, while the magnetic field doubles. Since the total electric field is zero, $\mathbf{E} \times \mathbf{B} = 0$, and the Poynting vector is zero.
- (c) Both methods give zero. This makes sense physically because we interpret the Poynting vector as a measure of the flow of energy. Energy is flowing in and out of the page at equal rates, so there is zero total flow.

Page 84, problem 9:

Although this is not a plane wave, if we take any small section of it, such as one of the squares in the figure, it can be approximated as a plane wave. Therefore we expect the electric and magnetic fields to be like those in a plane wave: perpendicular to each other and with $E = cB$. Since they are perpendicular to each other, the cross product occurring in the expression for the Poynting vector is equal to the product of the magnitudes EB , and we must have $EB \propto r^{-2}$. Because $E = cB$, the two fields must have the same dependence on r , and this means that we must have both $E \propto r^{-1}$ and $B \propto r^{-1}$. This is somewhat counterintuitive; it tells us that radiation fields fall off *more slowly* than the static field of a point source.

Solutions for chapter 5

Page 143, problem 12:

$$\begin{aligned}\sin(a+b) &= \left(e^{i(a+b)} - e^{-i(a+b)} \right) / 2i \\ &= \left(e^{ia} e^{ib} - e^{-ia} e^{-ib} \right) / 2i \\ &= [(\cos a + i \sin a)(\cos b + i \sin b) - (\cos a - i \sin a)(\cos b - i \sin b)] / 2i \\ &= \cos a \sin b + \sin a \cos b\end{aligned}$$

By a similar computation, we find $\cos(a+b) = \cos a \cos b - \sin a \sin b$.

Page 143, problem 13:

If $z^3 = 1$, then we know that $|z| = 1$, since cubing z cubes its magnitude. Cubing z triples its argument, so the argument of z must be a number that, when tripled, is equivalent to an angle of zero. There are three possibilities: $0 \times 3 = 0$, $(2\pi/3) \times 3 = 2\pi$, and $(4\pi/3) \times 3 = 4\pi$. (Other possibilities, such as $(32\pi/3)$, are equivalent to one of these.) The solutions are:

$$z = 1, e^{2\pi i/3}, e^{4\pi i/3}$$

Page 143, problem 14:

This function would be represented by the complex number 1, which lies on the positive real axis, one unit to the right of the origin. In this system of analogies, differentiation is represented by multiplication by $i\omega$, which here is $2i$. Taking a fourth derivative is represented by multiplying

four times by $2i$, i.e., we take our original point, 1, and make it into $1\dot{(2i)}^4 = 16$. Satisfying the differential equation then amounts to having $16 - 16\dot{1} = 0$, which is true.

Page 145, problem 22:

We have $n = \sin \phi / \sin \theta$. Doing implicit differentiation, we find $dn = -\sin \phi (\cos \theta / \sin^2 \theta) d\theta$, which can be rewritten as $dn = -n \cot \theta d\theta$. This can be minimized by making θ as big as possible. To make θ as big as possible, we want ϕ to be as close as possible to 90 degrees, i.e., almost grazing the surface of the tank.

This result makes sense, because we're depending on refraction in order to get a measurement of n . At $\phi = 0$, we get $\theta = 0$, which provides no information at all about the index of refraction — the error bars become infinite. The amount of refraction increases as the angles get bigger.

Solutions for chapter 6

Page 159, problem 3:

As in the example, we have as our starting point the relations $m^2 = E^2 - p^2$ and $v = p/E$. Here we want to eliminate p , so we substitute $p = vE$ into the definition of mass, which gives $m^2 = E^2 - v^2 E^2$. Solving this for E gives an expression that can be written most compactly as $E = m\gamma$.

Page 159, problem 5:

$\{\hat{\mathbf{x}}\}$ is not a basis, because there are vectors such as $\hat{\mathbf{y}}$ that we can't form as a linear combination (i.e., scalar multiple) of $\hat{\mathbf{x}}$. $\{\hat{\mathbf{x}}, \hat{\mathbf{y}}\}$ is the standard basis for this vector space. $\{-\hat{\mathbf{x}}, \hat{\mathbf{x}} + \hat{\mathbf{y}}\}$ also works as a basis, because the two vectors are linearly independent, and it's easy to check that any vector in the plane can be formed as a linear superposition of them. $\{\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{x}} + \hat{\mathbf{y}}\}$ is not a basis, because these three vectors are not linearly independent.

Page 160, problem 6:

- (a) The sketch for ℓ will be a 45-degree line through the origin, while r will be only the part of that line in the first quadrant. Of the two, only ℓ is a vector space. The set r isn't a vector space, because it doesn't have additive inverses.
- (b) We have $(1/2)(\pi + \pi) = 0$, but $(1/2)\pi + (1/2)\pi = \pi$.

Page 160, problem 7:

To do anything useful with these expressions describing units, we need to be able to talk about things like dividing meters by seconds to get meters per second. Thus “addition” needs to be multiplication, which corresponds to adding the exponents. Scalar “multiplication” actually has to be exponentiation, e.g., “multiplying” units of meters by the scalar 2 should give square meters.

Solutions for chapter 7

Page 180, problem 9:

- (a) Roughly speaking, the thermal energy is $\sim k_B T$ (where k_B is the Boltzmann constant), and we need this to be on the same order of magnitude as ke^2/r (where k is the Coulomb constant). For this type of rough estimate it's not especially crucial to get all the factors of two right, but let's do so anyway. Each proton's average kinetic energy due to motion along a particular axis is $(1/2)k_B T$. If two protons are colliding along a certain line in the center-of-mass frame, then their average combined kinetic energy due to motion along that axis is $2(1/2)k_B T = k_B T$. So in fact the factors of 2 cancel. We have $T = ke^2/k_B r$.
- (b) The units are $K = (J \cdot m/C^2)(C^2)/((J/K) \cdot m)$, which does work out.
- (c) The numerical result is $\sim 10^{10}$ K, which as suggested is much higher than the temperature

at the core of the sun.

Solutions for chapter 8

Page 197, problem 1:

(a) We have

$$\begin{aligned} dP &= \rho g dy \\ \Delta P &= \int \rho g dy, \end{aligned}$$

and since we're taking water to be incompressible, and g doesn't change very much over 11 km of height, we can treat ρ and g as constants and take them outside the integral.

$$\begin{aligned} \Delta P &= \rho g \Delta y \\ &= (1.0 \text{ g/cm}^3)(9.8 \text{ m/s}^2)(11.0 \text{ km}) \\ &= (1.0 \times 10^3 \text{ kg/m}^3)(9.8 \text{ m/s}^2)(1.10 \times 10^4 \text{ m}) \\ &= 1.0 \times 10^8 \text{ Pa} \\ &= 1.0 \times 10^3 \text{ atm}. \end{aligned}$$

The precision of the result is limited to a few percent, due to the compressibility of the water, so we have at most two significant figures. If the change in pressure were exactly a thousand atmospheres, then the pressure at the bottom would be 1001 atmospheres; however, this distinction is not relevant at the level of approximation we're attempting here.

(b) Since the air in the bubble is in thermal contact with the water, it's reasonable to assume that it keeps the same temperature the whole time. The ideal gas law is $PV = nkT$, and rewriting this as a proportionality gives

$$V \propto P^{-1},$$

or

$$\frac{V_f}{V_i} = \left(\frac{P_f}{P_i} \right)^{-1} \approx 10^3.$$

Since the volume is proportional to the cube of the linear dimensions, the growth in radius is about a factor of 10.

Page 197, problem 2:

(a) If the expression $1 + by$ is to make sense, then by has to be unitless, so b has units of m^{-1} . The input to the exponential function also has to be unitless, so k also has of m^{-1} . The only factor with units on the right-hand side is P_o , so P_o must have units of pressure, or Pa.

(b)

$$\begin{aligned} dP &= \rho g dy \\ \rho &= \frac{1}{g} \frac{dP}{dy} \\ &= \frac{P_o}{g} e^{-ky} (-k - kby + b) \end{aligned}$$

(c) The three terms inside the parentheses on the right all have units of m^{-1} , so it makes sense to add them, and the factor in parentheses has those units. The units of the result from b then

look like

$$\begin{aligned}\frac{\text{kg}}{\text{m}^3} &= \frac{\text{Pa}}{\text{m/s}^2} \text{m}^{-1} \\ &= \frac{\text{N/m}^2}{\text{m}^2/\text{s}^2} \\ &= \frac{\text{kg} \cdot \text{m}^{-1} \cdot \text{s}^{-2}}{\text{m}^2/\text{s}^2},\end{aligned}$$

which checks out.

Solutions for chapter 9

Page 209, problem 3:

If the full-sized brick A undergoes some process, such as heating it with a blowtorch, then we want to be able to apply the equation $\Delta S = Q/T$ to either the whole brick or half of it, which would be identical to B. When we redefine the boundary of the system to contain only half of the brick, the quantities ΔS and Q are each half as big, because entropy and energy are additive quantities. T , meanwhile, stays the same, because temperature isn't additive — two cups of coffee aren't twice as hot as one. These changes to the variables leave the equation consistent, since each side has been divided by 2.

Solutions for chapter 10

Page 228, problem 2:

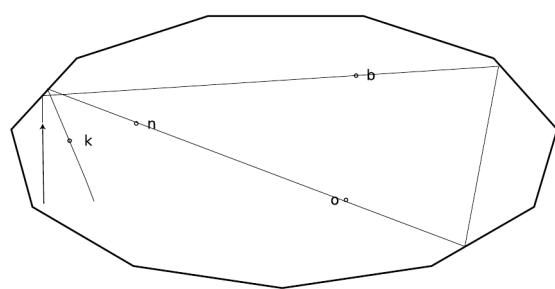
Because the surfaces are flat, you get specular reflection. In specular reflection, all the reflected rays go in one direction. Unless the plane is directly overhead, that direction won't be the right direction to make the rays come back to the radar station.



This is different from a normal plane, which has complicated, bumpy surfaces. These surfaces give diffuse reflection, which spreads the reflected rays randomly in more or less every possible direction.

Page 228, problem 3:

It spells "bonk."



Solutions for chapter 11

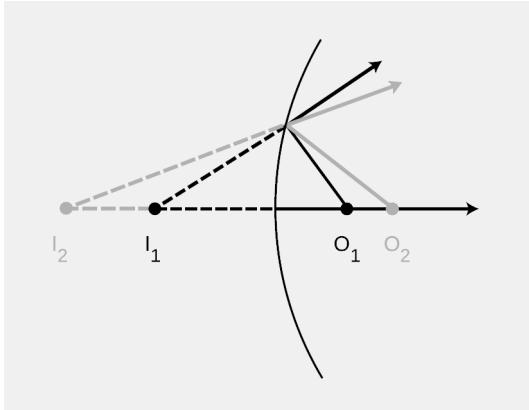
Page 250, problem 1:

For a flat mirror, d_i and d_o are equal, so the magnification is 1, i.e., the image is the same size

as the object.

Page 250, problem 3:

See the ray diagram below. Decreasing θ_o decreases θ_i , so the equation $\theta_f = \pm\theta_i + \pm\theta_o$ must have opposite signs on the right. Since θ_o is bigger than θ_i , the only way to get a positive θ_f is if the signs are $\theta_f = -\theta_i + \theta_o$. This gives $1/f = -1/d_i + 1/d_o$.



Page 250, problem 5:

- (a) The object distance is less than the focal length, so the image is virtual: because the object is so close, the cone of rays is diverging too strongly for the mirror to bring it back to a focus.
(b) Now the object distance is greater than the focal length, so the image is real. (c),(d) A diverging mirror can only make virtual images.

Page 250, problem 6:

- (a) In problem #2 we found that the equation relating the object and image distances was of the form $1/f = -1/d_i + 1/d_o$. Let's make $f = 1.00$ m. To get a virtual image we need $d_o < f$, so let $d_o = 0.50$ m. Solving for d_i , we find $d_i = 1/(1/d_o - 1/f) = 1.00$ m. The magnification is $M = d_i/d_o = 2.00$. If we change d_o to 0.55 m, the magnification becomes 2.22. The magnification changes somewhat with distance, so the store's ad must be assuming you'll use the mirror at a certain distance. It can't have a magnification of 5 at all distances.
(b) Theoretically yes, but in practical terms no. If you go through a calculation similar to the one in part a, you'll find that the images of both planets are formed at almost exactly the same d_i , $d_i = f$, since $1/d_o$ is pretty close to zero for any astronomical object. The more distant planet has an image half as big ($M = d_i/d_o$, and d_o is doubled), but we're talking about *angular* magnification here, so what we care about is the angular size of the image compared to the angular size of the object. The more distant planet has half the angular size, but its image has half the angular size as well, so the angular magnification is the same. If you think about it, it wouldn't make much sense for the angular magnification to depend on the planet's distance — if it did, then determining astronomical distances would be much easier than it actually is!

Page 251, problem 7:

- (a) This occurs when the d_i is infinite. Let's say it's a converging mirror creating a virtual image, as in problems 2 and 3. Then we'd get an infinite d_i if we put $d_o = f$, i.e., the object is at the focal point of the mirror. The image is infinitely large, but it's also infinitely far away, so its angular size isn't infinite; an angular size can never be more than about 180° since you can't see in back of your head!
(b) It's not possible to make the magnification infinite by having $d_o = 0$. The image location and object location are related by $1/f = 1/d_o - 1/d_i$, so $1/d_i = 1/d_o - 1/f$. If d_o is zero, then

$1/d_o$ is infinite, $1/d_i$ is infinite, and d_i is zero as well. In other words, as d_o approaches zero, so does d_i , and d_i/d_o doesn't blow up. Physically, the mirror's curvature becomes irrelevant from the point of view of a tiny flea sitting on its surface: the mirror seems flat to the flea. So physically the magnification would be 1, not infinity, for very small values of d_o .

Page 251, problem 9:

The magnification is the ratio of the image's size to the object's size. It has nothing to do with the person's location. The angular magnification, however, does depend on the person's location, because things farther away subtend smaller angles. The distance to the actual object is not changed significantly, since it's zillions of miles away in outer space, but the distance to the image does change if the observer's point of view changes. If you can get closer to the image, the angular magnification is greater.

Page 252, problem 11:

The refracted ray that was bent closer to the normal in the plastic when the plastic was in air will be bent farther from the normal in the plastic when the plastic is in water. It will become a diverging lens.

Page 252, problem 13:

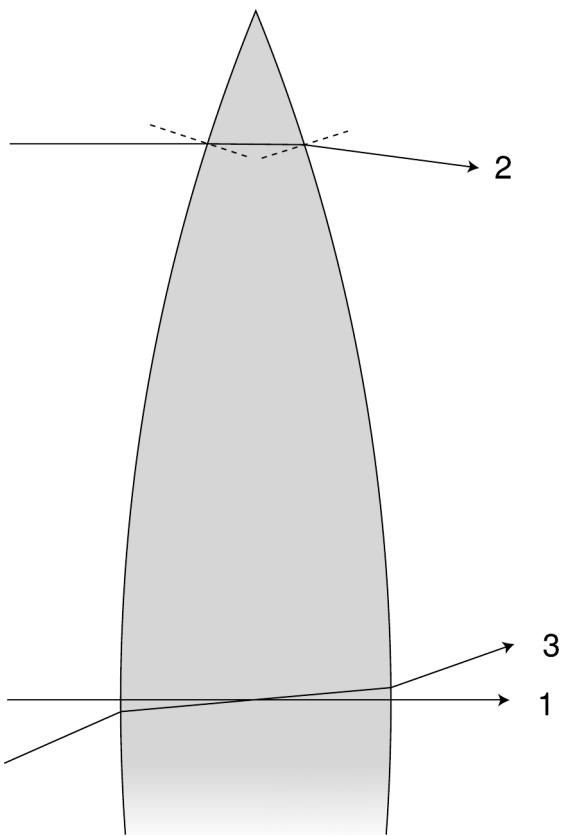
Refraction occurs only at the boundary between two substances, which in this case means the surface of the lens. Light doesn't get bent at all inside the lens, so the thickness of the lens isn't really what's important. What matters is the angles of the lens' surfaces at various points.

Ray 1 makes an angle of zero with respect to the normal as it enters the lens, so it doesn't get bent at all, and likewise at the back.

At the edge of the lens, 2, the front and back are not parallel, so a ray that traverses the lens at the edge ends up being bent quite a bit.

Although I drew both ray 1 and ray 2 coming in along the axis of the lens, it really doesn't matter. For instance, ray 3 bends on the way in, but bends an equal amount on the way out, so it still emerges from the lens moving in the same direction as the direction it originally had.

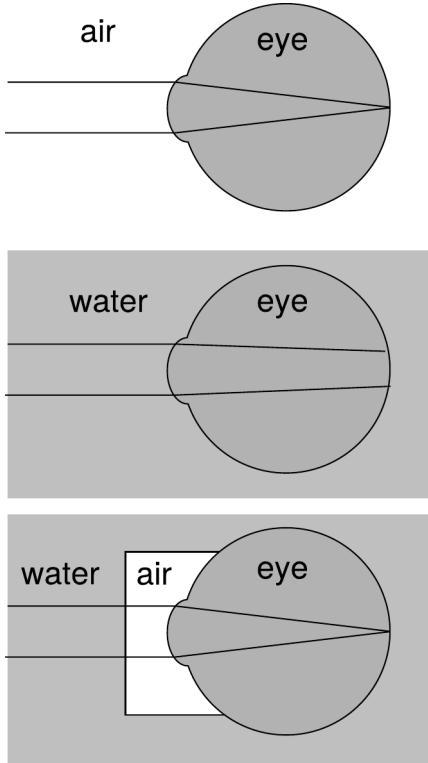
Summarizing and systematizing these observations, we can say that for a ray that enters the lens at the center, where the surfaces are parallel, the sum of the two deflection angles is zero. Since the total deflection is zero at the center, it must be larger away from the center.



Page 252, problem 15:

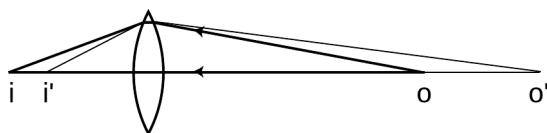
Normally, in air, your eyes do most of their focusing at the air-eye boundary. When you swim without goggles, there is almost no difference in speed at the water-eye interface, so light is not strongly refracted there (see figure), and the image is far behind the retina.

Goggles fix this problem for the following reason. The light rays cross a water-air boundary as they enter the goggles, but they're coming in along the normal, so they don't get bent. At the air-eye boundary, they get bent the same amount they normally would when you weren't swimming.



Page 253, problem 17:

- (a) The situation being described requires a real image, since the rays need to converge at a point on Becky's neck. See the ray diagram drawn with thick lines, showing object location o and image location i .



If we move the object farther away, to o' the cone of rays intercepted by the lens (thin lines) is less strongly diverging, and the lens is able to bring it to a closer focus, at i' . In the diagrams, we see that a smaller θ_o leads to a larger θ_i , so the signs in the equation $\pm\theta_o \pm \theta_i = \theta_f$ must be the same, and therefore both positive, since θ_f is positive by definition. The equation relating the image and object locations must be $1/f = 1/d_o + 1/d_i$.

- (b) The case with $d_i = f$ is not possible, because then we need $1/d_o = 0$, i.e., $d_o = \infty$. Although it is possible in principle to have an object so far away that it is practically at infinity, that is not possible in this situation, since Zahra can't take her lens very far away from the fire. By the way, this means that the *focal length* f is not where the *focus* happens — the focus happens at d_i .

For similar reasons, we can't have $d_o = f$.

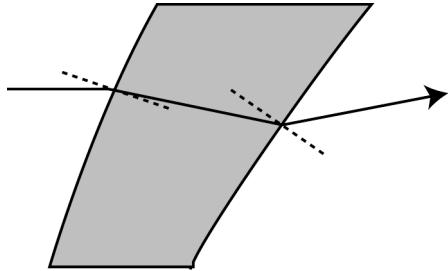
Since all the variables are positive, we must have $1/d_o$ and $1/d_i$ both less than $1/f$. This implies that $d_o > f$ and $d_i > f$. Of the nine logical possibilities in the table, only this one is actually possible for this real image.

Page 253, problem 18:

One surface is curved outward and one inward. Therefore the minus sign applies in the lens-maker's equation. Since the radii of curvature are equal, the quantity $1/r_1 - 1/r_2$ equals zero, and the resulting focal length is infinite. A big focal length indicates a weak lens. An infinite focal length tells us that the lens is infinitely weak — it doesn't focus or defocus rays at all.

Page 253, problem 19:

(a) See the figure below. The first refraction clearly bends it inward. However, the back surface of the lens is more slanted, so the ray makes a bigger angle with respect to the normal at the back surface. The bending at the back surface is therefore greater than the bending at the front surface, and the ray ends up being bent *outward* more than inward.



(b) Lens 2 must act the same as lens 1. It's diverging. One way of knowing this is time-reversal symmetry: if we flip the original figure over and then reverse the direction of the ray, it's still a valid diagram.

Lens 3 is diverging like lens 1 on top, and diverging like lens 2 on the bottom. It's a diverging lens.

As for lens 4, any close-up diagram we draw of a particular ray passing through it will look exactly like the corresponding close-up diagram for some part of lens 1. Lens 4 behaves the same as lens 1.

Page 254, problem 22:

Since d_o is much greater than d_i , the lens-film distance d_i is essentially the same as f . (a) Splitting the triangle inside the camera into two right triangles, straightforward trigonometry gives

$$\theta = 2 \tan^{-1} \frac{w}{2f}$$

for the field of view. This comes out to be 39° and 64° for the two lenses. (b) For small angles, the tangent is approximately the same as the angle itself, provided we measure everything in radians. The equation above then simplifies to

$$\theta = \frac{w}{f}$$

The results for the two lenses are $.70 \text{ rad} = 40^\circ$, and $1.25 \text{ rad} = 72^\circ$. This is a decent approximation.

(c) With the 28-mm lens, which is closer to the film, the entire field of view we had with the 50-mm lens is now confined to a small part of the film. Using our small-angle approximation $\theta = w/f$, the amount of light contained within the same angular width θ is now striking a piece of the film whose linear dimensions are smaller by the ratio $28/50$. Area depends on the square of the linear dimensions, so all other things being equal, the film would now be overexposed by a factor of $(50/28)^2 = 3.2$. To compensate, we need to shorten the exposure by a factor of 3.2.

Solutions for chapter 12

Page 277, problem 5:

You don't want the wave properties of light to cause all kinds of funny-looking diffraction effects. You want to see the thing you're looking at in the same way you'd see a big object. Diffraction effects are most pronounced when the wavelength of the light is relatively large compared to the size of the object the light is interacting with, so red would be the worst. Blue light is near the short-wavelength end of the visible spectrum, which would be the best.

Page 277, problem 6:

- (a) You can tell it's a single slit because of the double-width central fringe.
- (b) Four fringes on the top pattern are about 23.5 mm, while five fringes on the bottom one are about 14.5 mm. The spacings are 5.88 and 2.90 mm, with a ratio of 2.03. A smaller d leads to larger diffraction angles, so the width of the slit used to make the bottom pattern was almost exactly twice as wide as the one used to make the top one.

Page 278, problem 8:

For the size of the diffraction blob, we have:

$$\begin{aligned}\frac{\lambda}{d} &\sim \sin \theta \\ &\approx \theta \\ \theta &\sim \frac{700 \text{ nm}}{10 \text{ m}} \\ &\approx 10^{-7} \text{ radians}\end{aligned}$$

For the actual angular size of the star, the small-angle approximation gives

$$\begin{aligned}\theta &\sim \frac{10^9 \text{ m}}{10^{17} \text{ m}} \\ &= 10^{-8} \text{ radians}\end{aligned}$$

The diffraction blob is ten times bigger than the actual disk of the star, so we can never make an image of the star itself in this way.

Page 278, problem 9:

- (a) The patterns have two structures, a coarse one and a fine one. You can look up in the book which corresponds to w and which to d , or just use the fact that smaller features make bigger diffraction angles. The top and middle patterns have the same coarse spacing, so they have the same w . The fine structure in the top pattern has 7 fringes in 12.5 mm, for a spacing of 1.79 mm, while the middle pattern has 11 fringes in 41.5 mm, giving a spacing of 3.77 mm. The value of d for the middle pattern is therefore $(0.50 \text{ mm})(1.79/3.77) = 0.23 \text{ mm}$.
- (b) This one has about the same d as the top one (it's difficult to measure accurately because each group has only a small number of fringes), but the coarse spacing is different, indicating a different value of w . It has two coarse groupings in 23 mm, i.e., a spacing of 12.5 mm. The coarse groupings in the original pattern were about 23 mm apart, so there is a factor of two between the $w = 0.04 \text{ mm}$ of the top pattern and the $w = 0.08 \text{ mm}$ of the bottom one.

Page 279, problem 12:

The equation, solved for θ , is $\theta = \sin^{-1}(m\lambda/d)$. The sine function only ranges from -1 to $+1$, so the inverse sine is undefined for $|m\lambda/d| > 1$, i.e., $|m| > d/\lambda$. Physically, we only get fringes out to angles of 90 degrees (the inverse sine of 1) on both sides, corresponding to values of m less than d/λ .

Solutions for chapter 17

Page 423, problem 8:

The expressions $|\Psi|^2$ and $|\Psi^2|$ are identical, because the magnitude of a product is the product of the magnitudes. These expressions give positive real numbers as their results, which makes sense for a probability density. The expression Ψ^2 need not be real, and if it is real, it may be negative. It cannot be interpreted as a probability density. As a concrete example, suppose that $\Psi = bi$, where b is a real number with units. Then $|\Psi|^2 = |\Psi^2| = b^2$, which is real and positive, but $\Psi^2 = -b^2$, which clearly can't be interpreted probabilistically, because it's negative.

Answers to self-checks

Answers to self-checks for chapter 1

Page 17, self-check A:

The diagram for the house looks like | | because in the one dimension of space being represented, it has walls on both sides, and its existence also extends over a certain amount of time (left to right). If the dog is in the house at rest, then goes outside, and stays at rest in the back yard



for a while, the spacetime diagram looks like this: | | . An observer using another frame of reference has to agree that the dog went outside, because observers agree on intersections of world-lines, and the dog's world-line intersects the world-line of the house's back wall.

Page 18, self-check B:

For $d = 0$, the diagram becomes flat. The segments defining t and \mathcal{I} coincide, and we get $\mathcal{I} = t$. This makes sense, because that's what we expect to happen in the equation $\mathcal{I}^2 = t^2 - (\text{const.})d^2$, in the case where $d = 0$.

When $d < 0$, the triangle just flips over. By symmetry, we expect that the effect on time should be the same as for a positive d . This matches up in a sensible way with the behavior of the equation $\mathcal{I}^2 = t^2 - (\text{const.})d^2$. A negative d doesn't matter, because d appears as a square in the equation.

Page 24, self-check C:

If we turn the book sideways, slopes of world-lines represent velocities in natural units. World-line 1 has a speed less than c , 2 has speed c , and 3 has a speed greater than c . Only world-line 1 could be a material object, and only 2 could be a flash of light. Both 1 and 2 are straight lines, so both are inertial.

Answers to self-checks for chapter 2

Page 34, self-check A:

The leading edge is moving up, the trailing edge is moving down, and the top of the hump is motionless for one instant.

Page 41, self-check B:

(a) It doesn't have w or h in it. (b) Inertia is measured by μ , tightness by T . (c) Inertia would be measured by the density of the metal, tightness by its resistance to compression. Lead is more dense than aluminum, and this would tend to make the speed of the waves lower in lead. Lead is also softer, so it probably has less resistance to compression, and we would expect this to provide an additional effect in the same direction. Compressional waves will definitely be

slower in lead than in aluminum.

Answers to self-checks for chapter 3

Page 65, self-check A:

$\mathbf{E} \times \mathbf{E}$ or $\mathbf{B} \times \mathbf{B}$ are both zero, because the vector cross product vanishes when the vectors being multiplied lie along the same line.

Answers to self-checks for chapter 4

Page 94, self-check A:

The equation $D = \sqrt{(1+v)/(1-v)}$ is expressed in natural units. In SI units, it wouldn't make sense to add v , with units of m/s, to a unitless 1.

Page 94, self-check B:

For $v = -1$, we have $D = 0$. Multiplying this by any finite D will still give $D = 0$. Thus $v = -1$ is also invariant.

Page 98, self-check C:

When $v = 0$, the contracted length $L\sqrt{1-v^2}$ is the same as L . This makes sense. If $v = 0$, then there is no difference between the object's rest frame and the frame in which we're observing it.

Answers to self-checks for chapter 5

Page 115, self-check A:

When $\alpha = 1$, we have $R = 0$ and $T = 1$. This makes sense physically because $\alpha = 1$ means that the media have the same properties, so it's as if we haven't encountered a change of medium at all.

Page 116, self-check B:

The energy of a wave is usually proportional to the square of its amplitude. Squaring a negative number gives a positive result, so the energy is the same.

Page 119, self-check C:

The wave pattern will look like this:  Three quarters of a wavelength fit in the tube, so the wavelength is three times shorter than that of the lowest-frequency mode, in which one quarter of a wave fits. Since the wavelength is smaller by a factor of three, the frequency is three times higher. Instead of $f_0, 2f_0, 3f_0, 4f_0, \dots$, the pattern of wave frequencies of this air column goes $f_0, 3f_0, 5f_0, 7f_0, \dots$

Page 127, self-check D:

Say we're looking for $u = \sqrt{z}$, i.e., we want a number u that, multiplied by itself, equals z . Multiplication multiplies the magnitudes, so the magnitude of u can be found by taking the square root of the magnitude of z . Since multiplication also adds the arguments of the numbers, squaring a number doubles its argument. Therefore we can simply divide the argument of z by two to find the argument of u . This results in one of the square roots of z . There is another one, which is $-u$, since $(-u)^2$ is the same as u^2 . This may seem a little odd: if u was chosen so that doubling its argument gave the argument of z , then how can the same be true for $-u$? Well for example, suppose the argument of z is 4° . Then $\arg u = 2^\circ$, and $\arg(-u) = 182^\circ$. Doubling 182 gives 364, which is actually a synonym for 4 degrees.

Page 132, self-check E:

Only $\cos(6t - 4)$ can be represented by a complex number. Although the graph of $\cos^2 t$ does have a sinusoidal shape, it varies between 0 and 1, rather than -1 and 1 , and there is no way to represent that using complex numbers. The function $\tan t$ doesn't even have a sinusoidal shape.

Page 134, self-check F:

Energy is proportional to the square of the amplitude, so its energy is four times smaller after every cycle. It loses three quarters of its energy with each cycle.

Answers to self-checks for chapter 6

Page 148, self-check A:

In SI units, joules are equal to $\text{kg}\cdot\text{m}^2/\text{s}^2$, so to make the equation work in SI units, we need to write it as $m = E/c^2$. I.e., we just take the famous $E = mc^2$ and solve it for m .

Page 148, self-check B:

The flash of light that was twice as bright would also be on the diagonal, but twice as far from the origin. That is, the original energy-momentum vector \mathbf{p} would change to $2\mathbf{p}$.

Page 148, self-check C:

It makes sense if $v = p/E$, because an object at rest has $p = 0$ but $E \neq 0$ because its energy equals its mass. It wouldn't make sense to have $v = E/p$, because then for an object at rest with $p = 0$, we would have $v = \infty$.

Page 150, self-check D:

To make the units come out right in SI, we need to change $m^2 = E^2 - p^2$ to $m^2c^4 = E^2 - p^2c^2$. Of course there are multiple ways of expressing the same fact, e.g., we could divide by c^4 on both sides.

Answers to self-checks for chapter 7

Page 170, self-check A:

(1) Most people would think they were positively correlated, but it's possible that they're independent. (2) These must be independent, since there is no possible physical mechanism that could make one have any effect on the other. (3) These cannot be independent, since dying today guarantees that you won't die tomorrow.

Answers to self-checks for chapter 8

Page 186, self-check A:

Solids can exert shear forces. A solid could be in an equilibrium in which the shear forces were canceling the forces due to unequal pressures on the sides of the cube.

Page 187, self-check B:

(1) Not valid. The equation only applies to fluids. (2) Valid. The density of the air is nearly constant between the top and bottom of the building. (3) Not valid. There is a large difference in the density of the air between the top and the bottom of the mountain. (4) Not valid, because g isn't constant throughout the interior of the earth. (5) Not valid, because the air is flowing around the wing. The air is accelerating, so it is not in equilibrium.

Answers to self-checks for chapter 10

Page 219, self-check A:

You should have found from your ray diagram that an image is still formed, and it has simply moved down the same distance as the real face. However, this new image would only be visible from high up, and the person can no longer see his own image.

Page 223, self-check B:

Increasing the distance from the face to the mirror has decreased the distance from the image to the mirror. This is the opposite of what happened with the virtual image.

Answers to self-checks for chapter 11

Page 238, self-check A:

At the top of the graph, d_i approaches infinity when d_o approaches f . Interpretation: the rays just barely converge to the right of the mirror.

On the far right, d_i approaches f as d_o approaches infinity; this is the definition of the focal length.

At the bottom, d_i approaches negative infinity when d_o approaches f from the other side. Interpretation: the rays don't quite converge on the right side of the mirror, so they appear to have come from a virtual image point very far to the left of the mirror.

Page 243, self-check B:

(1) In 1, the rays cross the image, so it's real. In 2, the rays only appear to have come from the image point, so the image is virtual. (2) A ray is always closer to the normal in the medium with the higher index of refraction. The first left turn makes the ray closer to the normal, which is what should happen in glass. The second left turn makes the ray farther from the normal, and that's what should happen in air. (3) Take the topmost ray as an example. It will still take two right turns, but since it's entering the lens at a steeper angle, it will also leave at a steeper angle. Tracing backward to the image, the steeper lines will meet closer to the lens.

Answers to self-checks for chapter 12

Page 265, self-check A:

It would have to have a wavelength on the order of centimeters or meters, the same distance scale as that of your body. These would be microwaves or radio waves. (This effect can easily be noticed when a person affects a TV's reception by standing near the antenna.) None of this contradicts the correspondence principle, which only states that the wave model must agree with the ray model when the ray model is applicable. The ray model is not applicable here because λ/d is on the order of 1.

Page 267, self-check B:

At this point, both waves would have traveled nine and a half wavelengths. They would both be at a negative extreme, so there would be constructive interference.

Page 271, self-check C:

Judging by the distance from one bright wave crest to the next, the wavelength appears to be about $2/3$ or $3/4$ as great as the width of the slit.

Page 272, self-check D:

Since the wavelengths of radio waves are thousands of times longer, diffraction causes the resolution of a radio telescope to be thousands of times worse, all other things being equal. (To compensate for the wavelength, it's desirable to make the telescope very large, as in figure z on page 272.)

(1 rectangle = $5 \text{ cm} \times 0.005 \text{ cm}^{-1} = 0.025$), but that would have been pointless, because we were just going to compare the two areas.

Answers to self-checks for chapter 13

Page 308, self-check A:

Thomson was accelerating electrons, which are negatively charged. This apparatus is supposed to accelerate atoms with one electron stripped off, which have positive net charge. In both cases, a particle that is between the plates should be attracted by the forward plate and repelled

by the plate behind it.

Page 316, self-check B:

The hydrogen-1 nucleus is simply a proton. The binding energy is the energy required to tear a nucleus apart, but for a nucleus this simple there is nothing to tear apart.

Answers to self-checks for chapter 14

Page 326, self-check A:

The area under the curve from 130 to 135 cm is about 3/4 of a rectangle. The area from 135 to 140 cm is about 1.5 rectangles. The number of people in the second range is about twice as much. We could have converted these to actual probabilities (1 rectangle = $5 \text{ cm} \times 0.005 \text{ cm}^{-1} = 0.025$), but that would have been pointless because we were just going to compare the two areas.

Answers to self-checks for chapter 15

Page 348, self-check A:

The axes of the graph are frequency and photon energy, so its slope is Planck's constant. It doesn't matter if you graph $e\Delta V$ rather than $W + e\Delta V$, because that only changes the y-intercept, not the slope.

Answers to self-checks for chapter 16

Page 369, self-check A:

Wavelength is inversely proportional to momentum, so to produce a large wavelength we would need to use electrons with very small momenta and energies. (In practical terms, this isn't very easy to do, since ripping an electron out of an object is a violent process, and it's not so easy to calm the electrons down afterward.)

Page 381, self-check B:

Under the ordinary circumstances of life, the accuracy with which we can measure position and momentum of an object doesn't result in a value of $\Delta p \Delta x$ that is anywhere near the tiny order of magnitude of Planck's constant. We run up against the ordinary limitations on the accuracy of our measuring techniques long before the uncertainty principle becomes an issue.

Answers to self-checks for chapter 17

Page 395, self-check A:

No. The equation $KE = p^2/2m$ is nonrelativistic, so it can't be applied to an electron moving at relativistic speeds. Photons always move at relativistic speeds, so it can't be applied to them, either.

Page 397, self-check B:

Dividing by Planck's constant, a small number, gives a large negative result inside the exponential, so the probability will be very small.

Answers to self-checks for chapter 18

Page 426, self-check A:

If you trace a circle going around the center, you run into a series of eight complete wavelengths. Its angular momentum is $8\hbar$.

Answers

Answers for chapter 2**Page 61, problem 19:**

(a) $T = \mu\omega^2 r^2$

Answers for chapter 5**Page 140, problem 4:**

Check: The actual length of a flute is about 66 cm.

Page 141, problem 7:

(b) $g/2$

Page 143, problem 10:

(a) $f = 4\alpha/(1 + \alpha)^2$ (b) $v_2 = \sqrt{v_1 v_3}$

Page 143, problem 10:

(a) $f = 4\alpha/(1 + \alpha)^2$ (b) $v_2 = \sqrt{v_1 v_3}$

Answers for chapter 9**Page 209, problem 1:**(a) $\sim 2 - 10\%$ (b) 5% (c) The high end for the body's actual efficiency is higher than the limit imposed by the laws of thermodynamics. However, the high end of the 1-5 watt range quoted in the problem probably includes large people who aren't just lying around. Still, it's impressive that the human body comes so close to the thermodynamic limit.**Answers for chapter 11****Page 255, problem 27:**

f/ϵ

Page 257, problem 31:

$P = (1/2)(n^2 - 1)$

Answers for chapter 15**Page 362, problem 7:**about 10^{-34}

Photo Credits

Except as specifically noted below or in a parenthetical credit in the caption of a figure, all the illustrations in this book are under my own copyright, and are copyleft licensed under the same license as the rest of the book.

In some cases it's clear from the date that the figure is public domain, but I don't know the name of the artist or photographer; I would be grateful to anyone who could help me to give proper credit. I have assumed that images that come from U.S. government web pages are copyright-free, since products of federal agencies fall into the public domain. I've included some public-domain paintings; photographic reproductions of them are not copyrightable in the U.S. (Bridgeman Art Library, Ltd. v. Corel Corp., 36 F. Supp. 2d 191, S.D.N.Y. 1999).

When "PSSC Physics" is given as a credit, it indicates that the figure is from the first edition of the textbook entitled Physics, by the Physical Science Study Committee. The early editions of these books never had their copyrights renewed, and are now therefore in the public domain. There is also a blanket permission given in the later PSSC College Physics edition, which states on the copyright page that "The materials taken from the original and second editions and the Advanced Topics of PSSC PHYSICS included in this text will be available to all publishers for use in English after December 31, 1970, and in translations after December 31, 1975."

Credits to Millikan and Gale refer to the textbooks Practical Physics (1920) and Elements of Physics (1927). Both are public domain. (The 1927 version did not have its copyright renewed.) Since it is possible that some of the illustrations in the 1927 version had their copyrights renewed and are still under copyright, I have only used them when it was clear that they were originally taken from public domain sources.

In a few cases, I have made use of images under the fair use doctrine. However, I am not a lawyer, and the laws on fair use are vague, so you should not assume that it's legal for you to use these images. In particular, fair use law may give you less leeway than it gives me, because I'm using the images for educational purposes, and giving the book away for free. Likewise, if the photo credit says "courtesy of ...," that means the copyright owner gave me permission to use it, but that doesn't mean you have permission to use it.

?? *Huygens*: Contemporary painting?. ?? *Hindenburg*: Public domain product of the U.S. Navy. **13**
Football pass: From a photo by Wikipedia user RMelon, CC-BY-SA. **14** *Atomic clock on plane*: Copyright 1971, Associated press, used under U.S. fair use exception to copyright law. **16** *Hell*: Hieronymus Bosch, public domain. **16** *Cow*: Redrawn from a photo by Wikimedia Commons user Cgoodwin, CC-BY-SA. **16** *Car*: Redrawn from a photo by Wikimedia Commons user Auregann, CC-BY-SA. **19** *Horse*: From a public-domain photo by Eadweard Muybridge, 1872. **19** *Satellite*: From a public-domain artist's conception of a GPS satellite, product of NASA. **26** *Ripples*: Scott Robinson, CC-BY. **27** *Canis Major*: Wikimedia Commons user nubobo, CC-BY. **31** *Electric bass*: Brynjar Vik, CC-BY license. **48** *X-15 shock wave*: NASA, public domain. **49** *Reflection of ripples*: Wikimedia Commons user MikeRun, CC-BY-SA. **53** *Car crash*: Wikipedia user Pso, GFDL. **53** *Guitar*: Wikipedia user "Denis Diderot," CC-BY-SA. **85** *Pushmi-pully*: Hugh Lofting, 1920, public domain. **116** *Reflection of pulses*: PSSC Physics. **120** *Water wave refracting*: Original photo from PSSC. **124** *Ulcer*: Wikipedia user Aspersions, CC-BY-SA. **167** *Hot air balloon*: Randy Oostdyk, CC-BY-SA licensed. **167** *Humpty Dumpty*: L. Leslie Brooke, publication date unknown; second image believed PD, source unknown. **199** *Carnot*: contemporary. **206** *Space junk*: STK-generated images courtesy of CSSI (www.centerforspace.com). **219** *Praxinoscope*: Thomas B. Greenslade, Jr.. **224** *Flower*: Based on a photo by Wikimedia Commons user Fir0002, CC-BY-SA. **224** *Moon*: Wikimedia commons image. **247** *Fish-eye lens*: Martin D'urrschnabel, CC-BY-SA. **249** *Hubble space telescope*: NASA, public domain. **258** *Anamorphic image*: Wikipedia user Istvan Orosz, CC-BY. **261** *Counterfactual lack of diffraction of water waves*: Assembled from photos in PSSC. **261** *Diffraction of water waves*: Assembled from photos in PSSC. **263** *Dorothy Hodgkin*: 1964 Nobel Prize portrait. Used here under the U.S. fair use doctrine.. **263** *Lysozyme crystals*: Wikipedia user Zanecrc, CC-BY-SA. **263** *X-ray diffraction pattern*: Wikipedia user Del45, public domain. **263** *Structure of lysozyme*: The Protein Data Bank H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) Nucleic Acids Research, 28: 235-242. doi:10.1093/nar/28.1.235, stated on the web page to be "free for use". **264** *Scaling of diffraction*: Assembled from photos in PSSC. **265** *Diffraction of water waves*: Assembled from photos in PSSC. **266** *Young*: Wikimedia Commons, "After a portrait by Sir Thomas Lawrence, From: Arthur Shuster & Arthur E. Shipley: Britain's Heritage of Science. London, 1917". **271** *Single-slit diffraction of water waves*: PSSC. **271** *Simulation of a single slit using three sources*: PSSC. **272** *Pleiades*: NASA/ESA/AURA/Caltech, public domain. **272** *Radio telescope*: Wikipedia user Hajor, CC-BY-SA. **278** *Pleiades*: NASA/ESA/AURA/Caltech, public domain. **290** *Thomson*: Harper's Monthly, 1904. **297** *nuclear fuel pellets*: US DOE, public domain. **309** *nuclear power plant*: Wikipedia user Stefan Kuhn,

CC-BY-SA licensed. **315 GAMMASPHERE**: Courtesy of C.J. Lister and R.V.F. Janssens. **315 H bomb test**: public domain product of US DOE, Ivy Mike test. **315 Fatu Hiva Rainforest**: Wikipedia user Makemake, CC-BY-SA licensed. **315 fusion reactor**: “These images may be used free of charge for educational purposes but please use the acknowledgement ‘photograph courtesy of EFDA-JET’”. **315 sun**: SOHO (ESA and NASA). **318 Chernobyl map**: CIA Handbook of International Economic Statistics, 1996, public domain. **321 Horses**: (c) 2004 Elena Filatova. **321 Polar bear**: U.S. Fish and Wildlife Service, public domain. **331 Technetium injection**: Wikimedia Commons user Bionerd, CC-BY-SA. **365 Wicked witch**: W.W. Denslow, 1900. Quote from The Wizard of Oz, by L. Frank Baum, 1900. **373 Brain**: Based on a drawing by Patrick J. Lynch, CC-BY.. **425 Benzene**: Wikimedia Commons user Benjah-bmm27, public domain.. **437 Fermi and Bose**: Public domain.

Index

- Q* (quality factor), 136
 mechanical oscillator, 134
- aberration, 246
 chromatic, 244
- absorption, 377
- adiabatic, 193
- adiabatic gas constant, 193
- adiabatic index, *see* adiabatic gas constant
- aether, 78
- alpha decay, 311
 nature of emitted particle, 297
- alpha particle, *see* alpha decay
- anamorph, 258
- angular magnification, 224
- angular momentum
 and the uncertainty principle, 426
 in three dimensions, 426
 quantization of, 425
- antielectron, 313
- antimatter, 313
- atom
 raisin-cookie model of, 292
- atomic clock, 14
- atomic number
 defined, 301
- atoms
 helium, 437
 lithium, 437
 with many electrons, 437
- averages, 169
 rule for calculating, 169
- Avogadro's number, 176
- Balmer, Johann, 387
- basis for a vector space, 156
- basis of a vector space, 156
- beta decay, 313
 nature of emitted particle, 297
- beta particle, *see* beta decay
- binding energy
 nuclear, 316
- binomial coefficient, 178
- Bohr
- Niels, 264
- bond, *see* chemical bonds
- Bose, Satyendra Nath, 438
- boson, 438
- bottomonium, 391
- bound states, 376
- box
 particle in a, 376
- bra-ket notation, 372
- c*
 as speed limit for cause and effect, 24, 92
 as the speed of light, 22, 71
 as the universal speed limit, 20
 invariance, 21, 71, 94
- carbon-14 dating, 332
- Carnot engine, 200
- cathode rays, 288
- Celsius (unit), 174
- chain reaction, 311
- charge
 fundamental (*e*), 287
- charmonium, 391
- chemical bonds
 quantum explanation for hydrogen, 378
- Chernobyl, 318
- climate change, 321
- comet, 66
- completeness (quantum physics), 430
- complex numbers, 125
 in quantum physics, 403
- converging, 221
- correspondence principle, 264
 defined, 19
 for time dilation, 19
- D* (diagonal stretch factor in relativity), *see*
 stretch factor *D* in relativity
- Davisson
 C.J., 366
- de Broglie
 Louis, 366
- decay
 exponential, 329

- decoherence, 358, 381
- degeneracy, 52, 435
- degree of freedom, 173
- diffraction
 - defined, 262
 - double-slit, 266
 - fringe, 263
 - scaling of, 264
 - single-slit, 271
- diffraction grating, 271
- diffuse reflection, 214
- digital camera, 344
- dimension of a vector space, 156
- diopter, 236
- dispersion, 70, 136, 244, 373
- dissonance, 119
- DNA, 318
- Doppler shift, 47
 - relativistic, 95
- double-slit diffraction, 266
- duality
 - wave-particle, 351
- Dulong-Petit law, 174, 181
- Einstein, Albert, 343
- electromagnetic fields
 - energy and momentum density, 65
- electromagnetic wave
 - energy, 68
 - geometry, 68
 - momentum, 73
 - propagation at c , 70
- electromagnetism, 102
- electron, 291
 - as a wave, 366
 - wavefunction, 369
- electron capture, 313
- electron decay, 313
- emission spectrum, 377
- endoscope, 124
- energy
 - equivalence to mass, 76
 - of electromagnetic fields, 65
 - quantization of for bound states, 377
- engine
 - Carnot, 200
- entanglement, 356
 - of macroscopic objects, 359
- entropy
 - macroscopic definition, 201
 - microscopic definition, 204
- Euler's formula, 129
- Euler, Leonhard, 129
- exclusion principle, 437
- exponential decay, 329
- Fermi, Enrico, 437
- fermion, 437
- field
 - inertia of, 75
- focal angle, 234
- focal length, 235
- focal point, 235
- Fourier's theorem, 45
- fourier-spectra, 119
- frame of reference
 - none preferred in relativity, 104
- frequency
 - of waves, 42
- fringe
 - diffraction, 263
- fundamental, 119
- fundamental charge e , 287
- fundamental theorem of algebra, 128
- FWHM (full width at half maximum), 136
- γ , *see* gamma factor, *see* adiabatic gas constant
- Galilean transformation, 89
- gamma decay
 - nature of emitted particle, 297
- gamma factor γ in relativity, 98
- gamma ray, *see* gamma decay
- gas
 - spectrum of, 377
- Germer, L., 366
- Gisin's theorem, 409
- global warming, 321
- goiters, 329
- group velocity, 36, 137, 375
- Hafele-Keating experiment, 14
- half-life, 329
- Halley's comet, 66
- harmonics, 119
- heat, 191
- Heisenberg

Werner, 379
Heisenberg uncertainty principle, 379
 in three dimensions, 426
helium, 437
Helmholtz resonator, 194
Hertz, Heinrich, 71, 346
 Heinrich, 266
Hiroshima, 319
hormesis, 320
Huygens' principle, 265
hydrogen atom
 energies of states in, 384
hydrogen molecule, *see* chemical bonds

ideal gas law, 176
images
 formed by curved mirrors, 221
 formed by plane mirrors, 218
 location of, 233
 of images, 223
 real, 222
 virtual, 218
incoherent light, 262
independence
 statistical, 168
independent probabilities
 law of, 168
index of refraction, 122
inner product
 quantum mechanics, 409
 relativistic four-vectors, 156
interval, spacetime, 17
invariant quantities, 21
iodine, 329
isotopes, 308
Ives-Stilwell experiments, 97
kelvin (unit), 174

Laplacian, 54
length contraction, 98
lens, 242
lensmaker's equation, 243
lepton, 335
lepton number, 335
light
 electromagnetic wave, 22
 particle model of, 214
 ray model of, 214

speed, 22
wave model of, 214
light cone, 91
linear independence, 156
linear no-threshold, 320
linear operator, 156
Lipkin linkage, 258
LNT, 320
Lorentz transformation, 89

magnetism
 related to electricity, 102
magnification
 angular, 224
 by a converging mirror, 221
Maxwell, James Clerk, 266
median, 327
mirror
 converging, 233
molecules
 nonexistence in classical physics, 365
momentum
 of electromagnetic fields, 66

natural units, 22
neutron
 spin of, 433
Newton, Isaac
 Newtonian telescope, 223
 particle theory of light, 265
Nichols-Hull experiment on momentum of light, 67
normalization
 discrete probability, 169
 probability distribution, 326
nuclear forces, 310
nucleus
 discovery, 299

observable
 quantum
 operator, 412
operational definition, 188
operator
 for a quantum observable, 412
ozone layer, 343

partial derivative, 40
particle

definition of, 351
particle in a box, 376
particle model of light, 214, 266
pascal
 unit, 184
path of a photon undefined, 352
Pauli exclusion principle, 437
Peaucellier linkage, 258
period
 of waves, 42
periodic table, 301
phase in quantum mechanics
 not observable, 370, 408
phase space, 205
phase velocity, 36, 137, 375
photoelectric effect, 346
photon
 Einstein's early theory, 345
 energy of, 348
 in three dimensions, 359
 spin of, 433
Planck's constant, 348
Planck, Max, 348
positron, 77, 313
positron decay, 313
Poynting vector, 73
Poynting, John Henry, 73
praxinoscope, 219
probabilities
 addition of, 168
 normalization of, 169
probability distribution, 326
probability interpretation, 352
proton
 spin of, 433
quality factor (Q)
 mechanical oscillator, 134
quantum dot, 376
quantum moat, 425
quantum number, 372, 429
quark, 391
radar, 343
radiation hormesis, 320
radio, 343
raisin cookie model, 292
rapidity, 105
ray diagrams, 216
ray model of light, 214, 266
reflection
 diffuse, 214
refraction, 121
relativity
 general, 25
 origin of the term, 25
 special, 25
repetition of diffracting objects, 270
resonance, 134
retina, 223
RHIC accelerator, 99
Rutherford
 characterization of alpha particles, 297
 discovery of nucleus, 299
Schrödinger equation, 394
separability, 53
Sievert (unit), 318
simple harmonic motion, 130
simultaneity, 86
single-slit
 diffraction, 271
Sirius, 377
Snell's law, 121
Sommerfeld, Arnold, 181
sound
 speed of, 194
 waves, 42
specific heat
 electrons' contribution, 181
 solids, 174
spectrum
 absorption, 377
 emission, 377
spherical harmonics, 428
spin
 neutron's, 433
 photon's, 433
 proton's, 433
standing wave, 49
Star Trek, 377
states
 bound, 376
Stirling's formula, 178
stretch factor D in relativity, 94
strong nuclear force, 310
strong nuclear force, 310

superposition
 of waves, 32

synchronization, 86

telescope, 223, 272

temperature
 absolute zero, 174
 celsius, 174
 kelvin, 174
 macroscopic definition, 190
 microscopic definition, 174

thermodynamics
 first law of, 207
 laws of
 summarized, 207
 second law of, 203, 207
 third law of, 207
 zeroth law of, 189, 207

thermometer, 190

Thomson, J.J.
 cathode ray experiments, 290

time, 13
 not absolute, 15
 symmetry of the laws of physics under reversal, 13

total internal reflection, 124

tunneling, 394

ultraviolet light, 343

uncertainty principle, 379
 in three dimensions, 426

unitary evolution of the wavefunction, 415

unitary operator, 420

units
 natural relativistic, 22

vector space, 155

velocity
 combination in relativity, 94
 group, 375
 phase, 375

wave
 definition of, 351
 dispersive, 244, 373

wave model of light, 214, 266

wave-particle duality, 351
 probability interpretation of, 352

wave-vector, 51

wavefunction
 complex numbers in, 403
 of the electron, 369

wavelength, 44

wavenumber, 45

waves
 frequency of, 42
 medium not transported with, 34
 on a string, 38
 patterns, 36
 period of, 42
 sound, 42
 superposition of, 32
 velocity of, 35
 wavelength, 44

weak nuclear force, 312

Wicked Witch of the West, 365

Wigner, Eugene, 233

world-line, 17

Young, Thomas, 266

Useful Data

Metric Prefixes

M-	mega-	10^6
k-	kilo-	10^3
m-	milli-	10^{-3}
μ - (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}
p-	pico-	10^{-12}
f-	femto-	10^{-15}

(Centi-, 10^{-2} , is used only in the centimeter.)

Notation and Units

quantity	unit	symbol
distance	meter, m	$x, \Delta x$
time	second, s	$t, \Delta t$
mass	kilogram, kg	m
density	kg/m^3	ρ
velocity	m/s	\mathbf{v}
acceleration	m/s^2	\mathbf{a}
force	$\text{N} = \text{kg}\cdot\text{m}/\text{s}^2$	\mathbf{F}
pressure	$\text{Pa} = 1 \text{ N}/\text{m}^2$	P
energy	$\text{J} = \text{kg}\cdot\text{m}^2/\text{s}^2$	E
power	$\text{W} = 1 \text{ J}/\text{s}$	P
momentum	$\text{kg}\cdot\text{m}/\text{s}$	\mathbf{p}
angular momentum	$\text{kg}\cdot\text{m}^2/\text{s}$ or $\text{J}\cdot\text{s}$	L
period	s	T
wavelength	m	λ
frequency	s^{-1} or Hz	f
gamma factor	unitless	γ
probability	unitless	P
prob. distribution	various	D
electron wavefunction	$\text{m}^{-3/2}$	Ψ

The Greek Alphabet

α	A	alpha	ν	N	nu
β	B	beta	ξ	Ξ	xi
γ	Γ	gamma	\circ	O	omicron
δ	Δ	delta	π	Π	pi
ϵ	E	epsilon	ρ	P	rho
ζ	Z	zeta	σ	Σ	sigma
η	H	eta	τ	T	tau
θ	Θ	theta	v	Y	upsilon
ι	I	iota	ϕ	Φ	phi
κ	K	kappa	χ	X	chi
λ	Λ	lambda	ψ	Ψ	psi
μ	M	mu	ω	Ω	omega

Subatomic Particles

particle	mass (kg)	radius (fm)
electron	9.109×10^{-31}	$\lesssim 0.01$
proton	1.673×10^{-27}	~ 1.1
neutron	1.675×10^{-27}	~ 1.1

The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about a million fm in radius.

Earth, Moon, and Sun

body	mass (kg)	radius (km)	radius of orbit (km)
earth	5.97×10^{24}	6.4×10^3	1.49×10^8
moon	7.35×10^{22}	1.7×10^3	3.84×10^5
sun	1.99×10^{30}	7.0×10^5	—

Fundamental Constants

gravitational constant	$G = 6.67 \times 10^{-11} \text{ N}\cdot\text{m}^2/\text{kg}^2$
Coulomb constant	$k = 8.99 \times 10^9 \text{ N}\cdot\text{m}^2/\text{C}^2$
quantum of charge	$e = 1.60 \times 10^{-19} \text{ C}$
speed of light	$c = 3.00 \times 10^8 \text{ m}/\text{s}$
Planck's constant	$h = 6.63 \times 10^{-34} \text{ J}\cdot\text{s}$