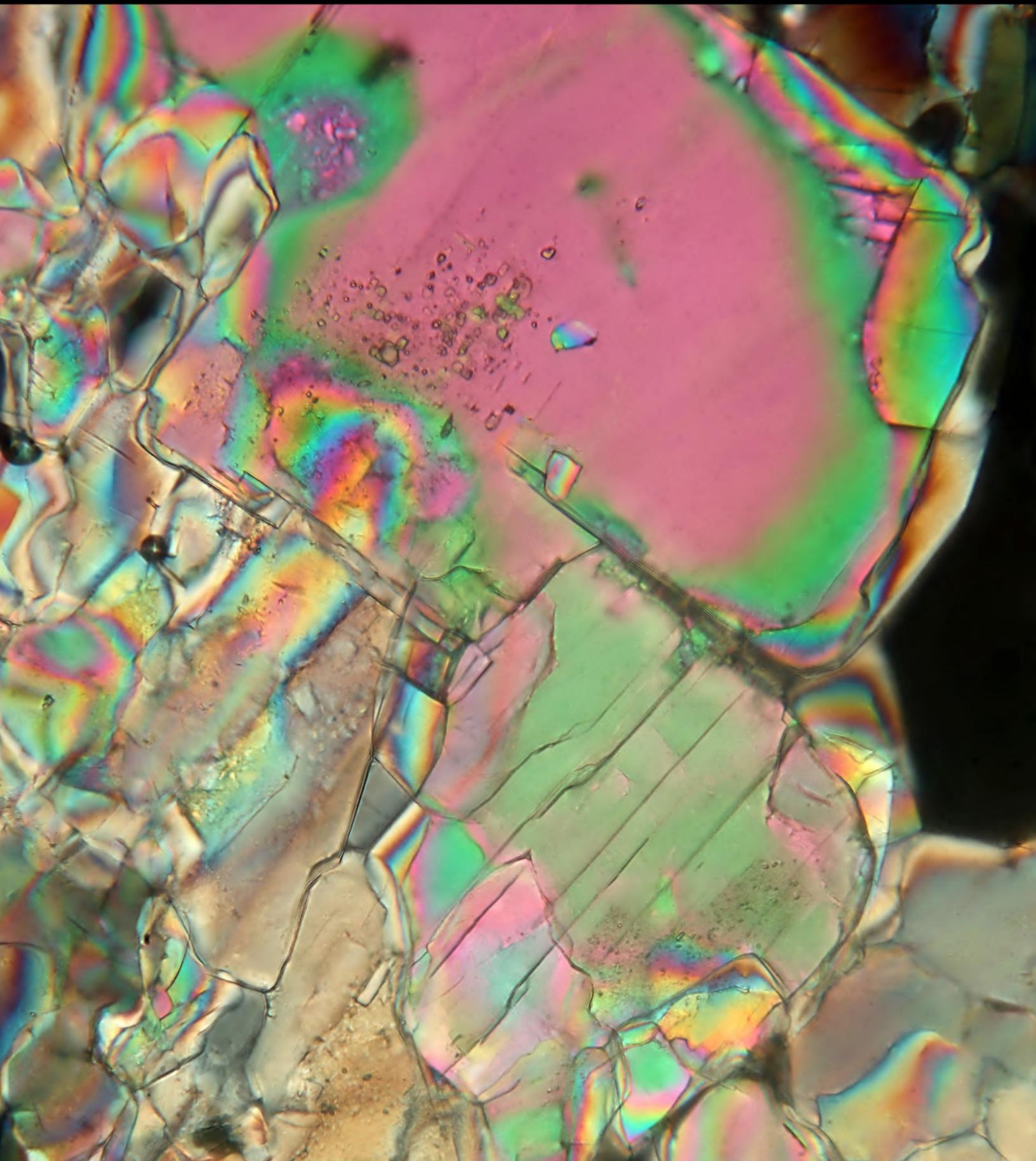


Fields and Circuits

Crowell



Fields and Circuits



Light and Matter

Fullerton, California

www.lightandmatter.com

copyright 2018 Benjamin Crowell

rev. November 10, 2020



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.

Brief Contents

Preface for the student	10
Preface for the instructor	11
<i>Electric and magnetic fields</i>	
1 The electric and magnetic fields	15
2 Gauss's law	41
3 Models of matter	73
4 The electric potential	85
5 Electromagnetism	119
6 Radiation	153
7 ★ More about relativity (optional)	175
<i>DC circuits</i>	
8 Electrical resistance	191
9 Parallel and series circuits	209
<i>Iterated integrals</i>	
10 Iterated integrals	233
<i>Sources of magnetism</i>	
11 Sources of magnetism	257
<i>AC circuits</i>	
12 Review of oscillations, resonance, and complex numbers	291
13 AC circuits	309
14 Impedance	331
<i>Stokes's theorem</i>	
15 Stokes's theorem	347
<i>Electromagnetic properties of materials</i>	
16 Electromagnetic properties of materials	361
<i>Relativity</i>	
17 Relativity (optional stand-alone chapter)	377

Contents

Preface for the student	10
Preface for the instructor	11
Electric and magnetic fields	
1 Electric and magnetic fields	
1.1 A surprising link to the nature of time	15
1.2 Basic properties of fields	18
1.3 Review of vectors	21
Definition of vectors and scalars, 21.— Components of vectors, 22.—Vector addition, 22.—Unit vector notation, 23.— Rotational invariance, 24.—Dot and cross product, 25.	
1.4 Field lines	26
1.5 Energy in fields and measurement of fields	27
Energy in fields, 27.—Units of the fields, 28.—Defining the magnitude and direction of a field, 29.	
Notes for chapter 1	32
Problems	33
Minilab 0: Magnetic interactions and en- ergy	37
Minilab 1: Magnetic field of a bar magnet as a function of distance	38
Exercise 1: Visualizing superposed fields	39
2 Gauss's law	
2.1 Gauss's law	41
Resolving the flip-the-arrowheads am- biguity, 41.—Sources and sinks, 42.— Gauss's law for field lines, in a vacuum, 43.	
2.2 A global form of Gauss's law	46
2.3 Field of a point charge at rest	50
2.4 Electric force on a charge	52
2.5 Coulomb's law	53
2.6 Charge	54
Gauss's law, not in vacuum, 54.— Invariance of charge, 56.—Quantization of charge, 56.—Conservation of charge, 57.	
2.7 Gauss's law when the number of di- mensions is effectively less than three . . .	58
2.8 Gauss's law in local form, with field vectors	60
Notes for chapter 2	64
Problems	67
Minilab 2: Gauss's law for magnetism . .	71
3 Models of matter	
3.1 Binding energy of matter: two exam- ples	73
3.2 The discovery of the electron, and the raisin cookie model	76
3.3 The energy scale for chemistry and atomic physics	78
3.4 The nucleus and the planetary model	79
3.5 The energy scale for nuclear physics	80
Notes for chapter 3	81
Problems	82
4 The electric potential	
4.1 Something is missing	85
4.2 The electric potential	89
4.3 Constant potential throughout a con- ductor	91
4.4 Potential related to field	93
One dimension, 93.—Two or three dimen- sions, 97.	
4.5 Summary of div, grad, and curl	98
4.6 Boundary conditions on a conductor	99
No component of the electric field parallel to the surface, 99.—The method of images, 100.	
Notes for chapter 4	101
Problems	103
Minilab 4A: Mapping electric fields . .	112
Exercise 4B: A preview of the electric po- tential and measurement of voltages . .	114
Exercise 4C: Charge density, field, and potential	115
Minilab 4D: Testing the curliness of the electric field	116
5 Electromagnetism	
5.1 Current and magnetic fields	119

5.2 Energy, pressure, tension, and momentum in fields	123
Momentum, 123.—Pressure and tension, 126.	
5.3 Force of a magnetic field on a charge	131
5.4 The dipole	132
5.5 $E=mc^2$	136
Fields carry inertia, 136.—Equivalence of mass and energy, 137.—A naughty infinity, 139.—Summary of relativity, 140.	
Notes for chapter 5	141
Problems	143
Minilab 5: The dipole and superposition	150
Exercise 5: Tension in the electric field .	152
6 Radiation	
6.1 Wave patterns	154
6.2 What it is that waves	155
6.3 Geometry of a plane wave	156
E and B perpendicular to the direction of propagation, 156.— E and B equal in energy, 156.— E and B perpendicular to each other, 157.	
6.4 Propagation at a fixed velocity	158
6.5 The electromagnetic spectrum	159
6.6 Momentum and rate of energy flow	161
Momentum of a plane wave, 161.—Rate of energy flow, 161.	
6.7 Maxwell's equations in a vacuum	163
6.8 Einstein's motorcycle	166
Notes for chapter 6	167
Problems	169
Exercise 6A: Polarization	171
Exercise 6B: Maxwell's equations applied to a plane wave	172
Minilab 6: A polarizing filter	173
7 *More about relativity (optional)	
7.1 Einstein's motorcycle: the resolution	175
7.2 Implications for the structure of space and time	178
Combination of velocities, 178.—Length contraction, 180.—Time dilation, 182.	
Notes for chapter 7	185
Problems	186

DC circuits

8 Electrical resistance

8.1 Circuits	191
Complete circuits, open circuits, 191.—Measuring the current in a circuit, 191.	
8.2 Power	194
8.3 Resistance	195
Resistance, 195.—Superconductors, 197.—Short circuits, 198.—Resistors, 198.	
8.4 Flow of energy	199
Notes for chapter 8	202
Problems	203
Exercise 8: Measuring voltage, current, and resistance	206
Minilab 8: Electrical measurements	208

9 DC circuits

9.1 Schematics	209
9.2 Parallel resistances and the junction rule	210
9.3 Series resistances and the loop rule	215
Notes for chapter 9	222
Problems	223
Lab 9: Voltage and current	226

Iterated integrals

10 Iterated integrals (optional)	
10.1 A warm-up: iterated sums	233
10.2 Iterated integrals	235
10.3 Varying limits of integration	236
10.4 How to set up applications	238
10.5 Electric field of a continuous charge distribution	241
10.6 Surface integrals	247

Curved surface, uniform field, 248.—Flat surface, varying field, 248.—The general case, 250.

Problems	251
--------------------	-----

Sources of magnetism

11 Sources of magnetism

11.1 The current density	257
Definition, 257.—Continuity equation, 259.—★Transformation properties (optional), 262.	
11.2 Maxwell's equations	263
Adding a current term, 263.—The view from the top of the mountain, 264.	
11.3 Ohm's law in local form	266
11.4 The magnetic dipole	267
Modeling the dipole using a current loop, 268.—Dipole moment related to angular momentum, 269.—Field of a dipole, 270.	
11.5 Magnetic fields found by summing dipoles	271
11.6 Magnetic fields for some practical examples	272
11.7 ★The Biot-Savart law (optional) . .	274
Notes for chapter 11	277
Problems	279
Lab 11: Charge-to-mass ratio of the electron	284
Exercise 11A: Currents and magnetic fields	286
Exercise 11B: The magnetic field of twin-lead cable	288

AC circuits

12 Review of oscillations, resonance, and complex numbers

12.1 Review of complex numbers	291
12.2 Euler's formula	294
12.3 Simple harmonic motion	296
12.4 Damped oscillations	299
12.5 Resonance	300
Problems	305

13 AC circuits

13.1 Capacitance and inductance	309
Capacitors, 309.—Inductors, 310.	
13.2 Oscillations	313
13.3 Voltage and current	315

13.4 Decay	321
The RC circuit, 321.—The RL circuit, 322.	
Notes for chapter 13	324
Problems	325
Minilab 13: Energy in electric and magnetic fields	328

14 Impedance

14.1 Impedance	331
14.2 Power	333
A resistor, 333.—RMS quantities, 335.—A capacitor, 336.—An inductor, 336.	
14.3 Impedance matching	336
14.4 Impedances in series and parallel .	338
14.5 Capacitors and inductors in series and parallel	340
Problems	341
Exercise 14: Impedance: see one, do one	344

Stokes's theorem

15 Stokes's theorem	347
15.1 A round-up of vector calculus	347
15.2 Stokes's theorem	349
15.3 Ampère's law	350
15.4 Faraday's law	351
Problems	357

Electromagnetic properties of materials

16 Electromagnetic properties of materials	361
16.1 Conductors	362
16.2 Dielectrics	362
16.3 Magnetic materials	364
Magnetic permeability, 364.—Ferromagnetism, 368.	
16.4 Electromagnetic waves in matter . .	371
Problems	374

Relativity

17 ★Relativity (optional stand-alone chapter)

17.1 Time is not absolute	377
The correspondence principle, 377.—Causality, 378.—Time distortion arising from motion and gravity, 378.	
17.2 Distortion of space and time	380
The Lorentz transformation, 380.—The γ factor, 385.—The universal speed c , 391.	
17.3 No action at a distance	396
The Newtonian picture, 396.—Time delays	

in forces exerted at a distance, 396.—More evidence that fields of force are real: they carry energy., 397.

17.4 The light cone	399
17.5 ★The spacetime interval	400
17.6 Four-vectors and the inner product	404
17.7 Dynamics	406
Momentum, 406.—Equivalence of mass and energy, 410.—★The energy-momentum four-vector, 414.—★Proofs, 417.	
Problems	420
Photo Credits	437

Preface for the student

Electricity and magnetism is more fun than mechanics, but it's also more mathematical. Most schools have their curriculum set up so that the typical engineering student takes electricity and magnetism, or "E&M," concurrently with a course in vector calculus (a.k.a. multivariable calculus). E&M books normally introduce the same math for the benefit of students who will not take vector calculus until later, or who will learn a relevant topic in their math course too late in the semester. That's what I've done in this book, but I've also delayed some of the mathematical heavy lifting until the very end of the book, so that you will be more likely to benefit from having seen the relevant material already in your math course.

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.

There's a saying among biologists that nothing in biology makes sense without evolution. Well, nothing in E&M makes sense without relativity. Although your school curriculum probably places relativity in a later semester, I've scattered a small amount of critical material about relativity throughout this book. If you prefer to see a systematic, stand-alone presentation of this material, I've provided one in ch. 17, p. 377.

Preface for the instructor

I've attempted an innovation in the order of topics for freshman E&M, the goal being to follow the logical sequence while also providing plenty of opportunities for relating abstract ideas to hands-on experience. The typical sequence starts by slogging through Coulomb's law, the electric field, and Gauss's law, none of which are well suited to practical exploration in the laboratory. In this book, each of the first 5 chapters is short and includes a laboratory exercise that can be completed in about an hour and a half. The approach I've taken is to introduce the electric and magnetic field on an equal footing (which is in fact the way the subject was developed historically). As empirically motivated postulates, we take some primitive ideas about relativity along with the expressions for the energy and momentum density of the fields.

Another goal is to introduce the laws of physics in their natural, local form, i.e., Maxwell's equations in differential rather than integral form, without getting bogged down in an extensive development of the toolbox of vector calculus that would be more appropriate in an honors text like Purcell. Much of the necessary apparatus of div, grad, and curl is developed first in visual or qualitative form. At the end of the book we circle back and do some problem solving using the integral forms of Maxwell's equations.

Electric and magnetic fields



A telescopic view of an energetic eruption on the sun, August 2012. The bright loops at the upper left are ionized gas following the sun's magnetic field lines.

Chapter 1

Electric and magnetic fields

A college campus. Many, perhaps most, of the people walking by are doing something with their phones. They are sending and receiving invisible radio signals, which are disturbances in the electric and magnetic *fields*, sort of like disturbances in “the force” in Star Wars. Surrounding and penetrating the landscape of people, trees, and buildings, there is a second, hidden landscape of these fields. This hidden aspect of reality has no mass and is not made of material particles such as electrons.

1.1 A surprising link to the nature of time

Bear with me for a page while I describe some seemingly unrelated ideas about the nature of time, which will turn out to tie back logically to our study of fields. The phone signals in the college campus scene are structured in complex ways and carry information using codes like English and binary, but often the signal can carry all the needed information simply because of the *time* when it arrives. My wife worries about my safety when I head out for a day of rock climbing, so when her phone rings late in the afternoon, the simple fact of the arrival of a signal from me carries the information that I’m safe. This suggests that we consider the nature of *time* as a basic question when embarking on our study of electricity and magnetism.

A colorful experiment demonstrating the nature of time was done



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

by Hafele and Keating in 1971 (figure a). 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.*¹ 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 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, and 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: when one observer is in motion relative to another observer, they will disagree on the rate at which time passes. A clock appears to tick at its normal rate according to an observer who is at rest relative to the clock, while observers moving relative to the clock say that the clock is slow.



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

We are now ready to connect these observations about time back to our main subjects of study, which are the electric and magnetic fields. We will prove that if time is relative, then disturbances in the electric and magnetic fields must propagate (travel or spread out) at some finite speed, not instantaneously. The logic works like this. Suppose that observers Alice and Betty are both aboard spaceships, and moving at velocities that are different from each other, but both constant. 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. Alice says Betty’s time is slowed down due to the effect of motion on time, but Betty says Alice is slowed down. This seems paradoxical, since it seems that they should be able to get in contact by radio and resolve the disagreement. But our experience talking on cell phones misleads us into assuming that radio communication is instantaneous. It isn’t. I haven’t demonstrated in mathematical detail exactly how logical consistency is restored by the fact that signals take time to

¹Science 329 (2010) 1630

propagate. (That would take us too far afield into a discussion of relativity, which is not our main topic right now.) But this example is sufficient to show that logical consistency cannot be preserved if there is some mechanism for sending signals instantaneously, as in the “subspace radio” of the Star Trek universe.

Thus we can never send signals instantaneously. The style of reasoning here is called proof by contradiction. In this example, we have propositions 1 (that time is relative) and 2 (that instantaneous communication is possible), and if both 1 and 2 hold, then we reach a paradox, or logical contradiction. We can therefore conclude that either 1 is false or 2 is false. But 1 has been proved experimentally, so 2 must be false.

In your previous study of mechanics, you were probably briefly exposed to the concept of a field through mentions of the gravitational field. From that experience, it would be easy to get the impression that fields are an optional concept, and that changes in the gravitational field (e.g., due to the motion of a planet in its orbit) would take effect immediately, throughout the universe. This is what Isaac Newton believed, but it is not how the universe actually turns out to work.

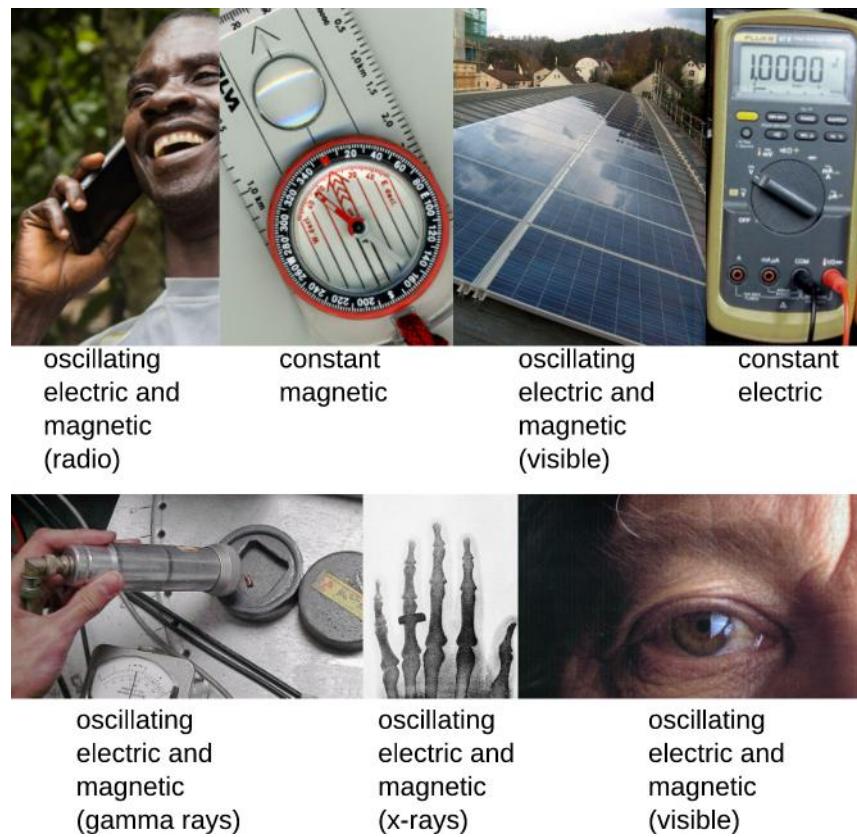
Since by now your physical intuition has been strongly influenced by Newton’s world-view, it’s worth noting that what I have given above as a tricky disproof of Newton is also backed up by empirical evidence. Nobody has ever succeeded in finding a method of instantaneously transmitting signals, or any type of cause and effect, with infinite speed.

Not only that but, experiments show that there is a fixed speed limit. This maximum speed at which signals can propagate is denoted c , and it 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 maximum speed of cause and effect, or as a sort of conversion factor between time and space units.

As an example of the kind of experimental work that has tested these claims, 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.

c / Tools for detecting, measuring, or exploiting electric and magnetic fields.



1.2 Basic properties of fields

In order to get started with our study of the electric and magnetic fields, we present three basic assumptions:

1. The electric and magnetic fields are *observable*. We can measure them.
2. The electric and magnetic fields are *vectors*.
3. *Superposition*: When fields are created by two effects, the fields contributed by the two effects at a particular point add

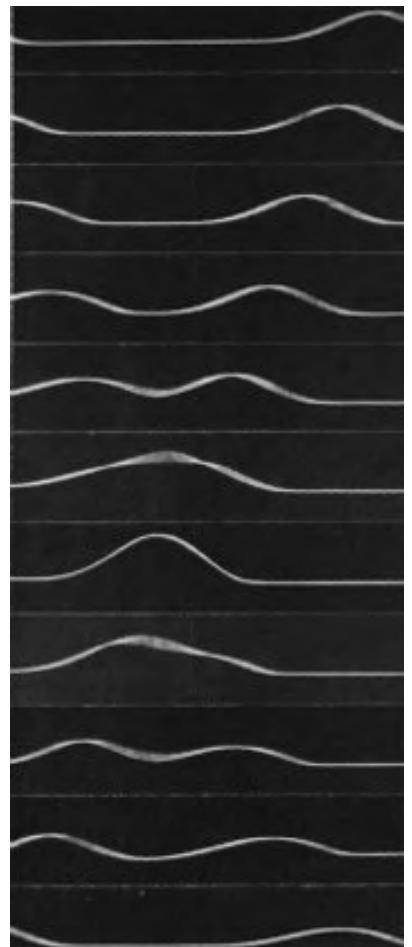
as vectors, and it is the vector sum that is actually observed at that location in space.

Figure c shows the wide variety of tools that are used for dealing with electric and magnetic fields. All of these can be used to detect the presence of fields, and some can actually measure the fields quantitatively. Many of them (the photovoltaic panels, Geiger counter, 19th-century x-ray film, and human eye) work by taking in the energy of the fields. Some of them are specialized for constant fields and others for oscillating ones. Among those that work with oscillating fields, there is further specialization by frequency. For example, the Geiger counter can detect gamma rays, which are a form of very high frequency light, but not radio waves.

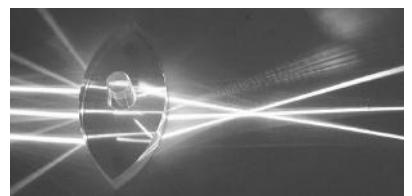
Assumption 3, superposition, is typical of wave phenomena, as shown in figure d. As an example involving electric and magnetic fields, the lens in figure e brings the three beams of light together at a point, and the waves pass through one another. The electric and magnetic fields in the region of intersection add as vectors but do not otherwise interact. The beams emerge unscathed on the other side.

To see the significance of assumptions 1 and 2, the observability and vector nature of the fields, it may be helpful to see examples of how they do not hold for other phenomena besides electricity and magnetism. These are a little exotic and are just for fun. If this sort of thing doesn't interest you for its own sake, feel free to skip ahead to section 1.3.

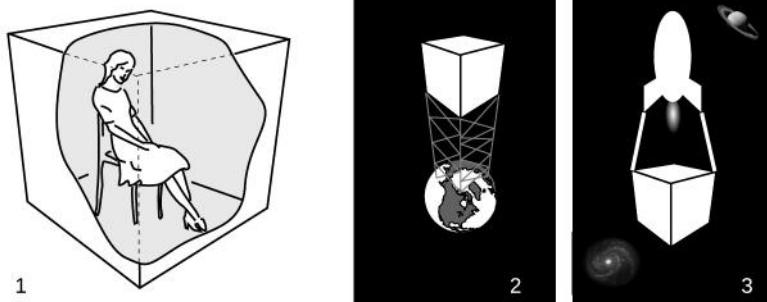
Our universe does contain fields that are scalars rather than vectors. One example is dark energy, discovered in 1998, which drives the accelerating expansion of the universe. Another example is the Higgs field, whose existence was predicted theoretically in 1964 and which was detected experimentally at the Large Hadron Collider in 2012.



d / Wave pulses on a coil spring travel toward each other, superpose, and reemerge without interacting.



e / Three rays of light converge.



f / The gravitational field is not observable.

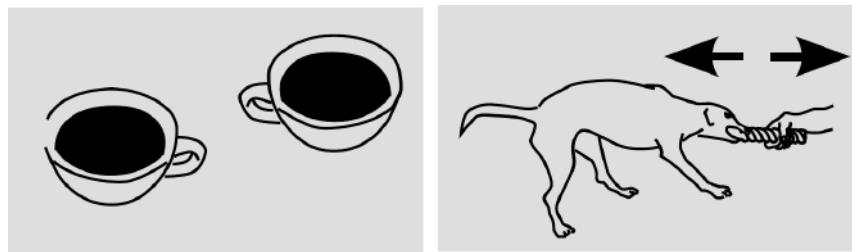
One of the considerations that led Einstein to his general theory of relativity is that the gravitational field is *not* necessarily observable. Consider Cathy, the girl in figure f. She is shown in a cutaway view, sealed inside a well-lit and air-conditioned box, which happens to be her favorite place to go and think about physics. (It's small, but it's private and it's all hers.) She feels pressure from the seat of the chair, and this would normally suggest that she was experiencing a gravitational force, perhaps from a nearby planet, $f/2$. But Cathy can't infer this without peeking outside at her surroundings. It's equally possible that the box is in deep space, $f/3$, with no gravity whatsoever, but is being towed by a rocket ship, so that it accelerates constantly. No possible experiment done inside the box can tell her which of these is the case.

1.3 Review of vectors

1.3.1 Definition of vectors and scalars

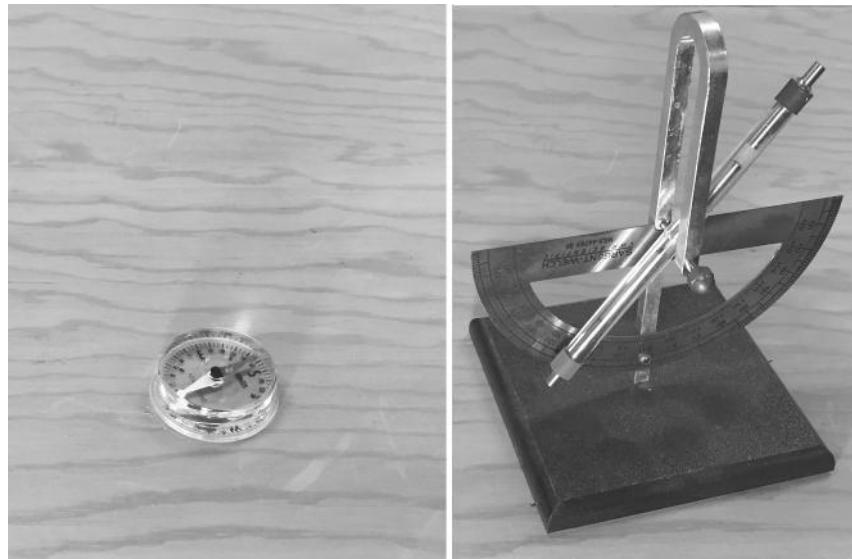
Because the electric and magnetic fields are vectors, we present a brief review of the topic, starting with a definition that may be different from the one you've seen previously, but is in better agreement with how physicists actually think about these things.

Most of the things we want to measure in physics fall into two categories, called vectors and scalars. A *scalar* is something that doesn't change when you turn it around, while a *vector* does change when you rotate it, and the way in which it changes is the same as the way in which a pointer such as a pencil or an arrow would change. Figure g shows two examples.



A vector \mathbf{A} has a magnitude $|\mathbf{A}|$, which means its size, length, or amount. Rotating a vector can change the vector, but will never change its magnitude.

g / Temperature is a scalar: a hot cup of coffee doesn't change its temperature when we turn it around. Force is a vector. When I play tug-of-war with my dog, her force and mine are the same in strength, but they're in opposite directions. If we swap positions, our forces reverse their directions, just as a pair of arrows would.



h / A magnetic compass and dip meter are used to find the direction of the earth's magnetic field in three dimensions.

In figure h, a compass is first used to find the vertical plane containing the earth's magnetic field. The needle defines the magnetic north-south direction (differing from true north, which is defined by the earth's rotation). Once this plane has been determined, a

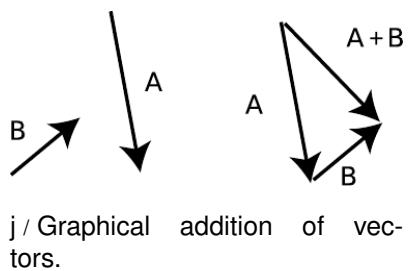
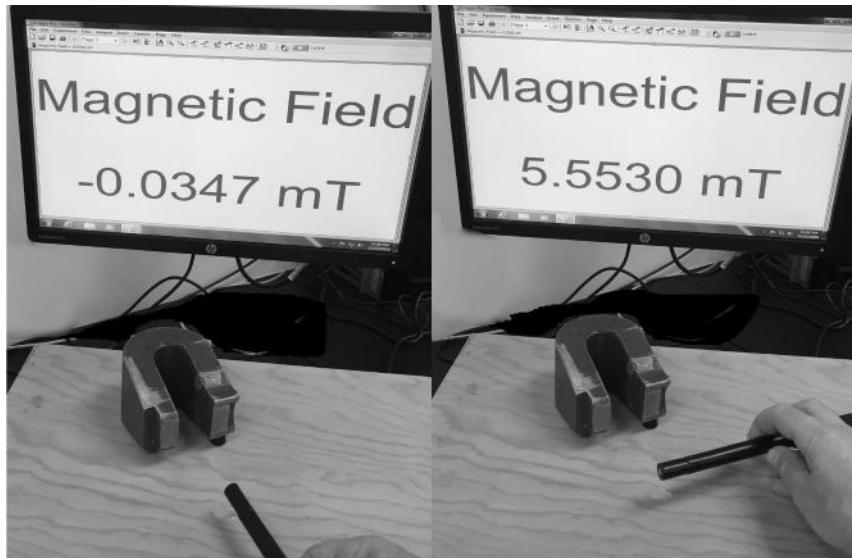
magnetic dip meter is used to find the three-dimensional direction of the field. The field has a large vertical component.

1.3.2 Components of vectors

A component of a vector is a signed real number giving its projection onto a line such as a coordinate axis. In two dimensions, when a vector \mathbf{A} lies in the x - y plane, at an angle θ counterclockwise from the x axis, its components are $A_x = A \cos \theta$ and $A_y = A \sin \theta$, where A is a shorthand notation for $|\mathbf{A}|$.

In figure i, the magnetic field sensor only measures the component of the field parallel to the wand, at its tip. At the point being tested, the component along the horseshoe's axis of symmetry is nearly zero. Rotating the sensor by 90 degrees gives a large reading, showing that the field is perpendicular to the axis of symmetry. If the magnetic field had been a scalar, then these two measurements would have had to give the same result.

i / A magnetic field sensor measures two perpendicular components of the field of a horseshoe magnet.



j / Graphical addition of vectors.

1.3.3 Vector addition

Scalars are just numbers, and we do arithmetic on them in the usual way. Vectors are different. Vectors can be added by placing them tip to tail, and then drawing a vector from the tail of the first vector to the tip of the second vector. A vector can be multiplied by a scalar to give a new vector. For instance, if \mathbf{A} is a vector, then $2\mathbf{A}$ is a vector that has the same direction but twice the magnitude. Multiplying by -1 is the same as flipping the vector, $-\mathbf{A} = (-1)\mathbf{A}$. Vector subtraction can be accomplished by flipping and adding.

The tip-to-tail method of adding vectors is called graphical addition. It is equivalent to adding components, which is called analytic addition. The following example demonstrates analytic addition in a case where it is necessary to do conversions back and forth between

the magnitude-direction and Cartesian descriptions of the vectors. If you have any trouble following this example, then you should review vector addition from a source that provides more detail than this brief review.

Analytic addition of vectors

example 1

▷ The displacement vector from San Diego to Los Angeles has magnitude 190 km and direction 129° counterclockwise from east. The one from LA to Las Vegas is 370 km at 38° counterclockwise from east. Find the distance and direction from San Diego to Las Vegas.

▷ The trig needed in order to find the components of the first leg (San Diego to LA) is:

$$\begin{aligned}\Delta x_1 &= (190 \text{ km}) \cos 129^\circ = -120 \text{ km} \\ \Delta y_1 &= (190 \text{ km}) \sin 129^\circ = 148 \text{ km}\end{aligned}$$

Applying the same pattern to the second leg (LA to Vegas), we have:

$$\begin{aligned}\Delta x_2 &= (370 \text{ km}) \cos 38^\circ = 292 \text{ km} \\ \Delta y_2 &= (370 \text{ km}) \sin 38^\circ = 228 \text{ km}\end{aligned}$$

For the vector directly from San Diego to Las Vegas, we have

$$\begin{aligned}\Delta x &= \Delta x_1 + \Delta x_2 = 172 \text{ km} \\ \Delta y &= \Delta y_1 + \Delta y_2 = 376 \text{ km}.\end{aligned}$$

The distance from San Diego to Las Vegas is found using the Pythagorean theorem,

$$\sqrt{(172 \text{ km})^2 + (376 \text{ km})^2} = 410 \text{ km}$$

(rounded to two sig figs). The direction is one of the two possible values of the inverse tangent

$$\tan^{-1}(\Delta y / \Delta x) = \{65^\circ, 245^\circ\}.$$

Consulting a sketch shows that the first of these values is the correct one.

1.3.4 Unit vector notation

Suppose we want to tell someone that a certain vector \mathbf{A} in two dimensions has components $A_x = 3$ and $A_y = 7$. A more compact way of notating this is $\mathbf{A} = 3\hat{\mathbf{x}} + 7\hat{\mathbf{y}}$, where $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, read “x-hat” and “y-hat,” are the vectors with magnitude one that point in the positive x and y directions. Some authors notate the unit vectors as $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ rather than $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$.



Example 1.

1.3.5 Rotational invariance

Certain vector operations are useful and others are not. Consider the operation of multiplying two vectors component by component to produce a third vector:

$$\begin{aligned}R_x &= P_x Q_x \\R_y &= P_y Q_y \\R_z &= P_z Q_z.\end{aligned}$$

This operation will never be useful in physics because it can give different results depending on our choice of coordinates. That is, if we change our coordinate system by rotating the axes, then the resulting vector **R** will of course have different components, but these will not (except in exceptional cases) be the components of the same vector expressed in the new coordinates. We say that this operation is not rotationally invariant.

The universe doesn't come equipped with coordinates, so if any vector operation is to be useful in physics, it must be rotationally invariant. Vector addition, for example, is rotationally invariant, since we can define it using tip-to-tail graphical addition, and this definition doesn't even refer to any coordinate system. This rotational invariance would still have held, but might not have been so obvious, if we had defined addition in terms of addition of components.

Calibrating an electronic compass

example 2

Some smart phones and GPS units contain electronic compasses that can sense the direction of the earth's magnetic field vector, notated **B**. Because all vectors work according to the same rules, you don't need to know anything yet about magnetism in order to understand this example. Unlike a traditional compass that uses a magnetized needle on a bearing, an electronic compass has no moving parts. It contains two sensors oriented perpendicular to one another, and each sensor is only sensitive to the component of the earth's field that lies along its own axis. Because a choice of coordinates is arbitrary, we can take one of these sensors as defining the *x* axis and the other the *y*. Given the two components B_x and B_y , the device's computer chip can compute the angle of magnetic north relative to its sensors, $\tan^{-1}(B_y/B_x)$.

All compasses are vulnerable to errors because of nearby magnetic materials, and in particular it may happen that some part of the compass's own housing becomes magnetized. In an electronic compass, rotational invariance provides a convenient way of calibrating away such effects by having the user rotate the device in a horizontal circle.

Suppose that when the compass is oriented in a certain way, it measures $B_x = 1.00$ and $B_y = 0.00$ (in certain units). We then expect that when it is rotated 90 degrees clockwise, the sensors will detect $B_x = 0.00$ and $B_y = 1.00$.

But imagine instead that we get $B_x = 0.20$ and $B_y = 0.80$. This would violate rotational invariance, since rotating the coordinate system is supposed to give a different description of the *same* vector. The magnitude appears to have changed from 1.00 to $\sqrt{0.20^2 + 0.80^2} = 0.82$, and a vector can't change its magnitude just because you rotate it. The compass's computer chip figures out that some effect, possibly a slight magnetization of its housing, must be adding an erroneous 0.2 units to all the B_x readings, because subtracting this amount from all the B_x values gives vectors that have the same magnitude, satisfying rotational invariance.

1.3.6 Dot and cross product

The vector dot product $\mathbf{A} \cdot \mathbf{B}$ is defined as the (signed) component of \mathbf{A} parallel to \mathbf{B} . It is a scalar. If we know the magnitudes of the vectors and the angle θ_{AB} between them, we can compute the dot product as $|\mathbf{A}||\mathbf{B}| \cos \theta_{AB}$. If we know the components of the vectors in a particular coordinate system, we can express the dot product as $A_x B_x + A_y B_y + A_z B_z$. The dot product of a vector with itself is the square of its magnitude, $|\mathbf{A}|^2 = \mathbf{A} \cdot \mathbf{A}$, and when we write A^2 as a notational shortcut, this is what we mean.

There is also a way of multiplying two vectors to obtain a vector result. This is called the vector cross product, $\mathbf{C} = \mathbf{A} \times \mathbf{B}$. The magnitude of the cross product is the area of the parallelogram illustrated in figure k. The direction of the cross product is perpendicular to the plane in which \mathbf{A} and \mathbf{B} lie. There are two such directions, and of these two, we choose the one defined by the right-hand rule illustrated in figure l.

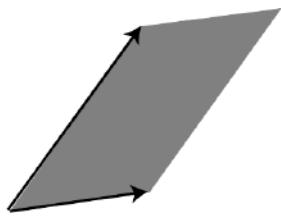
The dot and cross products are the *only* useful ways of multiplying two vectors to form a scalar or a vector, respectively. Any other definition of multiplication either will not yield a scalar or vector, will not be rotationally invariant, or will be a trivial variation on the dot or cross products in which the definitions are multiplied by some scalar, e.g., $7\mathbf{A} \cdot \mathbf{B}$ or $-\mathbf{A} \times \mathbf{B}$.

Unlike ordinary multiplication of real numbers, the cross product is anticommutative, $\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}$. The magnitude of the cross product can be expressed as $|\mathbf{A}||\mathbf{B}| \sin \theta_{AB}$. In terms of the components, we have

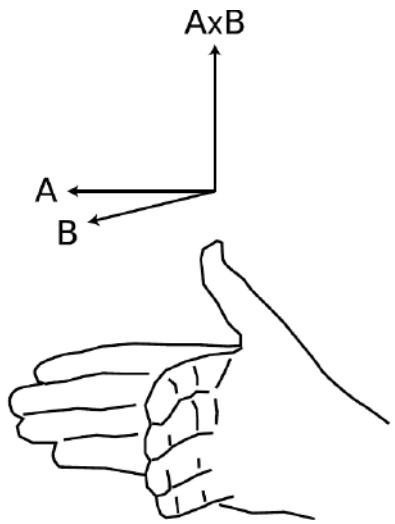
$$\begin{aligned}\mathbf{C}_x &= A_y B_z - B_y A_z \\ \mathbf{C}_y &= A_z B_x - B_z A_x \\ \mathbf{C}_z &= A_x B_y - B_x A_y.\end{aligned}$$

Discussion questions

- A** Suppose one magnetic field has a magnitude of 1 unit, another one 2 units. (We won't worry about what these units are called or how they fit into the SI until sec. 1.5.2, p. 28.) If you can superimpose these two fields in any orientation, what possible magnitudes can you get?



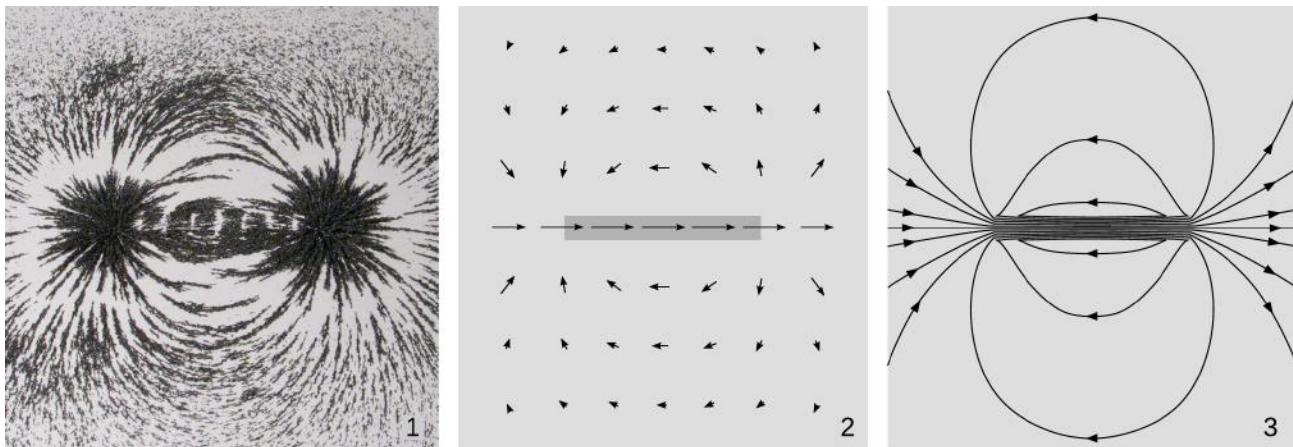
k / The magnitude of the cross product is the area of the shaded parallelogram.



l / The right-hand rule for the direction of the vector cross product.

B In certain kinds of wave disturbances in the electric and magnetic fields, it's possible for a field at a fixed point in space to rotate over time, exactly like the hand of a clock. For definiteness, let's say we're talking about the electric field, and that it's rotating in the x - y plane. (1) What can you say about the average values of the field's x and y components? (2) The energy density of the field is proportional to the square of its magnitude. What can you say about its average value?

C Electric field \mathbf{E}_1 has components $\langle 1, 1, 1 \rangle$ (in the appropriate units), and \mathbf{E}_2 has components $\langle 1, 0, 0 \rangle$. Let θ be the angle between them. Compare θ with 45 degrees and with 55 degrees.



m / Three representations of the magnetic field in and around a bar magnet.



n / Michael Faraday. When Queen Victoria asked him what the electrical devices in his lab were good for, he replied, "Madam, what good is a baby?"

1.4 Field lines

The photograph in figure m/1 provides a visually compelling image of the magnetic field of a bar magnet. The magnet itself is underneath a thin piece of white cardboard. What we see, on top, are iron filings, which become magnetized by the bar magnet, orient themselves along the field, and join up in chains. This type of visualization made a deep impression on the British physicist Michael Faraday (1791-1867), who was the mathematically untrained son of a poor blacksmith.

In general, we have two equally valid ways of conceptualizing an electric or magnetic field — not just visually, but mathematically. In the “sea of arrows” picture, m/2, the field at a particular point is represented by an arrow whose length and direction give the magnitude and direction of the field. In m/3, we instead have Faraday’s favored representation in terms of *field lines*.

The two pictures carry the same information. Given the sea of arrows representation, we essentially just link up the arrows to form the field lines ([Z32](#)). Given the field lines, we can also find the field

vectors: each field vector is tangent to the field line at the given point, and its magnitude is proportional to the *density* of the field lines. We'll have more to say in ch. 2 about why this works (it's not obvious), but for now we'll just note that it does. For example, the field lines are very dense on the interior of the bar magnet in figure m/3, so the field is very strong there. Near the top of m/3, the field lines are far apart, and indeed if we check back in m/2, the field vectors are very weak there.

Ch. 2 says more about how to define the density of field lines numerically. Because space has three dimensions and paper only two, we will often “fake it” by drawing cross-sections like figure m. This only works if the lines stay within the plane of the page, and even then we should not expect to get more than a rough conceptual idea of what actually happens in three dimensions, partly because density means something different in two dimensions than in three (figure o).

Field lines should not normally cross. If they did cross, then the direction of the field vector would have to be undefined. This could happen either because the field was undefined at that point or because it was zero. But in the case where the field is zero, the density of field lines is zero, and therefore we expect that no lines would actually penetrate to that location.

If a field has the same magnitude and direction everywhere in some region of space, then we say that the field is uniform. Figure p show the two representations of a uniform field. The earth's gravitational field is approximately uniform on small scales, but it does change its direction appreciably if we move far enough around the curve of the earth's surface, and its strength also falls off with elevation.

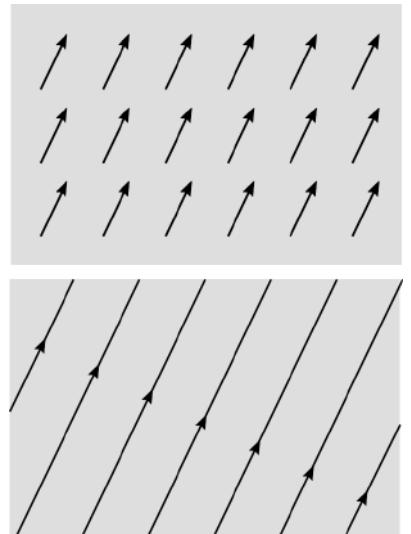
1.5 Energy in fields and measurement of fields

1.5.1 Energy in fields

We have not yet said anything about how to define and measure fields quantitatively, or what units would be used. One way to approach this is through the fact that fields carry energy. Energy is a scalar, but the fields are vectors. There is only one reasonable way to form a scalar out of a vector, and that is the dot product. Therefore a pure electric field \mathbf{E} , unaccompanied by any magnetic field, must have an energy density that depends only on $\mathbf{E} \cdot \mathbf{E}$, i.e., the field's squared magnitude E^2 . Similarly for a magnetic field \mathbf{B} , we expect an energy density that depends on $\mathbf{B} \cdot \mathbf{B}$. When the two fields are both present in a certain location, we could in principle have an additional term in the energy like $\mathbf{E} \cdot \mathbf{B}$, but such a term leads to predictions of phenomena that are not actually observed, e.g., an electrical device like a battery would tend to align itself with



o / In three dimensions, the density of bamboo stalks would be measured as stalks per unit area, while in two dimensions we would have stalks per unit length.



p / A uniform field, in the sea-of-arrows and field-line representations.



q / Charles Coulomb (1736–1806) was a French army officer, engineer, and physicist. The Coulomb constant k is named after him.

the earth's magnetic field, as a compass does.

Therefore we expect to have only an electric-field energy density that depends on E^2 and a corresponding magnetic one like B^2 . Supposing the relationships to be proportionalities (see discussion question A), it only remains to determine the constants of proportionality. These two constants depend on the units we define for the fields. Historically, there was a confusing variety of different and incompatible systems of units used for electricity and magnetism, and if you look through old books and scientific papers you can find a bewildering array of funky electrical units such as the abamp and the statcoulomb. Although some particle physicists and astrophysicists continue to use some of the less common systems (mostly the one referred to as cgs), today most scientists and engineers have settled on the version of the metric system called the SI. One way of notating the energy densities in SI units is the following.

$$dU_E = \frac{1}{8\pi k} E^2 dv \quad \text{and}$$
$$dU_B = \frac{c^2}{8\pi k} B^2 dv.$$

Here U stands for energy (to avoid a notational clash with \mathbf{E} for the electric field), k is a constant called the Coulomb constant, 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.

1.5.2 Units of the fields

If we had the luxury of being the first people to define the units for the fields, we would be free to choose $k = 1$, or even $1/8\pi k = 1$, which would simplify these expressions; the electric field would then have units of $J^{1/2}/m^{3/2}$. However, in the SI we have units for \mathbf{E} and \mathbf{B} that result in a value for k that has both nontrivial units and a nontrivial numerical value. Although this is a disadvantage when calculating energies, it does simplify certain other equations that we’ll encounter later. The SI units of the magnetic field are teslas (T), named after the Serbian-American inventor and prototypical mad scientist Nikola Tesla.

We can tell from the equations for the energy densities that E has the same units as cB , so that one way of expressing the units of the electric field would be T·m/s. In reality nobody does this, nor is there any convenient name or symbol for the relevant unit. It is instead expressed in terms of other electrical units, either the volt or

the coulomb, which we'll encounter later. The coulomb is a unit of electrical charge, a property of material objects that measures how strongly they participate in electrical interactions. The SI units of the electric field can be written as newtons per coulomb (N/C) or as volts per meter (V/m), and these expressions are equivalent, $1 N/C = 1 V/m$.

The Coulomb constant is approximately $k = 8.988 \times 10^9 \text{ Nm}^2/\text{C}^2$ in SI units. The SI is set up in such a way that k 's numerical value is equal to $10^{-7}c^2$, and since c has a defined value these days in the SI, k also has a defined value (which takes 17 decimal digits to express exactly). Because of the same pleasant coincidence that makes c so close to $3 \times 10^8 \text{ m/s}$, k can be approximated as 9×10^9 while incurring an error of only 0.1%, which is good enough for almost all applications.

When analyzing the units of an expression, we usually break the defined units down into the more basic ones, e.g., in a mechanical calculation we might replace newtons, N , with $\text{kg} \cdot \text{m/s}^2$. In the SI, the coulomb is (perhaps somewhat arbitrarily) considered to be more basic than the units of the fields. Therefore we would typically reduce electric field units to N/C , and if you're the type of person who enjoys crossword puzzles, you can convince yourself from the foregoing discussion that for the magnetic field, $1 \text{ T} = 1 \text{ N}\cdot\text{s}/\text{C}\cdot\text{m}$.

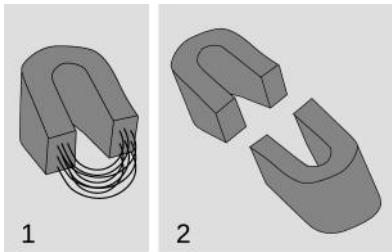
1.5.3 Defining the magnitude and direction of a field

If we are to be logically rigorous, the question arises of how to *define* the electric and magnetic fields. Rather than conceptual definitions in the style of a dictionary, physicists typically use operational definitions, meaning definitions that specify the operations that need to be carried out in order to measure the quantity. For example, an operational definition of time is that time is what a clock measures. The equations for the energy densities of the fields can be taken as operational definitions of the fields, except that they only seem to define the magnitudes of the fields, not their directions. Actually, the following example shows that they define more than they seem to define.

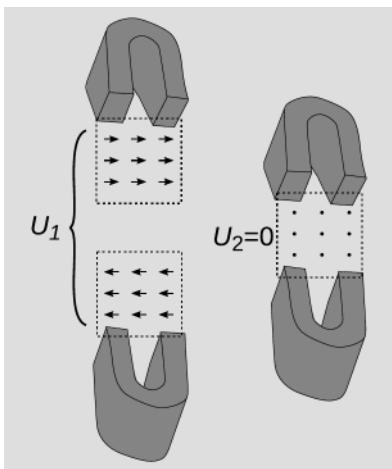
Direction of the field of a horseshoe magnet *example 3*

Figure r/1 shows the horseshoe magnet of figure i on p. 22, where we saw that in its plane of symmetry, the field was entirely perpendicular to the plane. In this figure we show the densely packed field lines that resulted in this observation, in the region near the pole tips where the field is very intense. There are no arrowheads drawn in the figure to show the direction of the field, because our definition of the field in terms of its energy density does not immediately seem to define any such thing or any way of determining it.

Now suppose that we have another such magnet, but there is



r / Example 3.



s / The case where the fields cancel, producing $U_2 = 0$.

still no indication of which pole is which. We bring the magnets close together, $r/2$, so that their fields superpose. There are two possibilities. Either the two magnets' contributions to the field are aligned, so that in the plane of symmetry the total field is $2B$, or they are reversed relative to one another, giving zero field in the midplane.

Although the field lines curve around, and it is only in the midplane that the two fields are collinear, we can tell that there will be a large difference between the energy stored in the field in these two cases. For simplicity, let's pretend that there is only some volume v near the pole tips where the field is large enough to matter, and let's approximate the field within this volume as being uniform and perpendicular to the midplane. When the magnets are far away from one another, each has energy $(c^2/8\pi k)B^2v$ stored in this region, so the total energy is double that,

$$U_1 = 2(c^2/8\pi k)B^2v. \quad [\text{total energy, apart}]$$

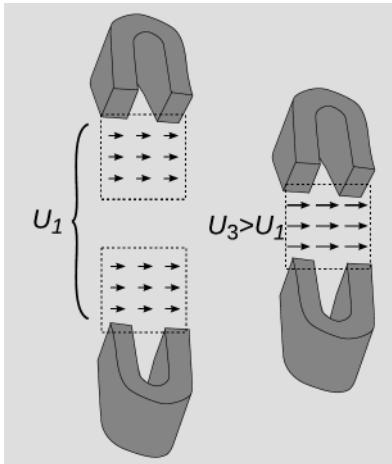
Now suppose that the magnets are brought near to each other, so that the volumes v coincide. In the case where the fields are reversed in relation to each other, the fields cancel (subject to our crude approximations), and we have

$$U_2 = 0. \quad [\text{total energy, reversed}]$$

With the fields cooperating, we replace B with $2B$, the result being

$$U_3 = 4(c^2/8\pi k)B^2v. \quad [\text{total energy, aligned}]$$

These three energies are all different. If the orientation of each magnet is reversed compared to the other, then they can reduce their magnetic energy from U_1 to U_2 by leaping toward each other. They attract, and the decrease in magnetic energy is transformed into kinetic energy. If the orientation is aligned, then energy is released if they fly apart, reducing their total energy from U_3 to U_1 . In this configuration, the magnets repel.



t / One of the magnets is turned around. The fields reinforce, producing the large energy U_3 .

Estimating the force from the field

example 4

Let's use the analysis of example 3 to find a rough numerical estimate of the force of attraction or repulsion. From figure i on p. 22, we take $B \sim 6$ mT, and we let the volume v be a cube, $v = \ell^3$, with $\ell = 4$ cm. The absolute value of the work done when the magnets are brought together or apart is $\Delta U = U_1 - U_2 = U_3 - U_1 = 2(c^2/8\pi k)B^2v$. (The fact that $U_1 - U_2 = U_3 - U_1$ tells us that the strength of the force will be of the same magnitude in the repulsive and attractive cases.) To the precision allowed by our crude approximations, we can approximate $F = dW/dx \approx$

$\Delta U/\Delta x = \Delta U/\ell$. Plugging in numbers, we get

$$\begin{aligned} F &\approx 2 \frac{c^2}{8\pi k} B^2 \ell^2 \\ &= 2 \frac{(3 \times 10^8 \text{ m/s})^2}{8\pi(9 \times 10^9 \text{ N}\cdot\text{m}^2/\text{C}^2)} (6 \times 10^{-3} \text{ T})^2 (4 \times 10^{-2} \text{ m})^2 \\ &= 0.05 \frac{\text{m}^2/\text{s}^2}{\text{N}\cdot\text{m}^2\cdot\text{C}^{-2}} \text{T}^2 \cdot \text{m}^2. \end{aligned}$$

To unroll the hairball of units, we use the fact that $1 \text{ T} = 1 \text{ N}\cdot\text{s}/\text{C}\cdot\text{m}$, and verify that the result turns out to be a force in units of newtons (problem 16, p. 35), which is as expected because when we plug in SI units, we should get a result in SI.

A force of 0.05 N is considerably too low for such massive magnets, and this is probably because the magnetic field in figure i was measured rather far away from the magnet in order to keep from overloading the sensor.

Generalizing the idea of example 3 (details [Z32](#)), we can define any field's direction relative to the direction of any other field. What we do not yet have is an absolute definition of the direction of a field, and in particular there is an ambiguity because we could always flip the direction of all our field vectors, simply as a matter of definition. This final ambiguity will be resolved in section 2.1.

Discussion question

A Strictly speaking, the argument on p. 27 only shows that the energy in a field *depends on* its squared magnitude, not that it is *proportional to* its squared magnitude. Recall that when we say y is proportional to x , $y \propto x$, we mean not just that when x increases, y increases as well, but also that y is *equal* to x multiplied by some constant. What goes wrong with the following proposed expressions for the energy density of the electric field?

$$\begin{aligned} &(\text{constant}) \sin E^2 \\ &(\text{constant})(E^2 + E^4) \end{aligned}$$

Notes for chapter 1

26 Linking up field vectors

A more careful definition of what it means to “link up” the field vectors to form field lines.

To define this more rigorously, imagine starting at a certain point in an electric field $\mathbf{E}(x, y, z)$. We now move some small distance ϵ in the direction of the field, to a new point. At this new point, we calculate the new and slightly different electric field, move ϵ in that direction, and continue indefinitely. It is intuitively plausible that if the field is smooth and well behaved, then in the limit as $\epsilon \rightarrow 0$, the path connecting these points converges to a uniquely defined curve such that the electric field is everywhere tangent to the curve.

The thing that we have constructed is called the integral curve of a vector field, and its existence and uniqueness, for smooth fields, follows from standard theorems about first-order differential equations. Physically, we may often consider fields that are not so smooth and well-behaved. For example, the field of a point charge (ch. 2) blows up to infinity at the point where the charge is, and the field lines can't be extended through that point.

31 Defining the direction of a field

If we know the energy density in terms of the field, we automatically also get a definition of one field's direction relative to another's.

We can abstract out the insight provided by example 3, p. 29, in the following way. If we have a way of measuring the magnitude of a field, say the electric field, then we automatically get a way of determining the orientation of any electric field vector \mathbf{E} in relation to some chosen reference field \mathbf{E}_o . This follows because we are free to superpose the two fields in any relative orientation we like, producing a total field $\mathbf{E} + \mathbf{E}_o$. The squared magnitude of this total field is $(\mathbf{E} + \mathbf{E}_o) \cdot (\mathbf{E} + \mathbf{E}_o)$, which we can measure. But this equals $\mathbf{E} \cdot \mathbf{E} + \mathbf{E}_o \cdot \mathbf{E}_o + 2|\mathbf{E}||\mathbf{E}_o| \cos \theta$, where θ is the angle between the vectors. (This is essentially the law of cosines, stated in vector language.) Since

all of the quantities in these expressions are assumed to be measurable, we can determine the unknown θ , which gives the orientation of \mathbf{E} relative to \mathbf{E}_o .

Problems

Key

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

1 In Fullerton, California, the earth's magnetic field is in the direction 59 degrees below horizontal. Find the ratio of the field's vertical component to its horizontal component. ✓

2 Suppose that, as in minilab 1 on p. 38, we have a known ambient magnetic field \mathbf{B}_a at right angles to a field \mathbf{B}_m whose magnitude is unknown. When the two fields are superposed, we measure the direction of the total field with a compass. Let θ be the (unsigned) angle by which the compass is deflected from the direction of the ambient field. (a) Find the ratio of the magnitudes B_m/B_a . ✓
(b) Interpreting your answer from part a, show that it makes sense in the special cases $\theta = 0, 45^\circ$, and 90° .

3 If you add a magnetic field with magnitude 1 T to a field of magnitude 2 T, what magnitudes are possible for the vector sum? (Consider the given values to be exact, ignoring significant figures.)

4 Two electric fields, both lying in the x - y plane, are superposed. Field \mathbf{E}_1 has a magnitude of 35.24 N/C and points in the direction 217.3° counterclockwise from the x axis. Field \mathbf{E}_2 has magnitude 48.01 N/C and direction 11.7° counterclockwise from the x axis. Find the magnitude and direction of the total field. ✓

5 Which of the following expressions make sense, and which are nonsense? For those that make sense, indicate whether the result is a vector or a scalar.

- (a) $(\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$
- (b) $(\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}$
- (c) $(\mathbf{A} \cdot \mathbf{B}) \times \mathbf{C}$

6 Give two reasons why it would not make sense for the energy density of the electric field to be proportional to the vector cross product $\mathbf{E} \times \mathbf{E}$ rather than the vector dot product $\mathbf{E} \cdot \mathbf{E}$.

▷ Solution, p. 426

7 Find the angle between the following two vectors:

$$\begin{aligned}\hat{\mathbf{x}} + 2\hat{\mathbf{y}} + 3\hat{\mathbf{z}} \\ 4\hat{\mathbf{x}} + 5\hat{\mathbf{y}} + 6\hat{\mathbf{z}}\end{aligned}$$

▷ Hint, p. 425 ✓

8 Find a vector that is perpendicular to both of the following two vectors:

$$\begin{aligned}\hat{\mathbf{x}} + 2\hat{\mathbf{y}} + 3\hat{\mathbf{z}} \\ 4\hat{\mathbf{x}} + 5\hat{\mathbf{y}} + 6\hat{\mathbf{z}}\end{aligned}$$

✓

9 The earth's magnetic field measured inside a certain classroom is $(10.0 \mu\text{T})\hat{\mathbf{x}} + (20.0 \mu\text{T})\hat{\mathbf{y}} + (20.0 \mu\text{T})\hat{\mathbf{z}}$. The volume of the classroom is 40.0 m^3 . (a) Find the magnitude of the field. ✓
(b) Find the energy of the magnetic field contained within the room. ✓

10 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. (We'll see in ch. 2 that this is a result of its greater electric charge.) Compare the energies stored in the two fields. ✓

11 The strongest magnetic field at the earth's surface is $66 \mu\text{T}$, off the coast of Antarctica, and the weakest is $23 \mu\text{T}$, in Paraguay. Find the ratio of the energy densities. ✓

12 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. ✓

13 In an electrical storm, an electric field builds up in the space between a cloud and the ground. Lightning occurs when the magnitude of the electric field reaches a critical value E_c , at which air is ionized.

- (a) Treat the cloud as a flat square with sides of length L . If it is at a height h above the ground, find the amount of energy released in the lightning strike. ✓
- (b) Based on your answer from part a, which is more dangerous, a lightning strike from a high-altitude cloud or a low-altitude one?
- (c) Make an order-of-magnitude estimate of the energy released by a typical lightning bolt, assuming reasonable values for its size and altitude. E_c is about 10^6 N/C.

See problem 14 for a note on how recent research affects this estimate.

14 In problem 13 on p. 35, you estimated the energy released in a bolt of lightning, based on the energy stored in the electric field immediately before the lightning occurs. The assumption was that the field would build up to a certain value, which is what is necessary to ionize air. However, real-life measurements always seemed to show electric fields strengths roughly 10 times smaller than those required in that model. For a long time, it wasn't clear whether the field measurements were wrong, or the model was wrong. Research carried out in 2003 seems to show that the model was wrong. It is now believed that the final triggering of the bolt of lightning comes from cosmic rays that enter the atmosphere and ionize some of the air. If the field is 10 times smaller than the value assumed in problem 13, what effect does this have on the final result of problem 13? ✓

15 Throughout this problem, we will use the standard notation ρ (Greek letter rho, which makes the "r" sound) for the energy density dU/dv . Suppose that electric field vector \mathbf{E}_1 exists at a certain point and creates an energy density ρ there. Field \mathbf{E}_2 has the same energy density ρ . When the two fields are superposed, the energy density is $\alpha\rho$, where α is a unitless number.

- (a) Find the angle between the two field vectors. ✓
- (b) Check that your answer to part a makes sense in the special cases $\alpha = 0, 2$, and 4 .

16 As claimed in example 4 on p. 30, show that units of

$$\frac{\text{m}^2/\text{s}^2}{\text{N}\cdot\text{m}^2\cdot\text{C}^{-2}} \text{T}^2\cdot\text{m}^2$$

are equivalent to newtons. You will need the fact that $1 \text{ T} = 1 \text{ N}\cdot\text{s}/\text{C}\cdot\text{m}$.

17 Consider the following statements:

1. If the dot product of two vectors is zero, then they are perpendicular.
2. If the cross product of two vectors is zero, then they are in the same direction.

Give one reason why statement 1 is false and two reasons why 2 is false.
▷ Solution, p. 426

18 Suppose that you are given vectors \mathbf{u} and \mathbf{v} , their components being specified in three dimensions. These vectors describe two sides of a triangle, and you want to find the triangle's area. This can be done using only the operation of the dot product, or using only the cross product. Outline both methods, and comment on which is more efficient.
▷ Solution, p. 426

19 Prove the anticommutative property of the vector cross product, $\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}$, using the expressions for the components of the cross product. Note that giving an example does not constitute a proof of a general rule.

20 Find three vectors with which you can demonstrate that the vector cross product need not be associative, i.e., that $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$ need not be the same as $(\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$.

21 Let a and b be any two numbers (not both zero), and let $\mathbf{u} = a\hat{\mathbf{x}} + b\hat{\mathbf{y}}$. Suppose we want to find a (nonzero) second vector \mathbf{v} in the x - y plane that is perpendicular to \mathbf{u} . Use the vector dot product to write down a condition for \mathbf{v} to satisfy, find a suitable \mathbf{v} , and check using the dot product that it is indeed a solution.

★

22 Suppose someone proposes a new operation in which a vector \mathbf{E} and a scalar B are added together to make a new vector \mathbf{C} like this:

$$\begin{aligned} C_x &= E_x + B \\ C_y &= E_y + B \\ C_z &= E_z + B \end{aligned}$$

For example, we could do this using the electric field vector for \mathbf{E} , and the magnitude of the magnetic field vector for B . Prove that this operation won't be useful in physics, because it's not rotationally invariant (sec. 1.3.5, p. 24). Note that this is a pure math problem requiring no knowledge of electricity and magnetism.
★

Minilab 0: Magnetic interactions and energy

This lab is designed to be done on the first day of class, without any prior reading or instruction. At the end of each part, the instructor will discuss the results with the class, and the different groups can compare notes. The general idea is to make observations using magnets. Based on your previous experience with magnets, any of the following might sound like possible mental pictures of how one magnet interacts with another:

- They interact through instantaneous action at a distance.
- The electromagnet makes a magnetic field, which affects the bar magnet.
- Both objects make fields. The fields overlap in space and interact.

By the end of this activity, we will be able to say something about which of these is right.

Apparatus

bar magnet
solenoid
power supply
string
compass

1. To start off with, we'll use two magnets: a bar magnet (permanent magnet) and an electromagnet, which is a hollow helical coil of wire. The electromagnet requires an external power supply. You will find that these two magnets interact with each other. Observe the interaction by holding the bar magnet delicately in your fingers, in various positions and orientations relative to the electromagnet. For sensitivity, you can also try hanging the bar magnet from a piece of string.

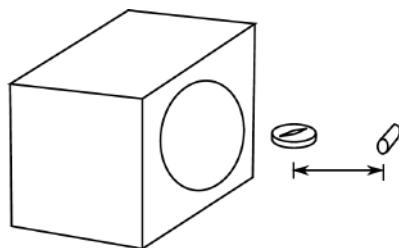
2. You should be able to feel both torques and forces. If testing hypotheses b and c, does either the torque or the force seem like a better indicator? Can you detect a force or torque due to the earth's magnetic field? For the interaction between the electromagnet and the bar magnet, with what location and orienta-

tion do you get the strongest interaction?

3. With the bar magnet at the position and in the orientation that give the strongest interaction, try turning off the power supply using the on-off switch. Try substituting the magnetic compass for the bar magnet. Try breaking the circuit rather than using the on-off switch.

4. Can you find a way to map out the magnetic effects in the space around the electromagnet? Around the bar magnet? What is a good way to represent these effects visually?

5. Put the apparatus in the configuration shown in the figure. Use books or other materials to get the compass and bar magnet up at the same height as the solenoid's axis. What happens as you vary the distance shown with the arrow?



a / Two different effects on the compass.

Class discussion: How would we set up a calculation to explain these observations quantitatively?

Carry out the calculation.

Notes for the instructor: The idea of part 3 is to observe the behavior described in more detail in example 5, p. 354. Although that example goes into Faraday's law, the result can be understood purely in terms of conservation of energy. This behavior may depend on the details of the construction of the power supply being used, or on how it is turned off (on-off switch versus a knob that controls the voltage).

Minilab 1: Magnetic field of a bar magnet as a function of distance

Apparatus

bar magnet
compass
2-meter stick

Goal: Find how the magnetic field of a magnet changes with distance along one of the magnet's lines of symmetry.

You can infer the strength of the bar magnet's field at a given point by putting the compass there and seeing how much it is deflected from the direction of the ambient field due to the earth and magnetic materials in the building.

The task can be simplified quite a bit if you pick one of the magnet's lines of symmetry and measure the field at points along that line. It should have an axis of symmetry that coincides with its center-line in the long direction, and another such axis for the short direction. The field at points on the first axis should be parallel to the axis, while the field on the second axis should be perpendicular to that axis.

1. Line up your magnet so it is pointing perpendicular to the ambient field, (nominally east-west). Choose one of the two symmetry axes, and measure the deflection of the compass at two points along that axis. For your first point, find the distance r at which the deflection is 70 degrees; this angle is chosen because it's about as big as it can be without giving very poor relative precision in the determination of the magnetic field. For your second data-point, use twice that distance. Using the result of homework problem 2, p. 33, by what factor does the field decrease when you double r ?

You will probably find that the ambient field in the room is strongly influenced by the magnetic field of the building and possibly the furniture. For example, in the lab room where I usually do this exercise, the lab benches contain iron or steel parts that distort the mag-

netic field, as my students can easily observe putting a compass on the top of the bench and sliding it around to different places. To work around this problem, we lay a 2-meter stick across the space between two lab benches, and carry out the experiment along the line formed by the stick.

It is also common to find that the magnetic field due to the building materials in the building is significant, and that this field varies from place to place. Therefore you should move the magnet while keeping the compass in one place. Then the field from the building becomes a fixed part of the background experienced by the compass, just like the earth's field.

Note that the measurements are very sensitive to the relative position and orientation of the bar magnet and compass.

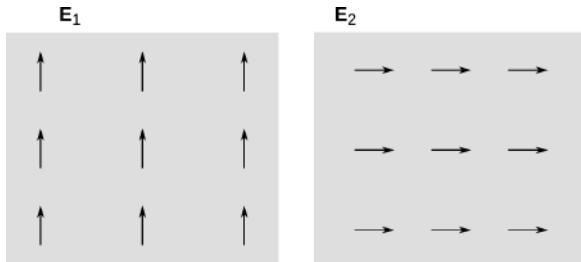
2. Based on your two data-points, form a hypothesis about the variation of the magnet's field with distance according to a power law $B \propto r^p$.
3. Take additional data at a range of distances, including the smallest and largest distances that it is practical to do. Graph the data on a log-log plot (i.e., with the log of B on one axis and the log of r on the other), and test whether your hypothesis actually holds.

Prelab

- P1. If the bar magnet's field follows the power law $B \propto r^p$, for some constant p , predict how the log-log plot should look.

Exercise 1: Visualizing superposed fields

Consider these two uniform electric fields, which have equal magnitudes.



- (1) Sketch their superposition in the same kind of sea-of-arrows representation.
- (2) Find its magnitude compared to $|\mathbf{E}_1|$ and $|\mathbf{E}_2|$.
- (3) Draw the field-line representation of \mathbf{E}_1 , \mathbf{E}_2 , and $\mathbf{E}_1 + \mathbf{E}_2$. (Just try to get the spacing qualitatively right. We'll do this kind of thing in more quantitative detail in ch. 2.)

Chapter 2

Gauss's law

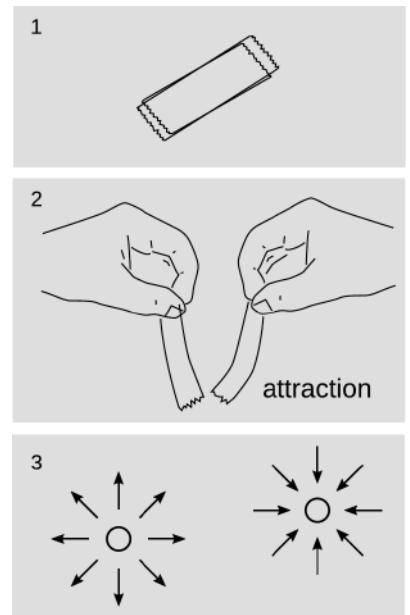
2.1 Gauss's law

2.1.1 Resolving the flip-the-arrowheads ambiguity

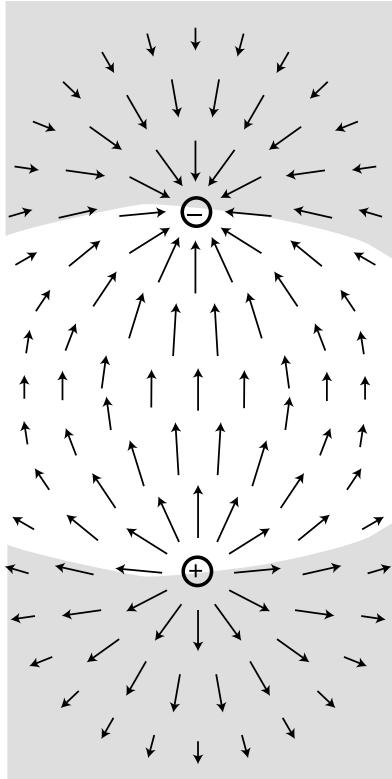
Continuing the train of thought from section 1.5 and 232, it would seem that by defining expressions for the energy density of the fields, we have effectively provided an operational definition for the fields, provided that we have some reference field somewhere that has a known direction. In fact, this reference field would be needed only in order to resolve the ambiguity that exists because we could always, e.g., define some field $\mathbf{F} = -\mathbf{E}$, which is the same as the electric field but with the opposite direction. This would be like flipping the arrowheads on all of our drawings. This “flip-the-arrowheads” ambiguity is entirely an arbitrary matter of definition, and for electric fields it was effectively fixed by Benjamin Franklin around 1750 (although we will give a description in a form that he would probably not have recognized).

It would have been a nuisance if Franklin had had to maintain some physical artifact in Philadelphia that had some electric field surrounding it, forcing other people to come there in order to consult it. What he did instead was to specify a procedure that could be followed, using materials commonly available at the time, in order to reproduce his standard. His prescription was to rub a piece of amber with wool. (Figures a/1 and a/2 show an easy way to do a similar procedure using more commonly available modern materials.) Once this has been done, we observe that the amber and the wool attract each other electrically, and we find that the attraction is about the same regardless of the orientations of the two objects. This suggests that each object must be surrounded by a sea of arrows that is approximately spherically symmetric, pointing either inward or outward, a/3. Franklin arbitrarily defined things so that the electric field surrounding the wool would point outward, while the field around the amber pointed inward.

By the way, almost any pair of substances will exhibit this kind of effect when rubbed or touched together, but amber is particularly good at producing a strong field. The Greek word for amber is “elektron,” which is the origin of English words like “electricity.”



a / 1. Place a piece of sticky tape on a tabletop, then stick another on top of it. 2. Lift them off the tabletop as a unit, then separate them. The result is that they attract one another. 3. Interpretation of this type of experiment in terms of the fields surrounding the two objects.



b / Example 1. The labels + and - are explained in the following subsection.

Explaining the attraction

example 1

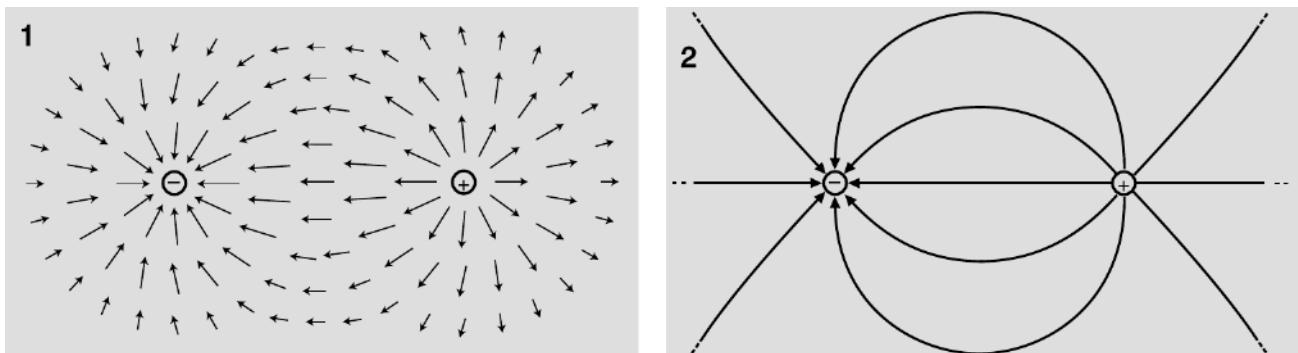
To explain the attraction between two objects in an observation like the one in figure a, we can employ an argument very similar to the one used in the case of the two horseshoe magnets, example 3, p. 29. We idealize each of the two objects as a point. When the objects are far apart, each has some electric field energy, and these add up to some nonzero amount U_1 . For our present purposes, we do not even need to know yet the function $E(r)$ that gives the magnitude of the electric field as a function of the distance away from the object. For simplicity, we will assume that the two objects have fields that are equal in strength but opposite in magnitude. (A qualitatively similar result is obtained even if we relax this requirement.) Now suppose the two objects are brought so close together that they are right on top of each other. The fields will cancel everywhere, and we will have the energy $U_2 = 0$. This loss of electric field energy means that we can do mechanical work by allowing the objects to come together, or if the objects were released in free space, they could convert this field energy into kinetic energy as they accelerated toward each other. Our conclusion is that the force is attractive.

We have not considered the intermediate case where the distance between the charges is nonzero but still small enough so that the fields overlap appreciably, as in figure b. In the region indicated approximately by the shading in the figure, the superposing fields of the two charges undergo partial cancellation because they are in opposing directions. The energy in the shaded region is reduced by this effect. In the unshaded region, the fields reinforce, so the energy there is increased.

It would be quite a project to do the integral in order to find the energy gained and lost in the two regions, but it is fairly easy to convince oneself that the energy is less when the charges are closer, as expected from interpolation between U_1 and U_2 . This is because bringing the charges together shrinks the high-energy unshaded region and enlarges the low-energy shaded region.

2.1.2 Sources and sinks

When the field lines flow out of an object, we call that object a source of the field, and when they flow in we call it a sink. The electric field has sources and sinks, but as far as we are able to tell, there are no magnetic sources or sinks in our universe. We can quantify the strength of an electric-field source by assigning an object a number called its electric charge, which is, roughly speaking, proportional to the number of field lines that begin or end on it. Of course the number of field lines is actually infinite, and it's just that we're picking a certain finite and representative sample of them to draw in our pictures, so a little more work is needed in



c / The sea-of-arrows and field-line representation of the field surrounding a source and a sink.

order to make this into a real definition. We will do so later in the chapter. The SI unit of charge is the coulomb (C). Positive charges are conventionally assigned to sources, negative to sinks. Material particles such as protons and electrons have charge. The electric and magnetic fields are themselves uncharged, and are not made of material particles.

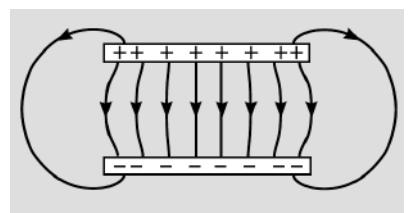
If we consider the possibility of pointlike charged particles, then in a figure like c, we have a bunch of electric field lines that all either begin or end at the same point. This is not, as it might seem, in contradiction with our proof on p. 27 that field lines never cross ([264](#)).

By the way, figures b and c are the simplest examples of a type of charge distribution called an electric *dipole*, which more generally is a kind of unbalanced arrangement of positive and negative charges. For example, some molecules are electric dipoles. Dipoles will be discussed in more detail in sec. 5.4, p. 132.

Another important way of setting up an arrangement of charges is the *capacitor*, consisting of two pieces of metal called the electrodes. One electrode is positively charged and the other negatively charged. Figure d is a side view of a capacitor consisting of two parallel, flat electrodes. The charges push and pull on each other and come to an equilibrium in which they are distributed in a certain way. The electric field is mostly confined to the interior of the capacitor, and is nearly uniform there.

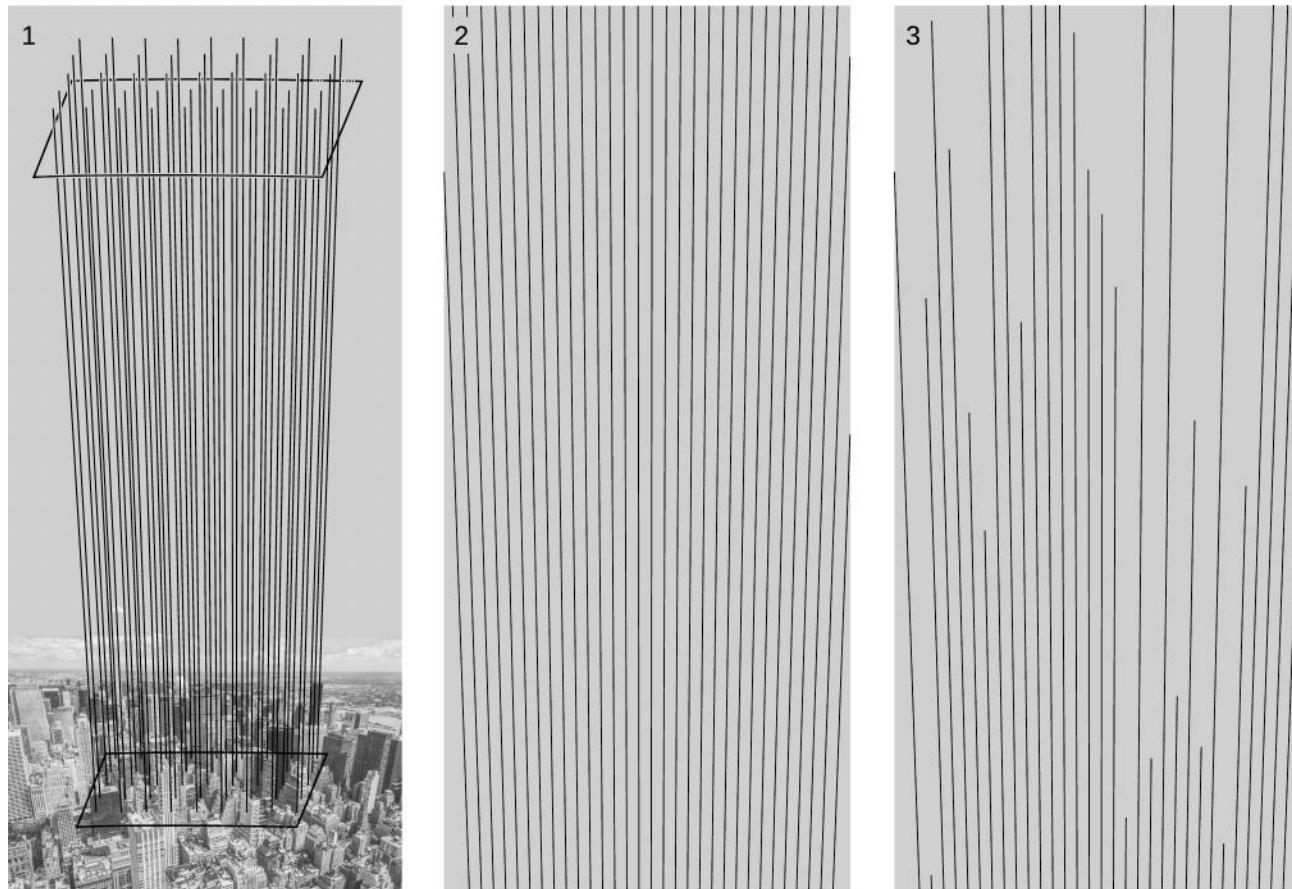
2.1.3 Gauss's law for field lines, in a vacuum

It is natural to want to know how the electric field of a particle such as a proton falls off with distance. This would be a law of physics that would play a role analogous to that of Newton's law of gravity in the case of the gravitational field. But in fact Newton's law of gravity is false, and any law of this type *must* ultimately be false (or at best be some kind of approximation), for the reasons described in section 1.1, p. 15. It is not possible for cause and effect to propagate faster than c , so we can't really have the kind



d / A parallel-plate capacitor.

of instantaneous action at a distance described by Newton's law of gravity, which has no t in it. The shortcomings of Newton's law of gravity turn out to be relatively inconsequential unless you're a physicist studying phenomena like black holes and gravitational waves, but the corresponding issues in electricity and magnetism are very real and practical. If electricity and magnetism worked the way Newton thought the universe worked, radio wouldn't exist.



e / A bundle of gravitational field lines rise up through New York City.

Since gravity is more familiar, let's see how we could find an alternative way of describing the strength of gravity that would not fall prey to these objections. In figure e/1, we see a bundle of gravitational field lines rising up through New York City. The earth is the source of these field lines, but we are looking at them outside the earth, which means that we are contemplating the behavior of gravitational fields in a vacuum. (The mass of the air is negligible.) Because the earth is round, the field lines spread farther apart as they rise. At the bottom of the picture, near the streets of Manhattan, these lines pass through a square with a certain area, while farther up, at the top of the drawing, the same number of lines pass through a square with a larger area.

Now the area of any geometrical shape is proportional to the square of its linear dimensions, so these areas are proportional to r^2 , where r is the distance from the center of the earth. (If this doesn't seem clear, it may help to imagine the situation for complete spherical surfaces.) The density of the field lines is the strength of the field, so we conclude that the earth's gravitational field falls off as $1/r^2$, as in Newton's law of gravity.

This derivation is so pat that it makes it seem a little mysterious how any field could *not* have $1/r^2$ behavior, and in fact there are such fields, including the fields associated with nuclear forces. To see how this would work, consider figures e/2 and e/3. In e/2 we have just simplified e/1 a little, making it easier to see what's going on by drawing only a single flat fan of field lines — but the other ones in front and behind are still there. We can have a field that falls off faster than $1/r^2$, but then we would need a picture like e/3, where some of the field lines simply die out, randomly, at some point in the air. In this example, we imagine that they are not terminating on material particles, but simply at random points in empty space. Figure e/3 is not inherently silly, and in fact it is a pretty good representation of a nuclear field, but it is not how electric and magnetic fields work. We state this as a law of physics.

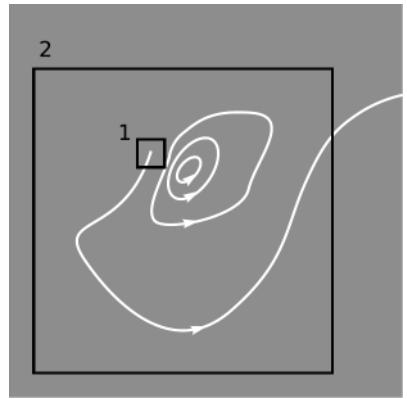
Gauss's law for field lines, in a vacuum

In a vacuum, electric and magnetic field lines never begin or end.

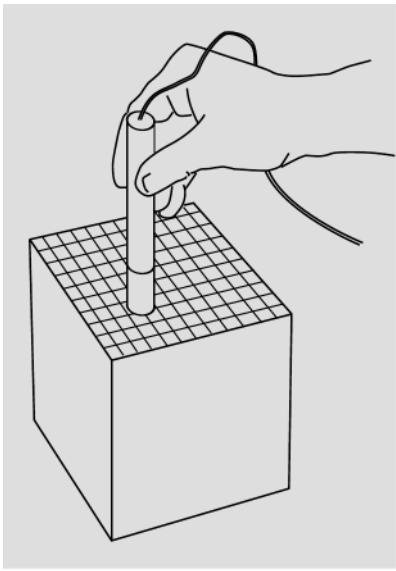
We have developed our statement and interpretation of Gauss's law for electricity and magnetism by exploiting the analogy with gravity. It seems as though Newton's law of gravity is logically equivalent to Gauss's law for gravity, so that it wouldn't matter much which one we used. But the two laws make different predictions in cases where masses are moving around, and they also have a very different character. Newton's law of gravity is *global*: it says that mass *here* has an effect right *now* on mass *there*, possibly very far away. Gauss's law, on the other hand, is *local*.

To see the distinction, consider the naughty field shown in figure f. Most of the field lines form closed loops, which is legal according to Gauss's law (and typical behavior for a magnetic field). But one of the field lines is disobeying Gauss's law: it starts at a point inside the tiny box marked 1. Even if we were restricted to a keyhole view through an extremely powerful microscope, we could still locate the violation if we carefully scanned the entire figure.

As a linguistic analogy for this distinction between local and global laws, consider this sentence: MY KATS EATS RATS. There are three errors, but if we're restricted to looking at the letters one at a time, without the larger context, the only one we can detect is the flipped letter "R." The rule against flipping letters is a local law.



f / A field that violates Gauss's law in a vacuum.



g / A practical experiment that tests Gauss's law for magnetism, globally.

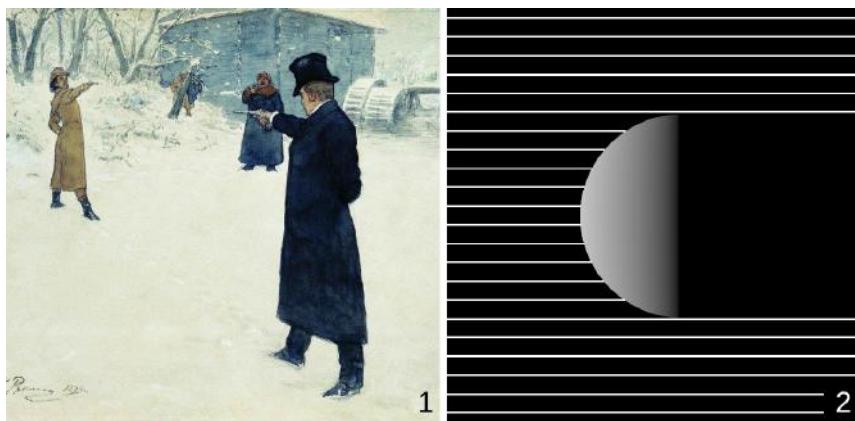
The form of Gauss's law stated above is very specialized. Later we will see how it can be generalized so it describes regions of space in which charged matter exists. We will also learn how to state it in terms of the field vectors rather than the field lines.

2.2 A global form of Gauss's law

There are other ways to detect the violation in f. We could, for example, check the larger box labeled 2, and note that although no field lines pass in through the edges of this box, one passes out. So a local law can lead to global predictions, and this can be helpful. A practical realization of this kind of global test is shown in figure g, and you will carry it out in Minilab 2, p. 71. A magnet is inside the cubical box. At one of the centimeter squares, we measure the component of the magnetic field perpendicular to the box. The intensity of the field is proportional to the number of field lines per cm^2 , so by measuring the field, we are essentially counting the number of field lines that exit the box through this square centimeter (or that enter it, if the field is inward). By adding up all of these measurements on a computer, for all six sides of the box, we get a count of the net number of field lines that exit the box, counting a line that enters as -1 . Gauss's law for magnetism predicts that the total is zero.

The sort of sum described above is called the *flux*, Φ (capital Greek letter phi). When it's not clear from context whether we're talking about the magnetic field's flux from the electric field's, we write Φ_E and Φ_B .

h / Two examples of flux from contexts other than electromagnetism.



Before stating formal definitions of electric and magnetic flux, let's build some intuition with two examples of the concept of a flux in other contexts. "Flux" is just the Latin word for "flow." In figure h/1, the men dueling with pistols stand sideways. This is to minimize the flux of bullets entering through the fronts of their chests. In h/2, parallel rays of sunlight hit a planet. The effect of the sun's

heating on a particular square kilometer of land is proportional to the number of rays intercepted, and this is different at different latitudes. At the poles, the land area is turned sideways to the sun's rays, as in the example of the duelers. (We should also note some ways in which these examples are *not* quite analogous to electromagnetic fluxes. Both the bullets and the sunbeams tend to be absorbed when they hit the object's surface; if they could pass through and reemerge from the other side, then the ingoing and outgoing fluxes would cancel — at least mathematically — the dueler would still probably be dead. And although these two examples involve things physically moving through space, electric and magnetic field lines do not represent stuff moving from place to place.)

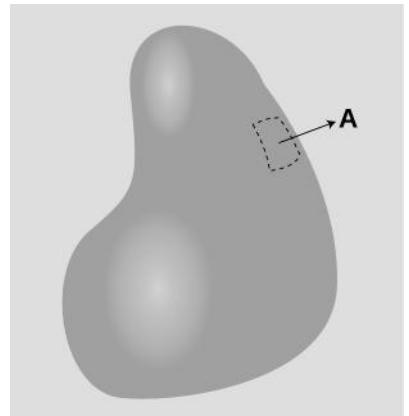
To more formally define the flux through a surface, we break up the surface into small areas A_1, A_2, \dots , such as the squares in figure g. For each of these areas, we define a unit normal vector $\hat{\mathbf{n}}_i$, which points outward. “Unit” means that $|\hat{\mathbf{n}}_i| = 1$, and “normal” means that it is perpendicular to the surface. What the wand measures in figure g is the component of the magnetic field perpendicular to the box, $\mathbf{B}_i \cdot \hat{\mathbf{n}}_i$. The magnetic flux Φ_B is then defined as the sum $A_1 \mathbf{B}_1 \cdot \hat{\mathbf{n}}_1 + A_2 \mathbf{B}_2 \cdot \hat{\mathbf{n}}_2 + \dots$. To make the notation less unwieldy, we can define area vectors $\mathbf{A}_i = A_i \hat{\mathbf{n}}_i$ and use sigma notation, $\Phi_B = \sum \mathbf{B}_i \cdot \mathbf{A}_i$. Using a large number of small areas gives an approximation to the flux, but the exact flux is given by the limit of such an expression as the number of area elements goes to infinity and each area becomes very small, $\Phi_B = \lim \sum \mathbf{B}_i \cdot \mathbf{A}_i$. This kind of continuous sum of infinitely many infinitesimal things is an integral, so we notate it as one,

$$\Phi_B = \int \mathbf{B} \cdot d\mathbf{A}. \quad [\text{definition of magnetic flux}]$$

A similar definition is used for the flux of the electric field. Because the dot product is a scalar, flux is a scalar. Like the kind of ordinary definite integral from freshman calculus, this notation defines a number. (If you're used to interpreting an integral sign without limits of integration as an indefinite integral, then you would be misled into thinking that this was an indefinite integral, which would be a function rather than a number. Also keep in mind that the notation $d\mathbf{A}$ does not mean that we're integrating with respect to some variable \mathbf{A} ; it just means an infinitesimal area.)

A closed surface is one that has no edges, like a box rather than a piece of paper. Gauss's law in a vacuum can be stated in terms of the field vectors, in global form, by saying that the magnetic flux through any closed surface is zero,

$$\Phi_B = 0, \quad \Phi_E = 0. \quad [\text{Gauss's law in a vacuum}]$$

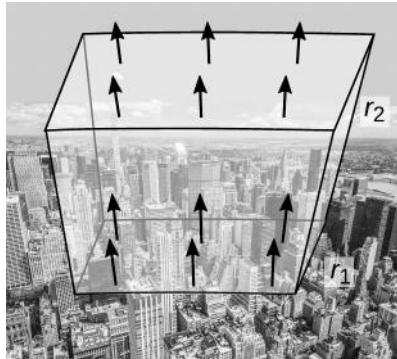


i / An area vector can be defined for a sufficiently small part of a curved surface.

Flux above New York

example 2

Often for problems that have a symmetry, we can apply Gauss's law without actually having to do an integral to calculate the flux. In figure j we have constructed a surface (a surface used for this purpose is referred to as a Gaussian surface) that takes advantage of the spherical symmetry of the earth's gravitational field \mathbf{g} . The top and bottom surfaces are sections of the spheres with radii r_1 and r_2 , while the sides are vertical. By this construction, the field is parallel to the surface at the sides and perpendicular to it at the top and bottom. The flux therefore has contributions only from the top and bottom, not from the sides.



j / A Gaussian surface in the earth's gravitational field.

If we consider a small portion of the area at the top, the area vector $d\mathbf{A}$ is parallel to the field, so for the integrand, $\mathbf{g} \cdot d\mathbf{A} = |\mathbf{g}| |d\mathbf{A}| \cos 0 = g dA$. When we evaluate the integral to find the flux over the top surface, we then have

$$\int_{\text{top}} g dA = g_2 \int_{\text{top}} dA,$$

since g on the top is a constant, which we call g_2 . But the remaining integral is simply the area of the top surface, A_2 , so we can find the integral without actually using any of the techniques of calculus. The calculation at the bottom surface is similar, except that the cosine factor is $\cos 180^\circ = -1$, so we pick up a minus sign, indicating that the flux passes in through the bottom. Adding up the positive flux through the top and the negative flux through the bottom, the result for the total flux is

$$\Phi_g = g_2 A_2 - g_1 A_1.$$

Setting this equal to zero gives

$$\frac{g_2}{g_1} = \frac{A_1}{A_2} = \left(\frac{r_2}{r_1}\right)^{-2},$$

i.e., $g \propto r^{-2}$, as expected.

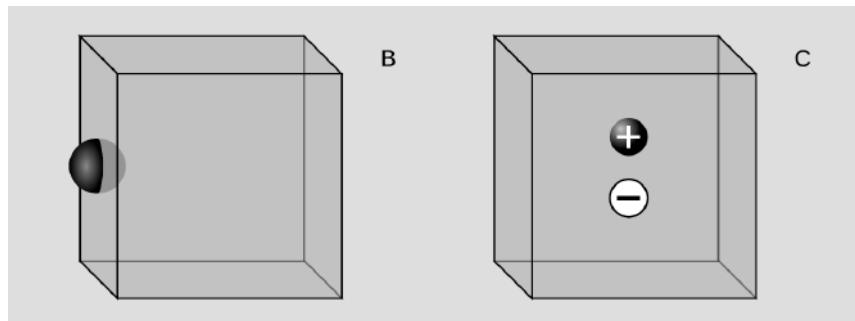
An integral of the form $\int \dots d\mathbf{A}$ is called a surface integral. In example 2, we were able to evaluate a surface integral without really knowing any fancy calculus, because the integrand was basically a constant, and we know that for any kind of integral, we can take the constant outside the integral. Techniques for evaluating more complicated area integral are discussed in ch. 10, p. 233.

Discussion questions

A Describe the duelers in figure h/1, p. 46, in terms of area vectors. Do the same for the snowplow in the figure, whose blade is angled. How does the rate of plowing (volume of snow per unit time) depend on the angle, if the vehicle moves at a fixed speed?



Discussion question A.



k / Discussion questions B and C.

B One question that might naturally occur to you about Gauss's law is what happens for charge that is exactly on the surface — should it be counted toward the enclosed charge, or not? If charges can be perfect, infinitesimal points, then this could be a physically meaningful question. Suppose we approach this question by way of a limit: start with charge q spread out over a sphere of finite size, and then make the size of the sphere approach zero. Figure k/B shows a uniformly charged sphere that's exactly half-way in and half-way out of the cubical Gaussian surface. What is the flux through the cube, compared to what it would be if the charge was entirely enclosed? (There are at least three ways to find this flux: by direct integration, by Gauss's law, or by the additivity of flux by region.)

C The dipole in figure k/C is completely enclosed in the cube. What does Gauss's law say about the flux through the cube? If you imagine the dipole's field pattern, can you verify that this makes sense?

2.3 Field of a point charge at rest

Our argument that the earth's field was proportional to $1/r^2$ depended on only two ingredients, Gauss's law and spherical symmetry. The spherical symmetry existed because the earth is (approximately) a sphere and because we were discussing the earth in the frame where it is at rest. If we consider a charged particle such as an electron as an idealized pointlike object, and discuss it in the frame of reference where it is at rest, then spherical symmetry again holds, and since Gauss's law is also valid for electric fields, we obtain exactly the same result. The electric field surrounding a point charge is proportional to $1/r^2$, where r is the distance from the charge. If the charge is q , then filling in the constants of proportionality gives

$$E = \frac{kq}{r^2} \quad [\text{point charge at rest}]$$

for the magnitude of the field. Here q is the charge in coulombs, and k is the same Coulomb constant that we originally introduced in the context of the energy density of a field. It's not hard to show that the same k should pop up in the way it does in this equation and in the equation for the energy density; proving the factors of π and such is a little extra work, and will be more easily taken care of later, when we have other techniques at our disposal. It may sometimes be useful to express not just the magnitude but also the direction of the electric field in this equation. We let $\hat{\mathbf{r}}$ be the unit vector in the direction from the charge to the point at which the electric field is being evaluated. Then the electric field at that point is

$$\mathbf{E} = \frac{kq}{r^2} \hat{\mathbf{r}}.$$

In the case where $q < 0$, the scalar factor is negative, and the electric field points inward.

There is a subtle point about the logic leading to these equations for the field of a point charge, which is that we obtained them without ever using any laws of physics that described the point charge itself. We used the only form of Gauss's law currently at our disposal, the vacuum form, which applies only to the empty space *around* the charge. By doing this, we found the proportionality $E \propto 1/r^2$, and it is in some sense arbitrary that we chose to refer to the proportionality constant q as "charge."

Field of an electric dipole, in its mid-plane

example 3

Consider the dipole in figure I. We wish to calculate the electric field in the dipole's mid-plane, at a point on the x axis with coordinates $(x, 0)$.

The principle of superposition says that the field at this point can be found by adding the fields due to the two charges, and adding means vector addition, because the fields are vectors. In analytic addition of vectors, we add components. The component E_z , perpendicular to the page, is clearly zero by symmetry — neither charge's field has any component out of the plane of the diagram. It would also be a waste of time to calculate the x components, since these too are guaranteed to cancel. This means that all we really need to do is calculate the y components and add them.

The distance from the top, positive charge to our point of interest is $r = \sqrt{x^2 + (\ell/2)^2}$. The magnitude of the field contributed by this charge is

$$E = \frac{kq}{r^2}$$

In terms of the angle θ that the field vector makes with the y axis, the y component is

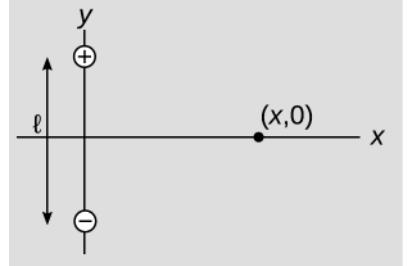
$$E_y = E \cos \theta = E \frac{\ell/2}{r} = \frac{kq\ell}{2r^3}.$$

The negative charge's field has an equal E_y , so the total field at our point is double this, or

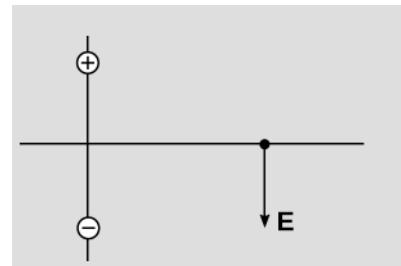
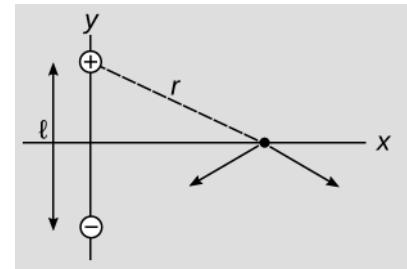
$$E_y = \frac{kq\ell}{r^3}.$$

If we wished to, we could substitute in our expression for r to get this in terms of x , but the expression is actually nicer in terms of r .

This result has the interesting property that for large distances, where $r \approx x$, it depends on the charge distribution only through the product $q\ell$, which is referred to as the dipole moment. The on-axis field of an electric dipole is calculated in problem 9, p. 69. Dipoles are discussed in more detail in secs. 5.4, p. 132, and 11.4, p. 267.



I / Example 3.



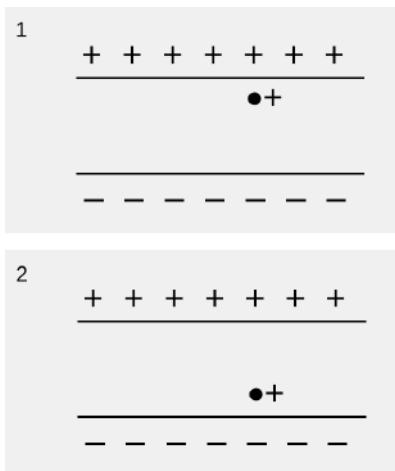
2.4 Electric force on a charge

In addition to describing the electric field made by a charge, it is natural to consider the action of some externally imposed field on a charge. We imagine that this ambient field is created by some other objects, and that the charge q is small enough so that its own reaction on these other objects is not enough to disturb them. That is, the background contribution \mathbf{E} to the field does not change just because we insert q . When q is small enough to make this a good approximation, we call it a test charge. In cases like example 4 on p. 30 and example 1 on p. 42, we have been able to explain the force between two objects in terms of the energy of their superposed fields. Using a similar style of reasoning ([264](#)), we find that the force of an electric field on a test charge is

$$\mathbf{F} = q\mathbf{E}.$$



m / A spark plug, example 4.



n / Discussion question A.

If we liked, we could take this as the definition of the electric field (and most books do). This also explains why, in the system of units we have constructed, the units of the electric field can be written as newtons per coulomb (N/C). If the charge is positive, the force is in the direction of the field, while a negative charge feels a force in the opposite direction.

A spark plug

example 4

▷ In a car with a gas-burning engine, the air-gas mixture is exploded by a spark from a spark plug. Suppose that the electric field required in order to create the spark is 2.0×10^7 N/C. Estimate the acceleration of an octane molecule having a charge of 3.2×10^{-19} C and a mass of 1.9×10^{-25} kg.

▷ The magnitude of the force is qE , and by Newton's second law the resulting acceleration is $a = (q/m)E = 3.4 \times 10^{13}$ m/s². This is an enormous acceleration! Note that the properties of the molecule influence its motion only through the ratio q/m of its charge to its mass. We'll learn more about the spark plug in problem 9, p. 106.

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?

2.5 Coulomb's law

Suppose that we have *two* point charges, q_1 and q_2 , both at rest and separated by a distance r . The electric field is the sum of the fields contributed by the two charges. If we want to find the force acting on one of the charges, say 2, then it doesn't make sense to try to take into account the contribution to the field from charge 2 itself (265), just the force that 1 makes on 2. Let $\hat{\mathbf{r}}_{21}$ be the unit vector in the direction from 1 to 2. The contribution to the field from charge 1, at the position of charge 2, is $\mathbf{E} = (kq_1/r^2)\hat{\mathbf{r}}_{21}$, and the resulting force on 2 is

$$\mathbf{F}_2 = \frac{kq_1q_2}{r^2}\hat{\mathbf{r}}_{21}.$$

The force is repulsive if the charges have the same sign, and attractive if they have opposite signs. For the force \mathbf{F}_1 acting on 1, we have the same expression but with $\hat{\mathbf{r}}_{21} = -\hat{\mathbf{r}}_{12}$. Newton's third law applies, which is not to be taken for granted and is not necessarily true if the charges are moving or have been moving at some time in the past.

This equation is known as Coulomb's law. It plays the same role in electrical interactions that Newton's laws of gravity plays in gravity, and it has the same form, but because charge, unlike mass, can have either sign, electrical forces can be either attractive or repulsive.

If one charge is kept fixed while another is moved toward or away from it, then the work done by the electric force is the definite integral of Coulomb's law, while the potential energy is minus the indefinite integral, or

$$U = \frac{kq_1q_2}{r},$$

where the constant of integration is arbitrarily chosen to be zero, so that $U \rightarrow 0$ as $r \rightarrow \infty$.

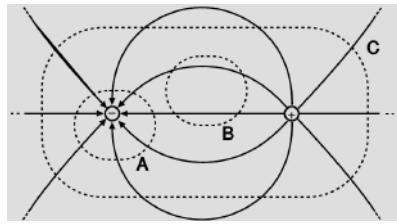
Discussion question

- A For two interacting point charges, the electrical potential energy is $U = kq_1q_2/r$. What is the corresponding expression for gravity? Explain how the signs work, and sketch graphs for the various cases.

2.6 Charge

2.6.1 Gauss's law, not in vacuum

The vacuum form of Gauss's law for electric fields can be generalized to handle the case where charges are present. In terms of field lines, Gauss's law states that field lines begin and end only on charges, and the number of field lines that begin or end on a charge is proportional to the charge.



- o / The number of field lines coming in and out of each region depends on the total charge it encloses.

Gauss's law with field lines and charges

example 5

Figure c/2 shows eight lines at each charge, so we know that $q_1/q_2 = (-8)/8 = -1$. Because lines never begin or end except on a charge, we can always find the total charge inside any given region by subtracting the number of lines that go in from the number that come out and multiplying by the appropriate constant of proportionality. Ignoring the constant, we can apply this technique to figure o to find $q_A = -8$, $q_B = 2 - 2 = 0$, and $q_C = 5 - 5 = 0$.

The global form of Gauss's law for electric fields says that the flux through a closed surface is

$$\Phi_E = 4\pi k q_{\text{in}},$$

where q_{in} is the total charge inside the surface.

A point charge

example 6

If we enclose a point charge q with a spherical Gaussian surface of radius r , Gauss's law gives

$$\begin{aligned} 4\pi k q &= \int \mathbf{E} \cdot d\mathbf{A} \\ &= \int \frac{kq}{r^2} dA. \end{aligned}$$

Taking the constant factor outside, we have

$$\begin{aligned} 4\pi k q &= \frac{kq}{r^2} \int dA \\ &= \frac{kq}{r^2} \cdot 4\pi r^2. \end{aligned}$$

The two sides of the equation are equal, so we have verified Gauss's law and proved that the proportionality constant of $4\pi k$ in Gauss's law is the correct one.

A uniform sphere of charge

example 7

Consider a sphere of radius b containing electric charge with a uniform density ρ , in units of C/m^3 , so that the total charge is $q = (\text{volume})\rho = (4\pi b^3 \rho/3)$.

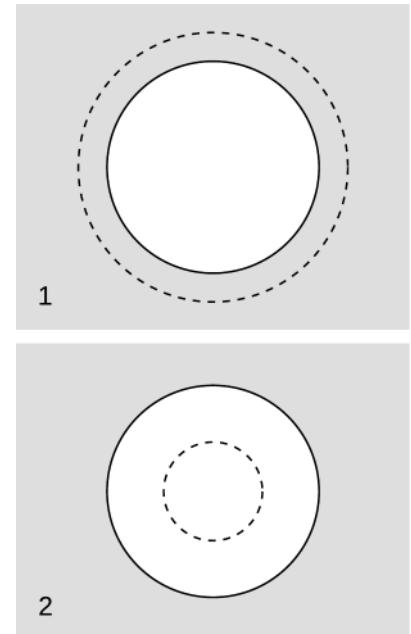
This can't be a metal sphere like a door knob, because charge can move easily through a metal, and the repulsion would then cause it all to spread to the surface rather than remaining uniformly spread throughout the interior. This is, however, a reasonable model of an atomic nucleus, which has a finite size and is usually spherical or approximately spherical (p. 80).

Let's find the electric field of this charge distribution.

By symmetry, the field points purely in the radial direction, and must also go to zero at the center. For any point outside the sphere, the calculation of example 6 carries through as before. We can turn around the logic of that example and treat E as the unknown, and the result will be as before, $E = kq/r^2 = (4\pi kb^3 \rho/3)/r^2$. So from the outside, we can't even tell that the sphere has a finite size, and the field has a $1/r^2$ form, exactly as for a point charge.

Inside the sphere, we will have a different result. We certainly can't have a $1/r^2$ field there, for that would give an infinite field at the center, and we've already found that the field at the center must be zero. The calculation of example 6 doesn't apply here, because if we make a spherical Gaussian surface with a radius $r < b$, then the charge q_{in} enclosed by the surface is smaller than q . For example, if we take $r = b/2$, then the volume enclosed is $1/8$ of the total (because volume is proportional to the cube of the linear dimensions), and $q_{in} = q/8$. In general, a sphere of radius $r < b$ will enclose an amount of charge $q_{in} = (\text{volume})\rho = (4\pi r^3 \rho/3)$. Since the field on our surface is constant in magnitude and always perpendicular to the surface, the calculation from example 6 requires only this modification of replacing q_{in} with this smaller value. Completing the algebra results in $E = (4\pi k\rho/3)r$.

Let's do some checks and interpretation. We expected by symmetry to have $E = 0$ at the center, and this does work out. The reason that the result is a simple proportionality to r (as opposed to a proportionality to some other power of r) is that there was a factor of r^3 for the enclosed charge, but a factor of r^{-2} because electricity is an inverse-square force. Connecting with your knowledge of Newtonian gravity, this result is also the one we would expect from the shell theorem: the shells at radii greater than r do not contribute to the field inside them. The analogy between electric and gravitational fields is exact in this example because they both go like $1/r^2$. We also see that b does not appear in our expression for the interior field. This is not an accidental can-



p / Example 7. 1. A spherical Gaussian surface with $r > b$. 2. A surface with $r < b$.

cellation. Gauss's law tells us that the flux through the surface at $r < b$ only depends on the charge inside that surface. It doesn't matter at all whether there is charge on the outside.

The technique demonstrated in this example can also be applied to other examples with a special symmetry, for example the field of a uniformly charged cylinder (problem 13, p. 69). The idea is to try to choose a Gaussian surface whose symmetry matches that of the problem, so that the field anywhere on the surface is of the same magnitude and perpendicular to the surface.

2.6.2 Invariance of charge

Ever since Galileo we have known that different observers could have frames of reference in motion relative to one another, and that the results of some measurements would depend on the frame of reference while others were not. Velocity is the classic example of a frame-dependent quantity. For example, your copy of this book is at rest in a frame of reference tied to your desk, but in the frame of reference of a Martian it's whizzing through space at high speed. A quantity that does not depend on the frame of reference is called invariant. Time is approximately invariant, but not exactly (sec. 1.1, p. 15).

Electric charge is, according to the best experimental evidence, perfectly invariant: observers agree on the charge of an object, regardless of their frame of reference. Another way of saying this is that when we start or stop an object's motion, its charge does not change at all. It is possible to do incredibly rigorous tests of this statement, because atoms are made of charged particles (protons and electrons), and the electrons in many atoms are orbiting at a significant fraction of the speed of light. If the charge of an electron depended on its state of motion, then dropping an electron into orbit in an atom would change its charge, but experiments¹ rule out such a change to the incredible precision of one part in 10^{21} .

2.6.3 Quantization of charge

In 1909, Robert Millikan and coworkers published experimental results showing that electric charge seemed to come only in integer multiples of a certain amount, notated e and referred 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, $e = 1.602 \times 10^{-19} \text{ C}$.

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 .

¹Marinelli and Morpugo, "The electric neutrality of matter: A summary," Physics Letters B137 (1984) 439.

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 most 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.

2.6.4 Conservation of charge

Electric charge is conserved in all known physical processes. When people say that their phone or laptop is “out of charge,” really it’s out of *energy* — even if we wanted to, it wouldn’t be possible to destroy or use up the electric charge in such a device. Usually in an electric circuit, we just send the same charge around and around, like the water being circulated through a fish tank or swimming pool by the filter pump.

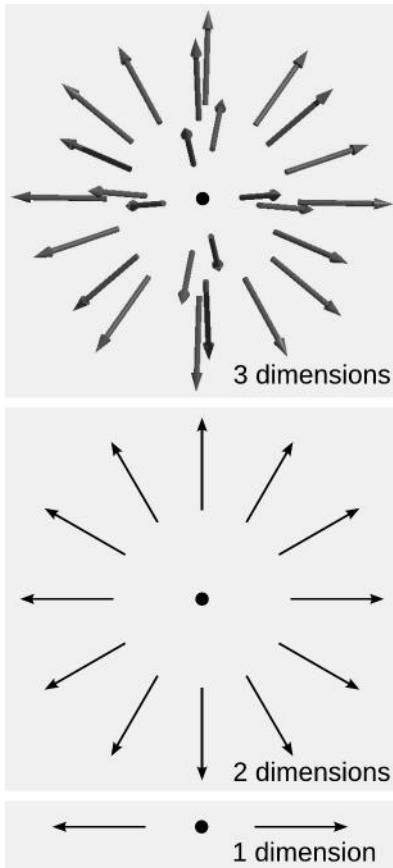
As a more exotic example, consider the neutron, a subatomic particle that is very common — neutrons make up about half the mass of your body. Neutrons are never found by themselves in nature, because in a matter of minutes, a free neutron will spontaneously undergo the radioactive decay process



You’ll be introduced to some of these particles more formally in chapter 3, but for now we just note the charges, in units of e ,

$$0 = 1 - 1 + 0,$$

which satisfy conservation of charge.



q / The field of a point charge in three, two, and one dimensions.



r / A coaxial cable.

2.7 Gauss's law when the number of dimensions is effectively less than three

Our world is three-dimensional, but it is full of physical systems that are effectively one- or two-dimensional. Pool and car racing are effectively two-dimensional sports. When a locomotive pulls a string of freight cars up a steep grade, the situation is effectively one-dimensional. Figure q sketches the field of a point charge in three, two, and one dimensions. These occur in practice. For example, the two-dimensional field pattern in figure q exists in the layer of transparent insulator in figure r, if we take a cross-section perpendicular to the cable's central axis.

In three dimensions, our previous analysis showed that the field of a point charge was proportional to $1/r^2$. This was because the field was proportional to the density of field lines, as measured by number of field lines per unit area, and area has units of distance squared.

In two dimensions, the density of field lines is measured by the number of field lines per unit length, so the field of a point charge is proportional to $1/r$. If charge is spread out along a line, like the central wire of the cable in figure r, then a cross-section perpendicular to the wire intersects the wire only at one point, and we can therefore treat the wire as a point charge.

The simplest analysis of all is in one dimension. Here the field lines can't spread at all — there is nowhere for them to escape. The “density” of field lines is simply the number of field lines, and this doesn't change with distance from the charge. The field varies as $1/r^0$, i.e., it is constant. This analysis applies to the field of a flat sheet of charge.

Electric field of a line of charge

example 8

The reasoning above about proportionalities was an extremely simple way to find the exponent with which the electric field varied as a function of distance in several cases. However, this technique will never determine unitless constants of proportionality such as 2 and π . In this example we apply Gauss's law in more detail to determine the correct numerical factors in the case of a line of charge.

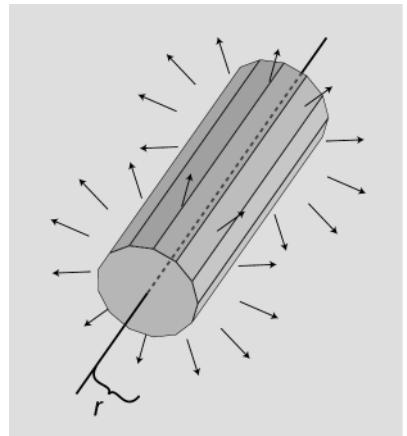
Consider the field of an infinitely long line of charge, holding a uniform charge per unit length λ . The problem has two types of symmetry. The line of charge, and therefore the resulting field pattern, look the same if we rotate them about the line. The second symmetry occurs because the line is infinite: if we slide the line along its own length, nothing changes. This sliding symmetry, known as a translation symmetry, tells us that the field must point directly away from the line at any given point.

Based on these symmetries, we choose the Gaussian surface shown in figure s. If we want to know the field at a distance R from the line, then we choose this surface to have a radius r , as shown in the figure. The length, L , of the surface is irrelevant.

The field is parallel to the surface on the end caps, and therefore perpendicular to the end caps' area vectors, so there is no contribution to the flux. On the long, thin strips that make up the rest of the surface, the field is perpendicular to the surface, and therefore parallel to the area vector of each strip, so that the dot product occurring in the definition of the flux is $\mathbf{E}_j \cdot \mathbf{A}_j = |\mathbf{E}_j||\mathbf{A}_j| \cos 0^\circ = |\mathbf{E}_j||\mathbf{A}_j|$. Gauss's law gives

$$4\pi kq_{in} = \sum \mathbf{E}_j \cdot \mathbf{A}_j$$

$$4\pi k\lambda L = \sum |\mathbf{E}_j||\mathbf{A}_j|.$$



s / Applying Gauss's law to an infinite line of charge.

The magnitude of the field is the same on every strip, so we can take it outside the sum.

$$4\pi k\lambda L = |\mathbf{E}| \sum |\mathbf{A}_j|$$

In the limit where the strips are infinitely narrow, the surface becomes a cylinder, with (area)=(circumference)(length)= $2\pi rL$.

$$4\pi k\lambda L = |\mathbf{E}| \times 2\pi rL$$

$$|\mathbf{E}| = \frac{2k\lambda}{r}$$

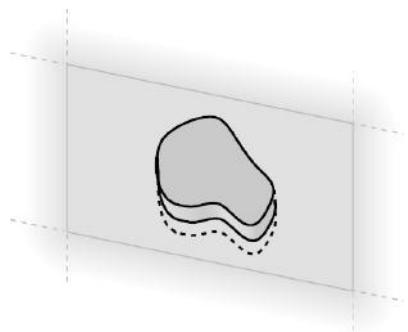
We had already found the $1/r$ nature of the field, and the factor of $k\lambda$ has to be there because of units, so the only new accomplishment in this calculation to determine that the unitless constant was 2.

Field of a charged plane

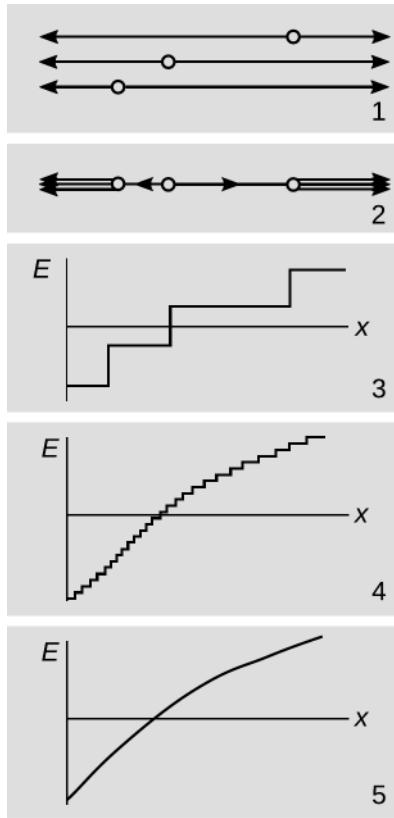
example 9

▷ Suppose we want to find the field of an infinite, uniformly sheet of charge, with charge density σ in units of coulombs per square meter. As in example 8, we know in advance, by visualizing the way the field lines spread out, that the result must be of a form with a certain exponent — the exponent is zero, so that the field is independent of the distance r . If we are to fix the constant of proportionality by the same technique, then we need to start by picking a Gaussian surface. What would be an advantageous choice?

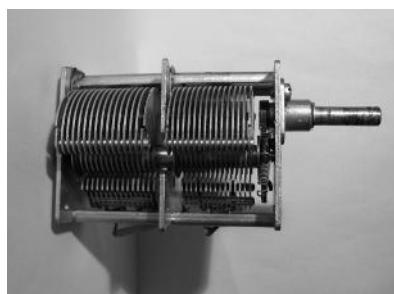
▷ Figure t shows a shape that works, called a Gaussian pillbox. It respects the planar symmetry of the problem, and is chosen so that the field on the two flat faces is always parallel to $d\mathbf{A}$. This calculation is completed in problem 10, p. 69.



t / Applying Gauss's law to an infinite sheet of charge.



u / 1. Three equal positive charges, in one dimensions, with their field lines. 2. The superposition of the fields. 3. A graph of the electric field, for the same example. 4. A similar graph with a larger number of charges. 5. The field of a continuous charge distribution.



v / This variable capacitor acts effectively like the kind of one-dimensional system described in figure u.

2.8 Gauss's law in local form, with field vectors

The laws of physics are ultimately local, but so far our only local way of stating Gauss's law locally is in terms of field lines: in a vacuum, they don't end. Reasoning about field lines is often less practical than working with field vectors, so we would like to have a local statement of Gauss's law that is expressed in terms of field vectors.

To figure out how to translate from the field-line picture to field vectors, it will be helpful to simplify to one dimension and then generalize back to three dimensions at the end. In one dimension, field lines can't spread out because there is nowhere for them to escape. If multiple field lines are superposed, we have to draw them slightly offset so that we can see them separately. The "density" of field lines is simply the number of field lines that pass through a given point.

In the example shown in figure u/1, three equal positive charges are shown with their field lines. The constants of proportionality are chosen such that each charge has exactly two field lines coming out of it, and a field line "density" of 1 is just an electric field of 1 unit. In one-dimensional examples like this, it's easy to take field-line patterns and superpose them, as in u/2. Figure u/3 is a graph of the field as a function of position. It looks like a staircase, with a discontinuity at each charge.

We would like to find a local law that describes the behavior of the field at any point. In the vacuum regions between the charges, this is easy: the field is constant, and if we like, we can describe this by saying that $dE/dx = 0$. At a charge, this becomes awkward, because the field behaves badly. We therefore consider the limit as the number of charges becomes large, u/4, and finally infinite, u/5. The slope of the graph is the charge density,

$$\frac{dE}{dx} \propto \frac{dq}{dx}.$$

In the notation dq , the d is to be interpreted as "a little bit of," i.e., we are talking about the infinitesimal amount of charge contained within the infinitesimal distance dx . The idea is that if dq/dx is large, we have a lot of densely packed stair steps, making the slope of the staircase dE/dx steeper. In a region of vacuum, the density of charge is zero, and we get a constant field. This is the local form of Gauss's law, for field vectors, in one dimension.

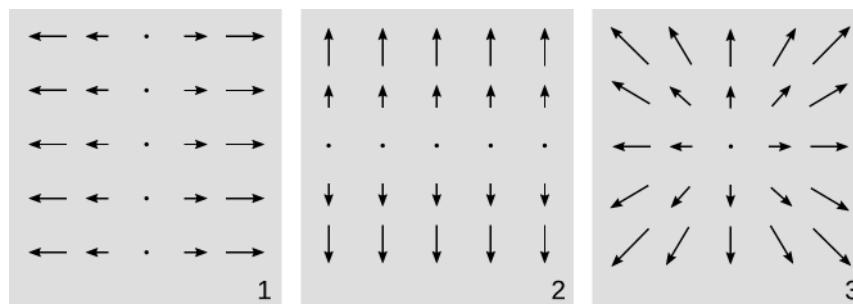
Although our main goal in developing this theory in one dimension was to immediately turn around and generalize it to three dimensions, effectively one-dimensional systems do exist in real life. For example in figure v, charge can be deposited on the series of metal plates, and if the charge on one of the plates is q , then it acts

as a charge density q/h , in units of coulombs per meter, where h is the spacing between the plates.

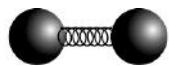
To visualize the meaning of the derivative dE/dx , we display in figure w a hypothetical device that measures this quantity. The extension of the springs is a measure of how much the field diverges. For example, if there are field lines coming out of the middle and going to the right, and also some field lines coming out of the middle and going to the left, then the forces on the two positive charges will be in opposite directions, and the spring will stretch. On the other hand, if the field is constant, $dE/dx = 0$, then the two charges feel equal forces in the *same* direction, and the spring does not stretch (although the device as a whole will accelerate to one side). Since this device measures how much the field diverges, we can refer to it as a “div-meter” (nobody actually uses this term), and we can also refer to dE/dx as the *divergence* of the electric field (people really do use this term).

We want our div-meter to be very small, because what it’s measuring is a kind of derivative, and derivatives are local things. As an example from a more familiar context, we might want to measure the angle or slope of some brickwork, which amounts to measuring the derivative of a function. This doesn’t really work if, as shown in figure x, our measuring device is bigger than the bricks. Similarly, we can get wrong results from the div-meter unless we shrink it down so small that any irregularities in the electric field vanish, and the field becomes smooth and well-behaved.

What about three dimensions? Conceptually, we just need to build the three-dimensional div-meter shown in figure y. What about a more mathematical description? Because the divergence is a kind of derivative, we want it to have the same kind of additive property that the plain old derivative has: the divergence of a sum should be the sum of the divergences. Physically, this is the only way we can guarantee that superposition will work. The following example shows that this essentially ties down the form of the divergence operator.



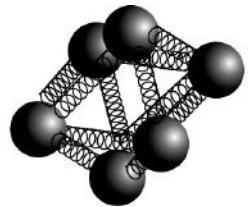
z / Example 10: some diverging fields in two dimensions.



w / The “div-meter:” an imaginary device for measuring the divergence of the electric field, dE/dx , in one dimension. Two positive charges are attached to the ends of a spring. If the field has a positive divergence, then the string stretches.



x / A misguided attempt at accurately measuring the slope of an old brick arch using a cell phone’s leveling app.



y / A three-dimensional div-meter. The expansion of the meter’s volume is a measure of the electric field’s divergence.

Superposition and rotation

example 10

The field in figure z/1 is effectively one-dimensional. It has the form $\mathbf{E} = bx\hat{x}$ (which could also be notated as $\langle bx, 0, 0 \rangle$ or $bx\hat{i}$), where b is a positive constant. For positive x , the field has a positive x component, and conversely on the negative side. We already know the form of the div operator in one dimension, which is that it's simply a derivative, so we have $\text{div } \mathbf{E} = \partial E_x / \partial x = b$. Here the symbol ∂ , called a partial derivative, is like the usual d in calculus, but it says that only the specified variable is being differentiated with respect to, while other variables are held constant. That is, we hold y constant while taking this derivative with respect to x . Since the divergence is constant, we conclude that this is an example in which the charge is distributed exactly evenly, like peanut butter spread evenly on a piece of bread.

Rotating the field by 90 degrees gives z/2. This example is also effectively one-dimensional, and we want our laws of physics to be rotationally invariant, so clearly the divergence must be given by the derivative with respect to y . The result for the charge distribution is the same, which makes sense because charge is a scalar, so it doesn't change when we rotate it.

Now suppose we want the divergence to be additive. If we add the fields in the first two examples, we get the field shown in z/3. It equals $bx\hat{x} + by\hat{y}$ (which can also be notated as $b\hat{r}$, where \hat{r} is a unit vector pointing in the radial direction). Because electric fields obey the law of superposition, the charge distribution in this example must be the sum of the ones in examples 1 and 2, i.e., the divergence must be $2b$.

Example 10 makes it clear that the form of the divergence operator in three dimensions is fixed, up to a constant of proportionality, by superposition, rotational invariance, and consistency with the one-dimensional expression. Detailed calculations with the divergence operator are not part of the logical backbone of this book, so the remaining details of the discussion are relegated to a note ([Z65](#)). Filling in the necessary constant of proportionality, in SI units, the local form of Gauss's law reads

$$\text{div } \mathbf{E} = 4\pi k\rho,$$

where ρ (Greek letter “rho,” which makes the “r” sound) is the density of charge, in units of coulombs per cubic meter.

Earnshaw's theorem

example 11

When kids first start playing with magnets, one of the first things they usually attempt, unsuccessfully, is to levitate a magnet in the air, by using a second magnet either to attract it from above or to repel it from below.

For the electrical version of this, we could place a negative charge in the field of figure z/3. The equilibrium at the center is stable,



aa / The magnetic repulsion between the two doughnut magnets cannot create a stable equilibrium by itself. An additional constraint has to be provided by the pencil.

because if the charge is displaced by some small distance in any direction, it will experience an electrical force pointing back toward the center. But Gauss's law guarantees that this field pattern can never exist in empty space; it can only exist when some density of positive charge is present. This is a special case of Earnshaw's theorem, which states that a system of point charges can never be in stable static equilibrium.

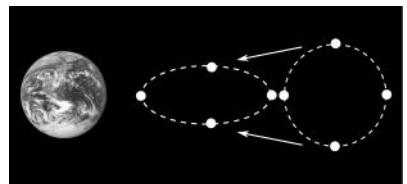
There are more general versions of the theorem that apply to magnetic forces made by permanent magnets such as the ones in figure aa.

Earnshaw makes it difficult to create models that explain why matter is stable. We'll return to this issue in chapter 3.

Gauss's law for gravity

example 12

Because Coulomb's law has the same form as Newton's law of gravity, Gauss's law holds for gravitational fields as well as electric fields, with the charge density replaced by a mass density. The figure shows a set of masses (white dots) released from rest at some distance from the earth. They function as a div-meter. As the cloud of masses free-falls, it becomes distorted. The waist gets narrower, because all the masses are converging on the center of the earth. But the opposite happens in the radial direction: the cloud becomes elongated, because the mass that was closest to the earth felt a stronger force, and the farthest mass a weaker one. The result is that the cloud's volume stays exactly the same. This makes sense, because the cloud is in a region of vacuum, where the density of mass is zero.

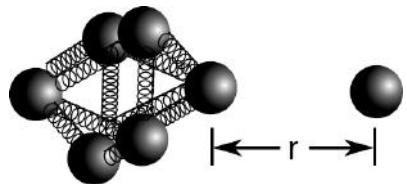


Example 12.

Discussion questions

A For practical reasons, the experiment in example 12 can't be done inside the earth. But suppose we could. What would happen if we released a set of masses all around the earth's surface, and they were able to free fall without being blocked by the dirt? Is this consistent with Gauss's law?

B The figure shows a div-meter next to a charge. All the charges in the figure are positive. A superficial analysis makes it seem as though the result of this experiment could plausibly be consistent with Gauss's law. (Do this analysis.) But what about the following counter-argument? The charge on the right is at some distance r from the nearest charge in the div-meter. If we make r small enough, we can make the force between these two charges grow to any desired size. But the interaction with the other five charges in the div-meter *doesn't* blow up in the limit as $r \rightarrow 0$. So can't we guarantee that Gauss's law will be violated when r is small?



Discussion question B.

Notes for chapter 2

[243](#) Field lines terminating at a point

Field lines don't cross. This is not in contradiction with the fact that multiple field lines can begin or end at a point charge.

We reasoned on p. 27 that field lines can never cross at a point where the field is well defined. However, if we draw the field-line representation of positive and negative point charges, as in figure c on p. 43, we find that the field lines begin at one point and terminate on the other. They do not literally cross, since they don't pass through one another, but our earlier logic still holds, and we find that the field must be undefined at these two points. It is undefined because it is infinite. At a pointlike source, the field blows up to infinity.

[252](#) 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 that can change if we move q around. When we move the charge by some small amount, the field \mathbf{E}_q only changes significantly in the space near the charge, where the ambient field \mathbf{E} has essentially a constant value, equal to its value at the position of the charge. From now on, we write \mathbf{E} to mean this field, and the work done on the charge is proportional to it. Since \mathbf{E}_q is proportional to q (as we can easily prove by Gauss's law), it follows that the work done in this small displacement 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 [2167](#).) 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. 52.

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.

253 No force on a charge from its own field

The force acting on a point charge is only that due to the field created by other charges.

In classical, as opposed to quantum, physics, an object always has a well-defined position in space at any given time. This book is about classical electromagnetism, and it is ultimately impossible to incorporate pointlike charged particles in such a theory. If a pointlike particle exists at a certain position in space, then the field blows up to infinity at that point, and this leads to all kinds of bad things. For example, the integral $\int E^2 dv$ diverges, so the energy of the field is infinite, and this implies via Einstein's $E = mc^2$ (sec. 5.5, p. 136) that the particle has infinite inertia, which it doesn't in reality. Quantum physics eliminates these problems, for example by replacing a pointlike electron at a definite point in space with a fuzzy, probabilistic electron cloud.

It is nevertheless convenient to be able to talk about point charges in classical electromagnetism, and we can get away with it as long as we don't try to describe any phenomena below a certain length scale (known as the classical electron radius) and keep in mind some fairly common-sensical rules. One of these rules is that it certainly wouldn't make sense to try to calculate the force of an electric field on

a charge without excluding the charge's own contribution to the field. Such a force would be infinite, and it wouldn't have a well-defined direction. (A less definitive argument for this rule is that according to Newton's third law, objects can't make forces on themselves. The trouble with this is that Newton's third law is in general false in electromagnetism. What is true is that momentum is conserved, and this momentum has to include the momentum of the fields themselves. See sec. 5.2.1, p. 123.)

262 Divergence operator in three dimensions

We present the form of the divergence operator and Gauss's law in three dimensions.

Example 10 on p. 62 suggests that the three-dimensional form of the divergence operator should be

$$\text{div}\mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}.$$

If the generalization of the one-dimensional operator to three dimensions exists, then it must have this form. If it is to be local, then it can only depend on the field and its derivatives. Superposition rules out expressions such as $(\mathbf{E} \cdot \mathbf{E})(\partial E_x / \partial x)$ in which the field is combined with its derivatives. Only first derivatives are possible, because otherwise the units would not work out. Rotational invariance rules out derivatives such as $\partial E_x / \partial y$. There is nothing left to try but terms having the forms of the three terms given above in the claimed form of the operator. We can make linear combinations of these, such as $7\partial E_x / \partial x - 3\partial E_y / \partial y$, but if the coefficients are unequal, it will violate rotational invariance. The only possibility left is the expression claimed above, or some constant multiple of it, so we arbitrarily fix the constant to be 1.

Any three-dimensional, local version of Gauss's law must therefore have the form

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \propto \rho.$$

The constant of proportionality can only be fixed by making contact with the global form

of Gauss's law. A theorem called Gauss's theorem says that the local and global forms of Gauss's law are equivalent, in a sense that is similar to the fundamental theorem of calculus. Explicitly, Gauss's theorem says that when we integrate the divergence of a field over a region of space, the result is the same as the flux through the surface of the region. To show that the constant of proportionality in Gauss's law is $4\pi k$, we can apply Gauss's theorem to a spherical region centered on a point charge.

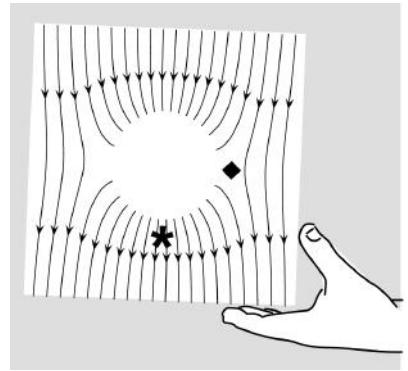
Problems

Key

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

1 Mayor Max is mayor of Idyllwild, California. He is a golden retriever. Because some of his fans are a little crazy, his security detail carefully examines all packages sent to him. The figure shows a cubical box he's received, which contains the electric field pattern indicated by the drawing of the field lines. (a) Where are electric charges located in the box, and what are their signs? (b) Compare the strength of the electric field at the points marked with the diamond and the star. Is there charge at these points? (c) What is the total charge of the box?

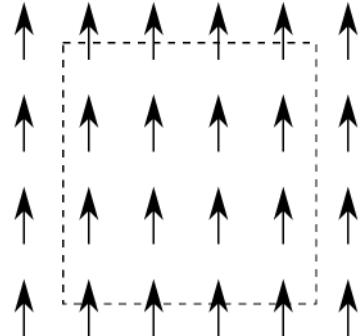
▷ Solution, p. 426



Problem 1.

2 The figure shows a uniform electric field of magnitude E , and a side view of a cubical imaginary surface with edges of length h .

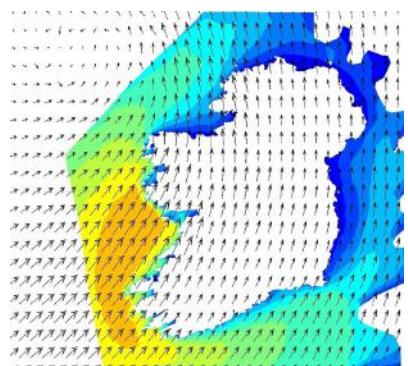
- (a) Find the flux through the top of the cube. ✓
- (b) Find the flux through each side of the cube. ✓
- (c) Find the flux through the bottom of the cube. ✓
- (d) Find the total flux through the cube. ✓
- (e) Find the electric charge contained inside the cube. ✓



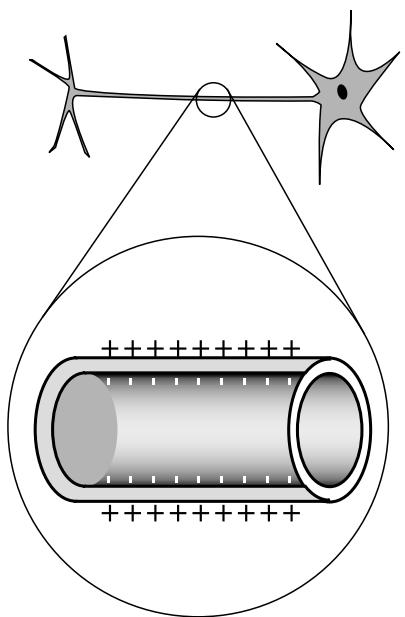
Problem 2.

3 Surfers on the western coasts of Europe in the winter of 2013–2014 experienced waves of historic proportions, up to 18 meters tall. Meteorologists deal with various quantities that are functions of position and can be plotted on a map. Some of these quantities, such as wave height, temperature, and pressure, are scalars. Others, such as wind velocity and ocean currents, are vectors, which we would notate in boldface. The map in the figure, from Ireland in December 2013, superimposes plots of wave height (colors) and wind velocity (arrows). Explain why each of the following expressions doesn't make sense for a meteorologist to talk about, purely on "grammatical" grounds. (Cf. problem 5, p. 33.)

- (a) $\operatorname{div} P$
- (b) $\mathbf{v} + \operatorname{div} \mathbf{w}$



Problem 3.



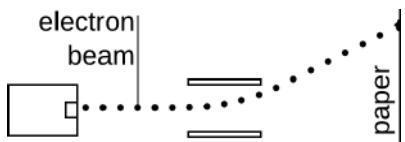
Problem 4. Top: A realistic picture of a neuron. Bottom: A simplified diagram of one segment of the tail (axon).

4 The figure shows a neuron, which is the type of cell your nerves are made of. Neurons serve to transmit sensory information to the brain, and commands from the brain to the muscles. All this data is transmitted electrically, but even when the cell is resting and not transmitting any information, there is a layer of negative electrical charge on the inside of the cell membrane, and a layer of positive charge just outside it. This charge is in the form of various ions dissolved in the interior and exterior fluids. Why would the negative charge remain plastered against the inside surface of the membrane, and likewise why doesn't the positive charge wander away from the outside surface?

5 The gravitational field g of a typical asteroid is so weak that it is just a loose pile of rubble, gravel, and dust. In this environment, the ambient electric field E can make a strong enough force on a small particle of dust so that it competes on fairly even terms with gravity.

(a) Suppose that a particle of mass m and charge q is disturbed so that it is shot upward from the surface with initial velocity v . Find the height h to which it rises. Use a coordinate system in which up is positive, so that g will be negative, v and h positive, and E either positive or negative depending on the direction of the field. ✓

(b) Evaluate your result for $v = 1.0 \times 10^{-1}$ m/s and the following data, which are meant to be fairly realistic for an asteroid the size of a football field: $g = -1.0 \times 10^{-4}$ m/s 2 , $E = 40$ V/m, $m = 1.0 \times 10^{-8}$ kg, $q = 1.0 \times 10^{-14}$ C. ✓



Problem 6.

6 (a) In an inkjet printer, a series of droplets of ink, each with mass m , are squirted out in rapid succession at speed u . They are then given charge q by the beam of an electron gun, and finally deflected by passing through a capacitor whose electric field E is in the perpendicular direction. Let the width of the capacitor be w , and assume that the electric field is uniform between its plates but zero outside. (We'll see in section 4.1, p. 85, that this cannot quite be true, and is at best an approximation.) Find the deflection angle θ . ✓

(b) Evaluate your result, in radians, for $m = 2.0 \times 10^{-10}$ kg, $u = 20$ m/s, $q = 2.0 \times 10^{-10}$ C, $E = 2.0 \times 10^5$ V/m, and $w = 3$ mm. ✓

7 (a) At time $t = 0$, a positively charged particle is placed, at rest, in a vacuum, in which there is a uniform electric field of magnitude E . Write an equation giving the particle's speed, v , in terms of t , E , and its mass and charge m and q . \checkmark

(b) If this is done with two different objects and they are observed to have the same motion, what can you conclude about their masses and charges? (For instance, when radioactivity was discovered, it was found that one form of it had the same motion as an electron in this type of experiment.)

8 Suppose that at some instant in time, a wire extending from $x = 0$ to $x = \infty$ holds a charge density, in units of coulombs per meter, given by ae^{-bx} . This type of charge density, dq/dx , is typically notated as λ (Greek letter lambda). Find the total charge on the wire. \checkmark

9 Example 3 on p. 51 calculated the field of a dipole in its mid-plane. (a) Calculate its field at a point on its axis (i.e., the line through the charges), at a distance r from the center ($r > \ell/2$). (b) Show that at large distances your result from part a is approximately proportional to $1/r^3$, as in the mid-plane. \triangleright Hint, p. 425

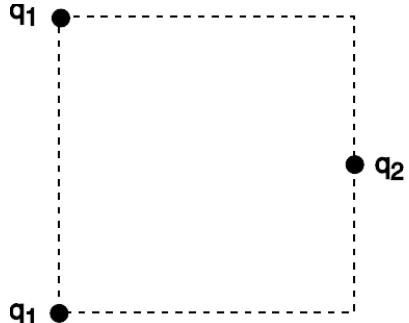
10 Example 9 on p. 59 described the choice of a Gaussian surface in order to calculate the electric field of a uniformly charged plane, with charge density σ in units of coulombs per square meter. (a) Complete the calculation in order to find the field. (b) Check that the units of your answer make sense. \triangleright Solution, p. 426

11 Suppose a capacitor consists of two parallel metal plates with area A , and the gap between them is h . The gap is small compared to the dimensions of the plates. Since the plates are metal, the charges on each plate are free to move, and will tend to cluster themselves more densely near the edges due to the mutual repulsion of the other charges in the same plate. However, it turns out that if the gap is small, this is a small effect, so we can get away with assuming uniform charge density on each plate. Find the field between the plates when the charges on the plates are q and $-q$. Use the result of problem 10, which has a solution in the back of the book. \checkmark

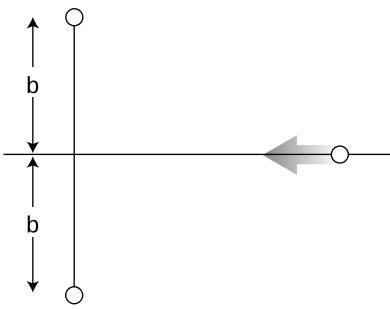
12 Three charges are arranged on a square as shown. All three charges are positive. What value of q_2/q_1 will produce zero electric field at the center of the square? \checkmark

13 (a) Use Gauss' law to find the field inside an infinite cylinder with radius b and uniform charge density ρ . Use the technique demonstrated in example 7, p. 55 \checkmark

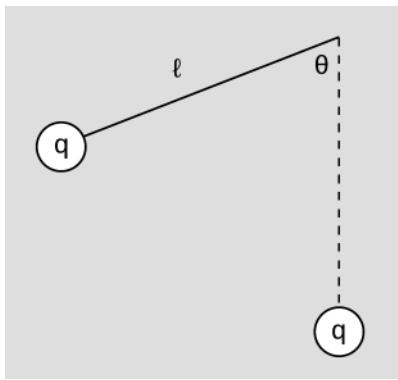
- (b) Check that your answer makes sense on the axis.
(c) Check that the units of your answer make sense.



Problem 12.



Problem 14.



Problem 16.

14 In the semifinals of an electrostatic croquet tournament, Jessica hits her positively charged ball, sending it across the playing field, rolling to the left along the x axis. It is repelled by two other positive charges. These two equal charges are fixed on the y axis at the locations shown in the figure. (a) Express the force on the ball in terms of the ball's position, x . (b) At what value of x does the ball experience the greatest deceleration? Express your answer in terms of b . [Based on a problem by Halliday and Resnick.] \checkmark

15 Several pointlike, interacting particles are released, all initially at rest, within the same finite region of space. We want to know whether, without violating conservation of energy, it is possible for one of these particles to be ejected to an infinite distance, while the others stay within the original region. Consider this question for each of the following systems. (a) Two particles of mass m , interacting gravitationally. (b) Three such particles. (c) Two particles, both with charge q . (d) Charges q and $-q$. (e) Charges q , q , and $-q$. \star

16 As shown in the figure, a particle of mass m and charge q hangs from a string of length ℓ , forming a pendulum fixed at a central point. Another charge q is fixed at the same distance ℓ , directly below the center. Find the equilibrium values of θ and determine whether they are stable or unstable. \star

17 The following point charges are located at the following positions in the Cartesian plane:

- $7q$ at $(0, a)$
- $7q$ at $(0, -a)$
- $5q$ at $(2a, 0)$

Both q and a are positive. Find the direction of the electric field at the point $(a, 0)$.

18 The following point charges are located at the following positions in the Cartesian plane:

- q at $(0, 0)$
- q at $(a, 0)$

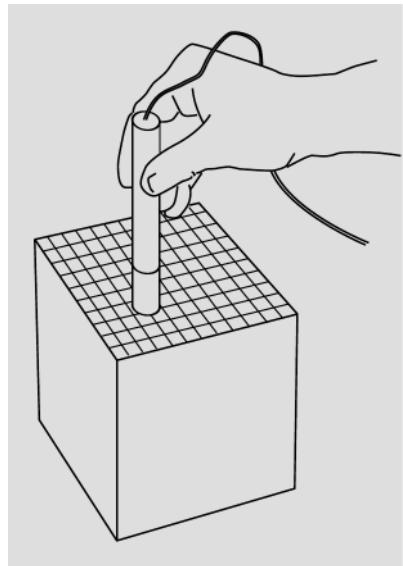
Both q and a are positive. (a) Find the direction of the electric field at the point (a, b) , where b is positive. State your result as an angle defined in the usual way, counterclockwise from the x direction. (b) Show that the units of your answer make sense. (c) Show that your answer has the correct limiting behavior for $b \rightarrow 0$ and $b \rightarrow \infty$.

Minilab 2: Gauss's law for magnetism

Apparatus

- magnetic field sensor
- box containing unknown magnet
- graph paper with 1 cm grid

Goal: Verify Gauss's law for magnetism, using an unknown magnet sealed inside a box.



a / Measuring the flux through the box.

The unknown will be provided to you inside a cubical cardboard box about 17 cm on a side. The magnetic field sensor is a wand like the one shown in figure i on p. 22, which displays the component of the field parallel to the wand's axis, at a point near the tip, as in figure g on p. 46.

The sensor has two scales, selected using a switch on the wand. You probably need it to be on the less sensitive scale in order to avoid overloading it in this experiment.

With the box moved far away, hold the sensor in the air, oriented vertically, in the position shown in the figure, and use the software to zero the sensor.² This means that the sensor will now consider the vertical component of

²If using LoggerPro, this is done using the blue zero icon.

the ambient field in the room to be zero.

The following technique can be used to determine the flux through one side of the box. Orient the box with that side up. Place the graph paper on top of it. Watching the second hand on a clock, get a one-second beat going. It may be helpful to have your partner rap their knuckles on the table. In the software, initiate a period data collection preset to last 17 seconds.³ Start collecting data, and simultaneously begin scanning the sensor down the rows of the graph paper, covering a 1 cm by 17 cm strip in one second. Use the software to average the data, which gives a measure of the average flux per unit area.⁴ Multiplying by the area gives the flux.

Adding the fluxes through all six sides of the cube provides a test of Gauss's law.

Theoretically the validity of Gauss's law should not depend on whether there are additional contributions to the magnetic field from the ambient field in the room. However, it is easier physically to rotate the box rather than scanning the different sides while the box stays in one orientation. This means that there can be a contribution from the ambient field. Zeroing the sensor approximately gets rid of this, but only on the assumption that the ambient field is uniform. Because the ambient field is mostly from building materials and magnetic materials inside the lab benches, the field is not exactly uniform. To do a more thorough job of getting rid of the effect, you can wave the wand in the air where the top of the box *would* be, but with the box removed. The result is the contribution to each flux measurement from the ambient field, so the final result can be corrected by subtracting six times this number.

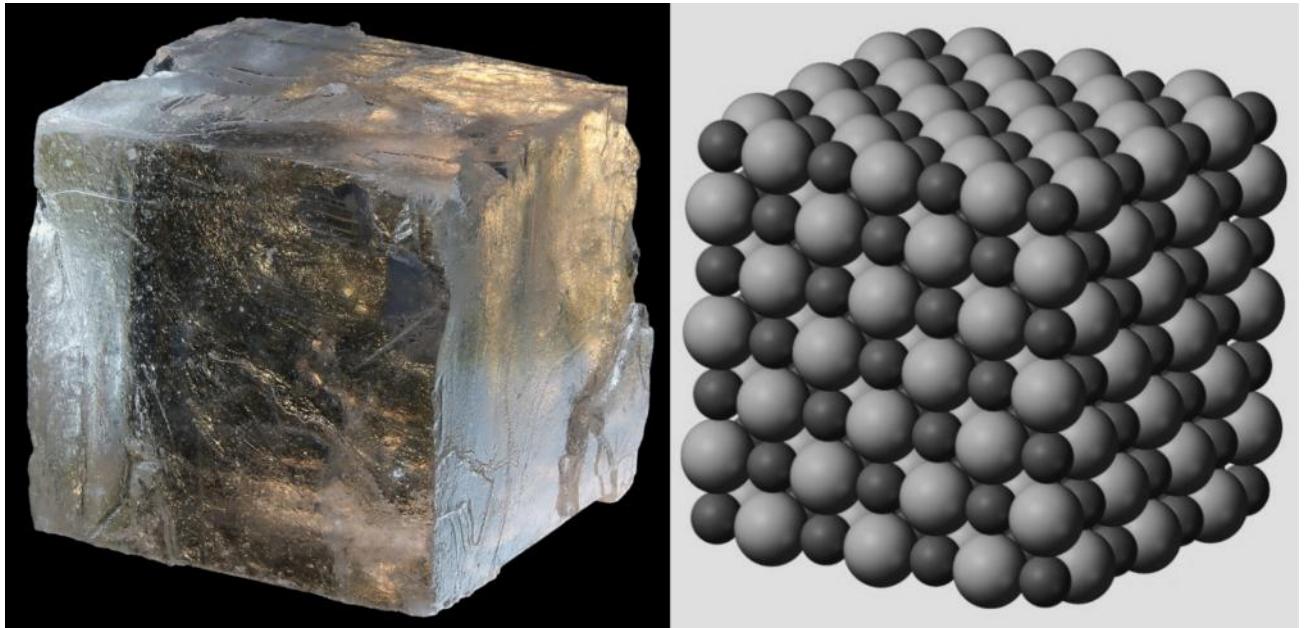
To estimate the experimental uncertainty, pick one side and do a series of measurements. The spread σ among these numbers gives an esti-

³In LoggerPro, do this using Experiment>Data collection.

⁴In LoggerPro, this is done through Analyze:Statistics.

mate of the uncertainty which is probably reasonable to apply independently to each of the six faces. Since it is the squares of errors that add in propagation of errors, the total error is $\sqrt{6}\sigma$.

If you have time, it's also a good idea to do additional measurements of all six fluxes, so as to be able to detect mistakes.



A salt crystal, and a model of its structure at the atomic level.

Chapter 3

Models of matter

3.1 Binding energy of matter: two examples

We know from everyday life that energy is required to break a stick, boil water, or drive photosynthesis in plants. In all of these examples, we have atoms or molecules bound together with electrical forces, and separating or rearranging them requires an increase in potential energy. The following two examples demonstrate how we can model this energy by using what we know about electrical potential energy.

Energy of a set of charges

example 1

▷ Consider the molecule shown in figure a, with an atom of a certain element in the middle, surrounded by five atoms of a different element, arranged along perpendicular axes, all at equal distances ℓ from the central one. Let's take the charge of the central atom to be $5q$, and the charge of each of the others to be $-q$, so that the molecule is electrically neutral. If we approximate the atoms as point charges, what is the total electrical energy released when the molecule is assembled, starting with its constituent parts all far away from each other?

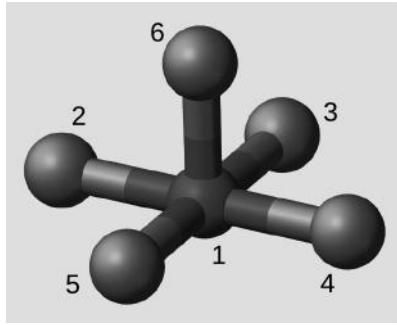
▷ Each charge interacts with *all* the others, not just the ones shown as connected to it in this ball-and-stick diagram. If we number the atoms 1 through 6, then we have to compute an interaction for each of the combinations, such as 12, 23, and so on. We don't want to count the self-energies like 11 or 22, and we also don't want to double-count any energies, e.g., 12 and 21 are not two separate terms in the sum. In sigma notation, we can notate this conveniently in either of the following equivalent forms:

$$\sum_i \sum_{j>i} U_{ij} = \frac{1}{2} \sum_i \sum_{j\neq i} U_{ij}.$$

In the second version, we double-count, but then divide by two at the end to compensate.

This particular example has $n = 6$ charges, but we can see in general, using the second form of the sum above, that the total number of energies will be $(n^2 - n)/2$, since we have n^2 choices for the indices i and j , n of which are ruled out by the condition $i \neq j$. This formula, which you may want to verify for $n = 1, 2$, and 3 , can also be written as $n(n - 1)/2$. It is the number of combinations of n things taken 2 at a time, disregarding their order, and is notated $\binom{n}{2}$, also known as a binomial coefficient. In our example, we have $\binom{6}{2} = 15$.

Numbering the atoms as shown in the figure, we have the following contributions to the sum:



a / A molecule.

$$\begin{aligned}
 U &= U_{12} + \text{four more similar terms, for a total of five} \\
 &\quad + U_{23} + \text{three more similar terms, for a total of four} \\
 &\quad + U_{24} + \text{one more similar term, for a total of two} \\
 &\quad + U_{26} + \text{three more similar terms, for a total of four} \\
 &= 5U_{12} + 4U_{23} + 2U_{24} + 4U_{26}.
 \end{aligned}$$

This is a total of 15 terms, which checks. Computing these ex-

plicitly, we have

$$\begin{aligned} U &= kq^2 \left(\frac{5(-5)}{r_{12}} + \frac{4}{r_{23}} + \frac{2}{r_{24}} + \frac{4}{r_{26}} \right) \\ &= \frac{kq^2}{\ell} \left(\frac{-25}{1} + \frac{4}{\sqrt{2}} + \frac{2}{2} + \frac{4}{\sqrt{2}} \right) \\ &= \frac{kq^2}{\ell} \left(-24 + \frac{8}{\sqrt{2}} \right). \end{aligned}$$

The negative sign indicates that the system is bound with respect to complete dissociation: it would not be possible without an input of energy to pull it apart and separate all the ions from one another.

Being able to do the calculation does not mean that this is a good model of any real-world molecule. There are molecules such as ClF_5 that have this geometry, but our model is not consistent with their having this shape as a stable equilibrium, for the energy could be made more negative either by making ℓ smaller or by moving the small charges so that they were more uniformly distributed on the sphere, rather than clustered in a half-sphere. This is an example of a more general fact, which is that classical (as opposed to quantum) physics cannot explain the stability of matter. The electrostatic energy we have calculated *does* come into play, and is roughly right for this molecule, but other factors are present as well.

A long molecule

example 2

Bulk matter, as opposed to an individual molecule, contains a large number of charges, and in many cases it makes sense to consider the material as infinite in extent and compute quantities such as electrical energy per unit of material. A simple example that has some of these characteristics is a long molecule, which is one-dimensional but can be idealized as infinite. An idealized, classical version of such a molecule is shown in figure b, in which positive and negative charges alternate along a straight line. There are real-world long molecules, although most are not exactly linear (for example, DNA or a type of plastic known by the trade name Delrin), and many (including DNA and Delrin) cannot reasonably be modeled by this type of charge distribution. However, there are some compounds such as cyclopentadienyl-lithium that are somewhat like figure b, albeit with some of the charges actually being molecules like pentane rather than individual atoms. In any case, let's play with this as a "toy model" of bulk matter.

... $\Theta \oplus \ominus \oplus \ominus \oplus \Theta \ominus \oplus \Theta \oplus \ominus \dots$

b / An idealized model of a long molecule.

As in example 1, the total energy is a sum

$$U = \sum_i \sum_{j>i} U_{ij}.$$

Surprisingly, this comes out to be quite a bit simpler to evaluate than the one in example 1. We are actually interested not in the total energy, which would be infinite, but in the energy per atom. If we fix the index i to some arbitrary value, then we have singled out one atom, and the inside sum

$$\sum_{j>i} U_{ij}$$

can be interpreted as the energy per atom. If we label each atom by an integer, in sequence, then the restriction of the sum to $j > i$, to avoid double-counting, means that we can evaluate the interaction of this atom only with the ones on its right. If the charges are $\pm q$ and the inter-atomic spacing ℓ , then the total energy is

$$U_{\text{per atom}} = \frac{kq^2}{\ell} \left(-1 + \frac{1}{2} - \frac{1}{3} + \dots \right).$$

The sum in parentheses is a famous one with a name, the alternating harmonic series, and it has the value $-\ln 2$, which can be proved, if you know about Taylor series, by taking the Taylor series of $\ln(1+x)$ and evaluating it at $x=1$. The result is

$$U_{\text{per atom}} = -\frac{kq^2 \ln 2}{\ell}.$$

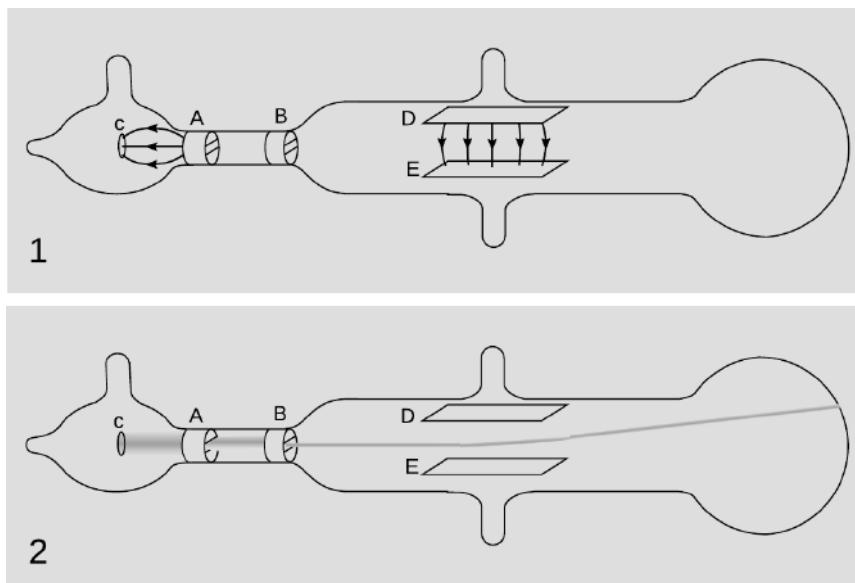
Although this toy model has the same shortcomings as the one in example 1, it does provide some interesting insight by showing us, correctly, that the electrical energy of matter is not very localized, as you might have expected from ball-and-stick models. The sum contains contributions from pairs of atoms at arbitrarily large distances, and contributions from very distant terms in the series make quite big contributions — if you try evaluating the alternating harmonic series on a calculator, you will find that it is extremely slow to converge.

For a similar description of a three-dimensional solid, see Purcell, Electricity and Magnetism, section 1.6.

3.2 The discovery of the electron, and the raisin cookie model

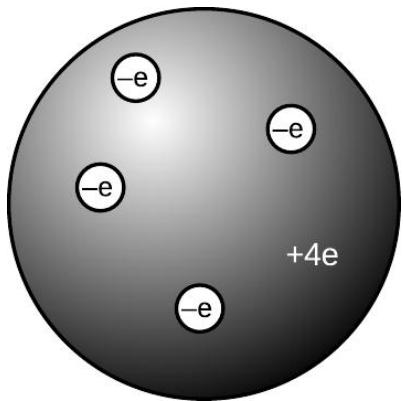
Static electricity experiments such as the one with the sticky tape (figure a, p. 41) work with essentially any substance (although they do produce bigger effects with some substances than with others). This suggests that electric charge is a built-in feature of all matter. Quantization of charge seems like a hint that this charge occurs because matter is made out of particles, and those particles have charges that are integer multiples of e .

Until about 1897, it was assumed that all such particles were about the size and mass of an atom. But in that year, J.J. Thomson did the experiment shown in figure c, in which he produced a beam of charged particles in a vacuum tube, subjected them to a transverse electric field, and measured their deflection. The particles are initially accelerated from C to A, at which point they are moving along the x axis with velocity u . Let's assume that this velocity is known, although the actual technique Thomson used to determine it is one that we will not describe until sec. 5.3, p. 131. The electric field between D and E is in the negative y direction. This field produces a force along the y axis, but does not affect the x motion, just as the earth's gravity does not affect the horizontal velocity of a baseball. To travel the length ℓ of electrodes D and E, the time required is $t = \ell/u$. Now t turns out to be many orders of magnitude too short to be detectable by eye, but during this interval the acceleration $a_y = qE/m$ caused by the electric field deflects the beam by a distance $y = (1/2)a_y t^2$. Since y is known, we can infer the value of q/m . As we have seen in example 4, p. 52 (the spark plug), this charge-to-mass-ratio is the *only* property of a particle that we can ever infer from its motion in an electric field.



c / Thomson's experiment proving cathode rays had electric charge (redrawn and simplified from his original paper). Panel 1 shows the glass vacuum tube, metal electrodes, and electric field lines. Electric charge is introduced to the electrodes through wires (not shown) that feed through the nipples in the glass. This creates the electric fields. Panel 2 shows the beam of electrons. Because the electrons are negatively charged, they accelerate in the direction opposite to that of the electric field.

The value that Thomson found for the q/m of his mystery parti-



d / The raisin cookie model of the atom with four units of charge, which we now know to be beryllium.

cles was negative, and, more importantly, thousands of times bigger than the q/m ratio of charged atoms (ions) as determined from chemistry experiments. Thomson guessed (without actually knowing for sure) that the charge of the particles was $-e$, and that therefore they had masses thousands of times smaller than the mass of an atom. He interpreted this as evidence that he had discovered the first subatomic particle, which was later named the electron.

Based on his experiments, Thomson proposed a picture of the atom which became known as the raisin cookie model. In the neutral atom, figure 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.

If you’re inclined to dismiss the raisin cookie model as silly, note that it evades Earnshaw’s theorem (example 11, p. 62), since it is not made solely out of point charges.

3.3 The energy scale for chemistry and atomic physics

The sizes of atoms were originally estimated based on a variety of techniques, one example being the determination of the thinnest layer of oil that cold be deposited on the surface of water. The results are on the order of a fraction of a nanometer. The smallest atom, hydrogen, has a radius of roughly 0.05 nm. A hydrogen atom is composed of charges $+e$ and $-e$, so we can estimate the electrical potential energy of a hydrogen atom to be something like 5×10^{-18} J. To within an order of magnitude, this is the energy scale of all of chemistry and atomic physics. For example, a human body contains some number of atoms in the form of fat molecules, and multiplying this by the energy scale found above results in a rough estimate of the size of the energy reserve we all carry around with us — typically on the order of 100,000 food calories, or several months’ worth of survival without food.

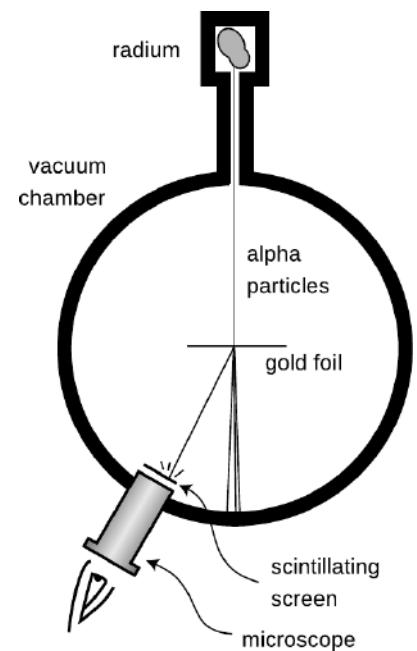
3.4 The nucleus and the planetary model

In 1909, a new experiment by Ernest Rutherford and collaborators forced a revision of the raisin cookie model. As shown in figure e, the experiment used a lump of radium, which was known to emit radioactivity in a form known as alpha particles, another name for a helium atom that has been stripped of both its electrons. The alpha particles, moving at a significant fraction of the speed of light, struck a very thin gold foil, and most of them passed almost straight through it. But some were deflected by measurable angles, and the experimenters found, to their great surprise, that some rebounded at angles approaching 180 degrees. 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 to hit you.”

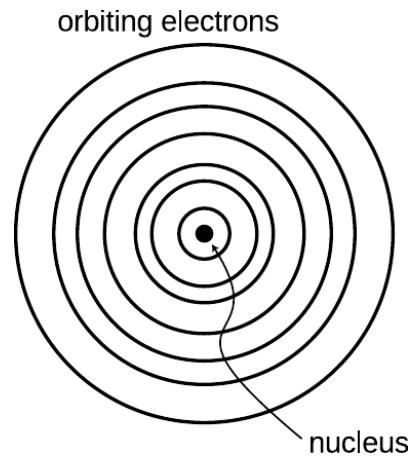
This observation proved impossible to explain in the raisin cookie model, because the kinetic energies of the alpha particles were many orders of magnitude greater than the energy scale estimated in sec. 3.3. At this point, the Rutherford group dusted off an unpopular and neglected model of the atom, 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 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.

The nucleus was later found to be a cluster of particles called protons, with charge $+e$, and electrically neutral particles called neutrons. The number of protons is notated Z , and referred to as the atomic number. The total number of neutrons and protons is notated A , and for light, stable nuclei is usually about equal to $2Z$, i.e., the numbers of neutrons and protons are roughly equal.

There is a problem with the stability of the planetary model. Although it evades Earnshaw’s theorem, because the electrons are orbiting rather than sitting in equilibrium, there is still a problem because, as we will see in ch. 6, an accelerating electric charge radiates light. This would have caused the orbiting electrons to lose energy and spiral into the nucleus. The resolution of this problem would have to wait until the classical picture of the atom was replaced by one involving the new field of quantum physics.



e / Rutherford’s apparatus.



f / The planetary model of the atom.

3.5 The energy scale for nuclear physics

For nuclear reactions, experiments show that the energy scale is about 10^{-13} J per neutron or proton, or about 6 orders of magnitudes more than the scale for chemical reactions discussed in sec. 3.3. Let's compare this with the electrical energy that we expect a nucleus to have.

Often, as in figure f, p. 79, we visualize the nucleus of an atom as a tiny dot, a point with no size. But this is impossible, because then the electrical potential energy of the protons would be infinite. Actually, the radius of a medium-sized nucleus is on the order of $r \sim 3 \times 10^{-15}$ m, and we can take this as a typical distance between the electrically interacting protons. Using this distance as an input, we can estimate (281) the electrical energy of a nucleus, divided by the number of neutrons and protons, to be $\sim (+2 \times 10^{-14}$ J)($Z - 1$).

Subject to all the crude assumptions that went into this order-of-magnitude estimate, this is in the right ballpark. It would therefore be tempting therefore to try to interpret nuclear processes as involving purely electrical interactions, just like chemical processes. On this interpretation, the vast disparity between chemical and nuclear energy scales would be solely due to the very different values of $1/r$.

But there are two things that prevent this from working. First, we do observe energetic nuclear reactions and excitations that occur when there is only one proton on the scene, but in such cases the electrical energy, which is proportional to $Z - 1$, would be zero — there is nothing for the proton to interact with electrically. Second, the sign of the energy is positive, indicating that energy is required in order to assemble the system and could be released by letting it fly apart. The purely repulsive electrical interaction between the like charges of the protons can never explain why nuclei are *bound*. We conclude that although electrical energy may often be big enough to matter in nuclear processes, there must be some other interaction involved as well. This is called the strong nuclear force.

Notes for chapter 3

280 Estimate of the electrical energy in a nucleus

Starting from a rough estimate of the size of a typical nucleus, we make an order-of-magnitude estimate of its electrical energy.

To estimate the energy, we need to know how many pairs of protons are repelling one another. This is the binomial coefficient $\binom{Z}{2} = Z(Z - 1)/2$, as in example 1, p. 74.

To estimate the electrical energy of a nucleus, per neutron or proton, we can therefore calculate

$$\begin{aligned}\frac{k[Z(Z - 1)/2]e^2}{r} \cdot \frac{1}{A} &\approx \frac{ke^2}{r} \cdot \frac{Z(Z - 1)}{2} \cdot \frac{1}{2Z} \\ &= \frac{k(Z - 1)e^2}{4r} \\ &\sim (+2 \times 10^{-14} \text{ J})(Z - 1).\end{aligned}$$

Problems

Key

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

1 Verify the calculation in sec. 3.3, p. 78, that estimated the energy scale for chemistry and atomic physics.

2 A carbon dioxide molecule has a symmetrical, straight shape, with the carbon atom sitting at the midpoint between the two oxygens. Suppose that we model this as a set of three point charges, like this: $Q - q - Q$. (a) Find the ratio q/Q that produces a static equilibrium. ✓

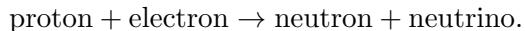
(b) Show that the total charge is nonzero.

(c) Find the total electrical potential energy. ✓

Remark: This is not a satisfactory model for several reasons. The CO_2 molecules in our environment are normally electrically neutral and nevertheless don't spontaneously break up. The equilibrium would also be unstable (determinable directly, or from Earnshaw's theorem without the need for a calculation). This is an example of the fact that classical theories of matter fail, and we need quantum physics.

3 The subatomic particles called muons behave exactly like electrons, except that a muon's mass is greater by a factor of 206.77. Muons are continually bombarding the Earth as part of the stream of particles from space known as cosmic rays. When a muon strikes an atom, it can displace one of its electrons. If the atom happens to be a hydrogen atom, then the muon takes up an orbit that is on the average 206.77 times closer to the proton than the orbit of the ejected electron. How many times greater is the electric force experienced by the muon than that previously felt by the electron? ✓

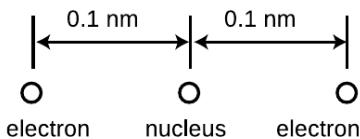
4 The radioactive decay of the neutron was discussed on p. 57. A similar process is electron capture,



The only particle here that has not been previously mentioned is the neutrino, but as you might guess, it's electrically neutral, just like the antineutrino. Verify that electric charge is conserved in this type of decay.

Remark: Luckily for us, this process doesn't occur spontaneously in the hydrogen atom — it requires an input of energy to make it go. It does occur spontaneously, however, for some nuclei.

5 A helium atom finds itself momentarily in this arrangement. Find the direction and magnitude of the force acting on the right-hand electron. The two protons in the nucleus are so close together ($\sim 1 \text{ fm}$) that you can consider them as being right on top of each other. ✓

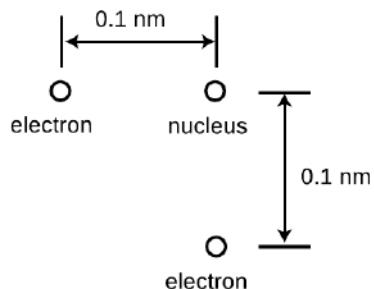


Problem 5.

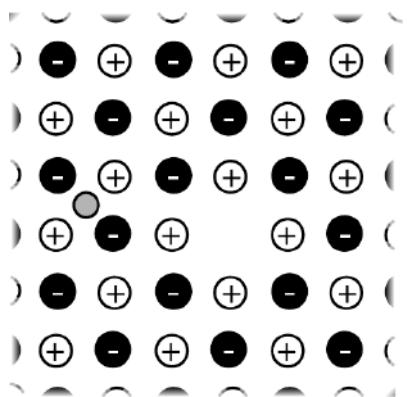
6 The helium atom of problem 5 has some new experiences, goes through some life changes, and later on finds itself in the configuration shown here. What are the direction and magnitude of the force acting on the bottom electron? (Draw a sketch to make clear the definition you are using for the angle that gives direction.) ✓

7 The figure shows one layer of the three-dimensional structure of a salt crystal. The atoms extend much farther off in all directions, but only a six-by-six square is shown here. The larger circles are the chlorine ions, which have charges of $-e$, where $e = 1.60 \times 10^{-19}$ C. The smaller circles are sodium ions, with charges of $+e$. The center-to-center distance between neighboring ions is about 0.3 nm. Real crystals are never perfect, and the crystal shown here has two defects: a missing atom at one location, and an extra lithium atom, shown as a grey circle, inserted in one of the small gaps. If the lithium atom has a charge of $+e$, what is the direction and magnitude of the total force on it? Assume there are no other defects nearby in the crystal besides the two shown here.

▷ Hint, p. 425 ✓ *



Problem 6.



Problem 7.

Chapter 4

The electric potential

4.1 Something is missing

Sometimes in physics, as in life, we may not realize that there is a piece missing from a puzzle. To see that there is such a missing piece in our description of the electric field, consider the following seemingly minor technical issue. In figures a/1 and a/2, we have two fields, either of which we could imagine was the correct field for a parallel-plate capacitor in a state of static equilibrium.

We would expect some physical law to tell us which was correct, but either is consistent with Gauss's law, because the field lines begin on positive charges and end on negative ones. This is a hint that we need some additional law of physics if we are to be able to fully predict the behavior of electric fields.

Experiments show that a/2 is right and a/1 is wrong, and a hint as to what's going on is provided by the following argument. Suppose we take a positive test charge and move it around the rectangular closed path ABCDA, shown in figure a/3 superimposed on the simpler-looking field pattern from a/1, in which the field is zero outside the capacitor. The field does positive work on the charge from B to C, but zero work along the other edges of the rectangle. Energy is released, but nothing has changed about the field pattern, so we can repeat the cycle as many times as we like. This is a perpetual motion machine, and it violates conservation of energy.

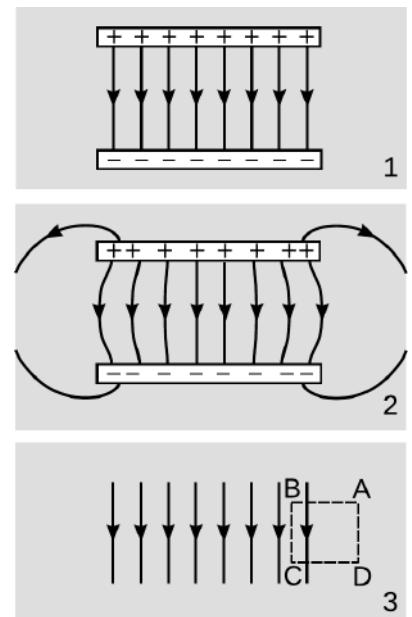
The field in figure a/2 fixes the problem. Although the field is weaker outside the capacitor, it does some negative work along DA. Furthermore, the field has horizontal components that do negative work along AB and CD. The field has been carefully contrived so that the total work done by the field around the rectangle is zero.

For a static field, i.e., one that doesn't change as a function of time, we want this to be true in general:

The work done by a static electric field on a test charge, as the charge moves around a closed loop, is always zero.

A field that satisfies this criterion is referred to as conservative or, for reasons that will become clear soon, irrotational.

Non-static electric fields do not have to be conservative. For example, when you go out in the sun on a hot day, the sunlight heats



a / Which is the correct electric field for the parallel-plate capacitor?

your skin. As the light wave oscillates, it does work on the charged particles in your body, and it does so over and over again, for as long as you continue your sunbath. Although we refer to this electric field as “nonconservative,” there is no violation of conservation of energy. The energy gained by your body as heat is balanced by the loss of the electric and magnetic energy of the light waves, which pass into nonexistence as they are absorbed.

For the rest of this chapter we restrict our attention to static electric fields, which are conservative. We expect the laws of physics to be expressible in a way that is purely local, so it should be good enough if the conservative property of a static electric field holds in the limit of very small closed loops. In fact, we will see that if the conservative property holds for small loops, then it is also automatically satisfied for large ones as well. This is guaranteed by a theorem which, in its various flavors and levels of generality, is usually called Stokes’s theorem or Green’s theorem. Stokes’s theorem looks scary when expressed in fancy mathematical notation, but expresses a fact that turns out to be visually obvious.

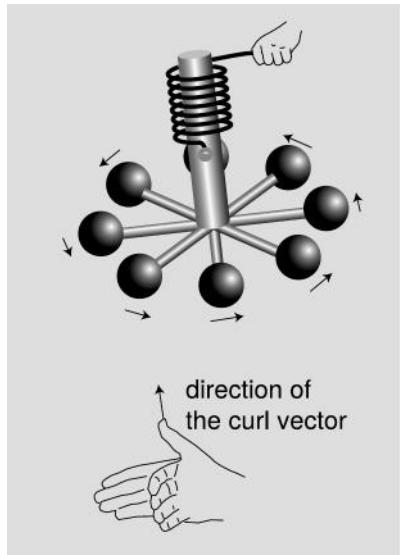
To make these ideas more concrete, we define an imaginary device called the curl-meter, shown in figure b. A set of positive charges is mounted around the circumference of a wheel, and a coil spring connects the wheel to its supporting axle, so that in the absence of any electrical forces, the wheel turns to a certain equilibrium orientation. If the curl-meter is immersed in a nonconservative electric field, then it may respond by rotating, and the torque that it measures is an indication of how badly nonconservative the field is. This is obviously not a practical, real-world device, but it is easy to make one in the real world out of an oscilloscope and a coil of wire. The basic idea of such devices is that they are like paddlewheels that will spin if inserted in a whirlpool.

The reading on the curl-meter, called the “curl” of the field, depends on its orientation, and this suggests that its reading is a vector, not a scalar like the divergence. Indeed, in the physical realization of figure b, the reading comes out as a torque, and torque is a vector. Therefore we have a right-hand rule for expressing the curl, which is the same as the right-hand rule for torques. It is possible to express the curl in terms of the partial derivatives of the field’s components, but detailed mathematical manipulations in that style are not necessary in order to follow the logical backbone of the presentation in this book, so we relegate this material to a note ([2101](#)). Figure c shows some visual examples.

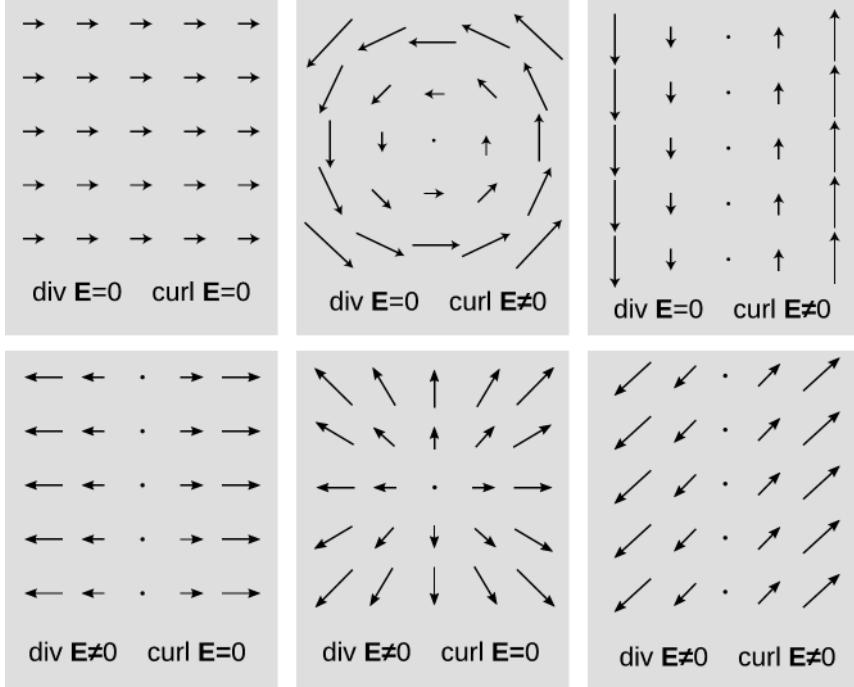
In turns of the curl, we have the following local law of physics:

For a static electric field, $\text{curl } \mathbf{E} = 0$.

This means that the work done around any *small* closed loop is zero. Let’s see how this generalizes to large loops. Stokes’s theorem



b / The curl-meter, and the right-hand rule for making its reading into a vector.

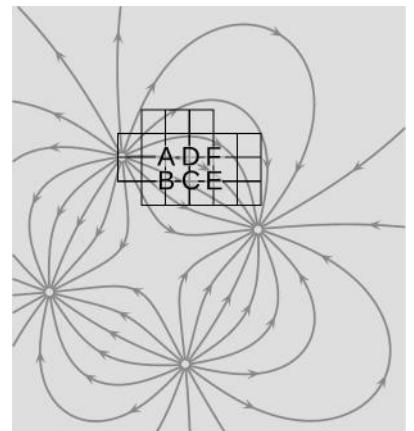


c / Some electric fields with zero and nonzero divergences and curls.

comes in a variety of versions and different levels of generality (e.g., there is a version that works in four dimensions, which is helpful in relativity, where time is treated as a dimension). For now we'll content ourselves with the special case that covers two dimensions, in the case where the curl is zero. (More general versions of Stokes's theorem tell us interesting things in the more general case where the curl can be nonzero.) For a field in a two-dimensional plane, the curl would point in the direction perpendicular to the plane, but we're now talking about the case where the curl is zero, so that issue doesn't arise.

In figure d, we construct the small square path ABCDA. This is small enough so that our local law $\text{curl } \mathbf{E} = 0$ guarantees, to good enough precision, that the work done on a charge around this path is zero. This approximation can be made as good as desired by making the squares small enough. The same thing works for the adjoining square DCEFD, using the same counterclockwise direction of motion. Now when we join the two squares together to make the rectangle ABEFA, something nice happens: the work done around the rectangle equals the *sum* of the works done around the two squares. The reason this is true is that in the first square we go up from C to D, while in the other we go down, so when we add the works for the two squares, these two contributions cancel. By the way, there is nothing special about squares — this holds for any shapes having an adjoining boundary.

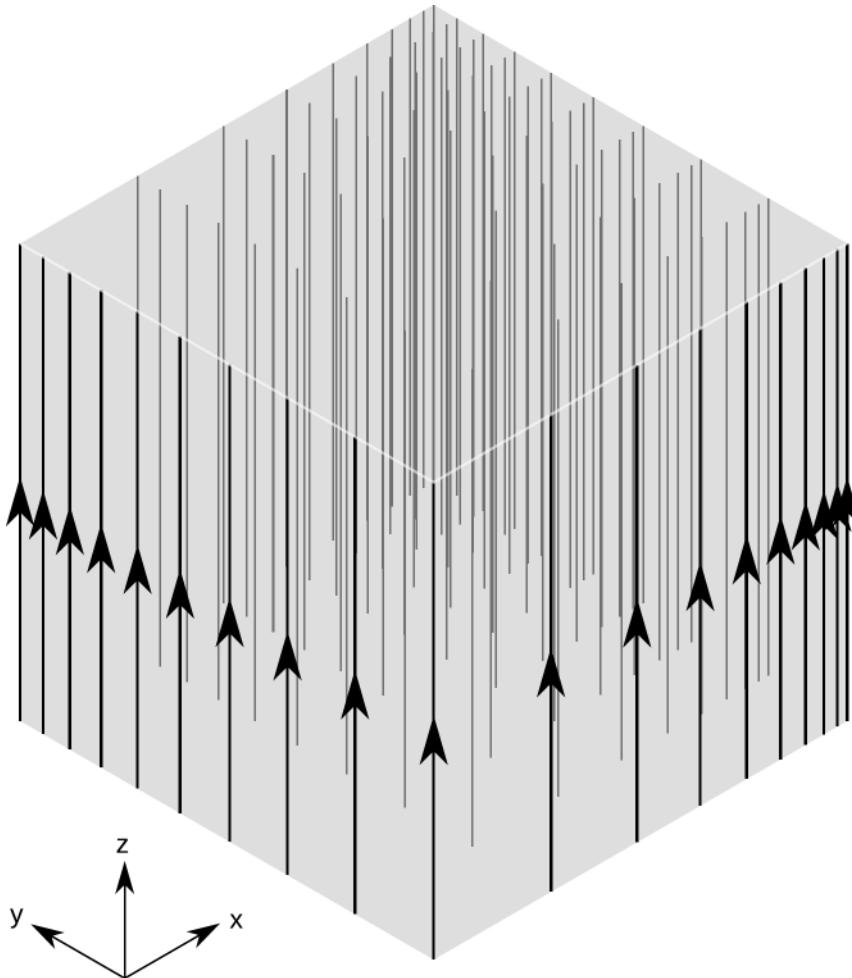
It is now clear that we can continue in this way and construct



d / A basic version of Stokes's theorem. The electric field in the background is just an example; any static field would work.

a grid that covers a piece of the plane of unlimited size, so that, for example, we know there is zero work done around the perimeter of the irregular shape shown in the figure. By making the grid fine enough, we can approximate any shape to the desired level of approximation, so the work done around any closed path is zero.

Although our main topic in this chapter is the electric field, it is also true that static magnetic fields in a vacuum obey the same equations: $\text{div } \mathbf{B} = 0$ and $\text{curl } \mathbf{B} = 0$. This is the underlying reason why, at large distances, electric dipole and magnetic dipole fields have the same universal form: the same equations have the same solutions. The magnetic dipole is discussed in more detail in sec. 11.4, p. 267.



e / An electric field pattern, discussion question A.

Discussion question

A Figure e shows a field pattern in three dimensions, represented using field lines. A coordinate system is defined for reference. Discuss the signs and relative sizes of the curl's x , y , and z components.

4.2 The electric potential

The earth's gravitational field has zero curl, and this is why it makes sense to define the gravitational potential energy of a test mass in the earth's field. If we lift the ball in figure f along path 1, then lower it by completing the loop along path 2, the total work done by the gravitational field is zero. This implies that if we lift along path 2, reversing the sign of the work, the amount of work done is the same as along path 1. That is, the work done by gravity is *path-independent*. We can therefore pick some reference point, such as the lower position, define the ball's gravitational potential energy to be zero there, and define the potential energy at any other point to be the work done by the hand to get from the reference point to that point. The definition is unambiguous because its result does not depend on the path taken.

In the context of the gravitational field, it is of interest to define the gravitational potential energy per unit mass; we don't need a special scientific name for this, because it is simply the height relative to the reference point.

Making the analogy with static electric fields, we define an electric potential ϕ , which is the energy per unit charge required in order to move a test charge to a certain point, from a reference point. It has units of joules per coulomb, which can be abbreviated as volts, $1 \text{ V}=1 \text{ J/C}$. In real-life electrical work, we may take the reference point to be a convenient object such as a water pipe. In theoretical contexts, it is often convenient to take the reference point to be infinitely far away. In informal contexts, people often refer to potential as voltage, and the notation V can also be used instead of ϕ .

Potential surrounding a point charge

example 1

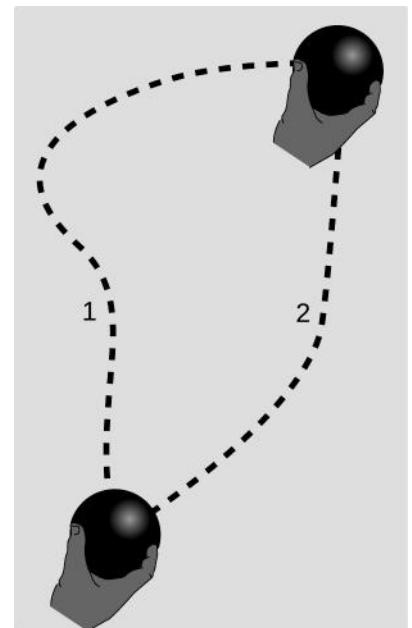
Given a point charge Q , the potential energy of a test charge q at a distance r is kQq/r . Dividing by the test charge, we find that the potential is

$$\phi = \frac{kQ}{r}.$$

A battery is a device that maintains a desired potential difference between its two terminals.

The instrument for measuring electric potentials is called a voltmeter (these days usually implemented as one of the functions of a multimeter). The voltmeter has no way of knowing what mental choice we had in mind as our reference potential, so all it can do is tell us the potential *difference* between two points,

$$\Delta\phi = \frac{\Delta U}{q}.$$



f / The work done by gravity along paths 1 and 2 is the same.



g / A multimeter set up to be used as a voltmeter, and a person using such a meter.

The meter has two plugs, which can be connected with wires to the two points of interest. The meter allows a tiny trickle of charge to flow in one wire, through the meter, and out the other, and it measures the work done by the electric field along this path. The potential difference $\Delta\phi$ is then minus the work divided by the charge.

Figure g shows a multimeter set up to be used as a voltmeter, and a person using such a meter to measure a potential difference. In the close-up, we see that one wire has been plugged into the COM plug (so called because it is used for all functions of the meter), while the other wire is plugged into the one marked V. The rotary dial is turned to V for a voltage measurement. (The position marked V with the sine wave icon would be used for measuring an alternating current, i.e., a voltage that was oscillating over time rather than static. The one marked mV would be used for measuring small voltages.) The digital readout reads the potential difference (V plug minus COM plug), in units of volts. Such meters are more or less standardized, so this description applies to all multimeters, with minor variations such as the possible use of an analog display with a needle, or pushbuttons rather than a rotary dial.

In the right-hand side of figure g, the woman using the voltmeter is touching the metal tips of the two probes to two different points, a piece of exposed wire in the rat's nest of wiring, and a point on the chassis. The reading on the meter will be the difference in electric potential between these two points.

Figure h/1 shows a map of part of Europe. Suppose you decide to walk all the way around the border of France, eventually returning to your starting point. The work done on you by the gravitational field is zero. In other words, any climbing you do on this hike will be canceled out by the elevation losses from going downhill. If we break the route up into pieces, then for example the segment from A and B will have a significant change in elevation, because B is near the crest of the Alps. Hiking from A to B will be an elevation gain, whereas if you went the opposite way and hiked from B to A, the elevation gain would be negative. Thus if you wanted to check that the total elevation gain on this loop was zero, you would need to decide on a clockwise or counterclockwise orientation for your hike, and you would also need to make sure that when you measured elevation differences, you defined their signs in a way that was consistent with this orientation.

In figure h/2, we make the map into a circuit containing a bunch of circuit elements.¹ There are various national borders we can trace in this circuit. There is one around France, containing six circuit elements, one around Switzerland that has three, and so on. There are also larger loops we could draw. For instance, we could start from the west coast of France and “hike” clockwise around the outer perimeter of the whole figure, which would take us through nine components. Regardless of which loop we choose, we expect that the sum of the voltage differences measured around any such closed loop will be zero.²

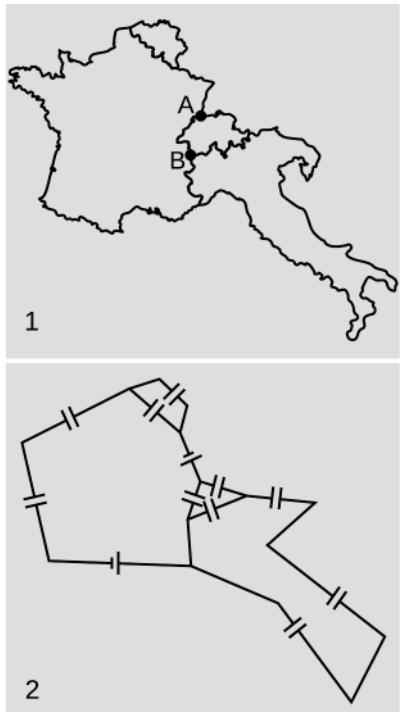
4.3 Constant potential throughout a conductor

In figure g, the woman is using a point on the metal chassis of the electrical box as a reference. It looks like she’s chosen a screw hole as a convenient place to make good electrical contact. This choice of a reference potential is arbitrary, and so we might imagine that although it would be equally valid to pick some other point on the chassis as a reference, the readings might then all be offset by some constant. In fact, we would find that the readings were all the same, because there will be no difference in potential between one point on the box and another.

The reason for this has to do with the fact that a metal contains a large number of electrons that are free to move. A substance like this that has many free charge carriers is called a *conductor*,

¹These symbols represent capacitors and batteries, but that doesn’t really matter very much — all that matters is that any electric fields in this circuit are static, and therefore not curly.

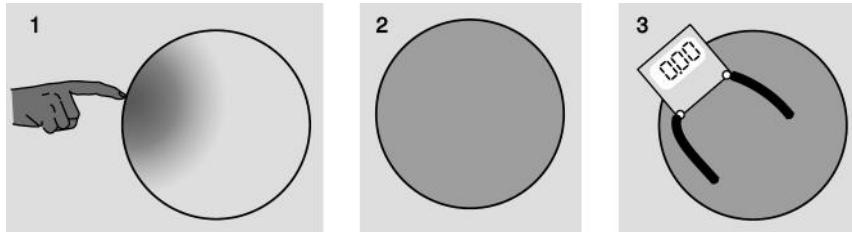
²If we want to check this in an actual experiment, then it turns out that if we check only the *inner* loops, i.e., the border of every country, then no more information is obtained by checking other loops such as the big outer loop. If we write down the four equations asserting that each inner loop adds up to zero, then it is possible to prove that every loop in the circuit adds up to zero.



h/1. Any hike around the complete boundary of a country results in zero net elevation gain, since we come back to our starting point at the same elevation.

the opposite being an insulator. The living cells in your body are pretty good conductors, because they have free charge carriers such as positively charged sodium ions and negatively charged chlorine ions, coming from the salt in your body. The outer layer of your skin is an insulator because it consists of dry, dead cells in which ions are not free to move because the material is a solid rather than a liquid. These distinctions between conductors and insulators are quantitative rather than absolute, as we will see in more detail in section 8.3, p. 195.

A good conductor cannot sustain a potential difference between different points within itself. If such a difference were set up, the charges would immediately begin to flow. Any positive charges would move toward the lower potential (as water, with positive mass, flows downhill), while negative charges would flow toward the higher potential. The system would rapidly reach an equilibrium in which the potential differences were eliminated. In the study of electrostatics, there are by definition no currents or time-varying fields, and therefore all charges must be in equilibrium, and this would be impossible if there were any differences in potential within a conductor.



i / 1. The finger deposits charges on the solid, spherical, metal doorknob and is then withdrawn. 2. Almost instantaneously, the charges' mutual repulsion makes them redistribute themselves uniformly on the surface of the sphere. The only excess charge is on the surface; charges do exist in the atoms that form the interior of the sphere, but they are balanced. Charges on the interior feel zero total electrical force from the ones at the surface. Charges at the surface experience a net outward repulsion, but this is canceled out by the force that keeps them from escaping into the air. 3. A voltmeter shows zero difference in voltage between any two points on the interior or surface of the sphere. If the voltage difference wasn't zero, then energy could be released by the flow of charge from one point to the other; this only happens before equilibrium is reached.

Excess charge placed on a conductor, once it reaches its equilibrium configuration, is entirely on the surface, not on the interior (proof [2101](#)). This should be intuitively reasonable in figure i, for example, since the charges are all repelling each other.

The lightning rod

example 2

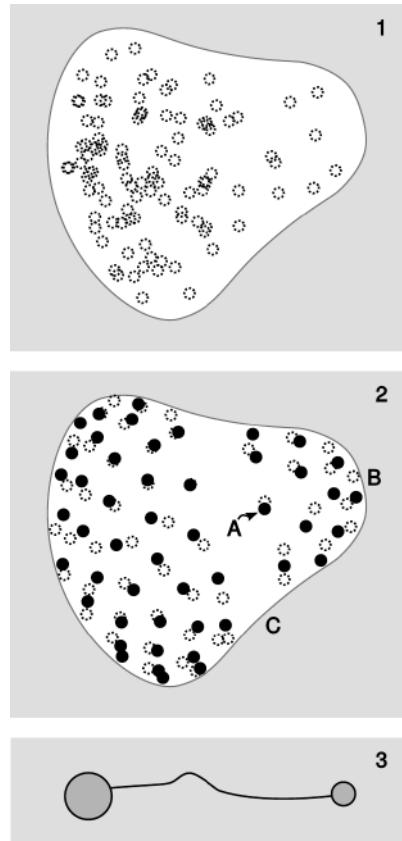
Suppose you have a pear-shaped conductor like the one in figure j/1. Since the pear is a conductor, there are free charges everywhere inside it. Panels 1 and 2 of the figure show a computer simulation with 100 identical electric charges. In 1, the charges are released at random positions inside the pear. Repulsion causes them all to fly outward onto the surface and then settle down into an orderly but nonuniform pattern.

We might not have been able to guess the pattern in advance, but we can verify that some of its features make sense. For example, charge A has more neighbors on the right than on the left, which would tend to make it accelerate off to the left. But when we look at the picture as a whole, it appears reasonable that this is prevented by the larger number of more distant charges on its left than on its right.

There also seems to be a pattern to the nonuniformity: the charges collect more densely in areas like B, where the surface is strongly curved, and less densely in flatter areas like C.

To understand the reason for this pattern, consider j/3. Two conducting spheres are connected by a conducting wire. Since the whole apparatus is conducting, it must all be at the same potential. As shown in problem 17 on p. 108, the density of charge is greater on the smaller sphere. This is an example of a more general fact observed in j/2, which is that the charge on a conductor packs itself more densely in areas that are more sharply curved.

Similar reasoning shows why Benjamin Franklin used a sharp tip when he invented the lightning rod. The charged stormclouds induce positive and negative charges to move to opposite ends of the rod. At the pointed upper end of the rod, the charge tends to concentrate at the point, and this charge attracts the lightning. The same effect can sometimes be seen when a scrap of aluminum foil is inadvertently put in a microwave oven. Modern experiments (Moore *et al.*, Journal of Applied Meteorology 39 (1999) 593) show that although a sharp tip is best at starting a spark, a more moderate curve, like the right-hand tip of the pear in this example, is better at successfully sustaining the spark for long enough to connect a discharge to the clouds.



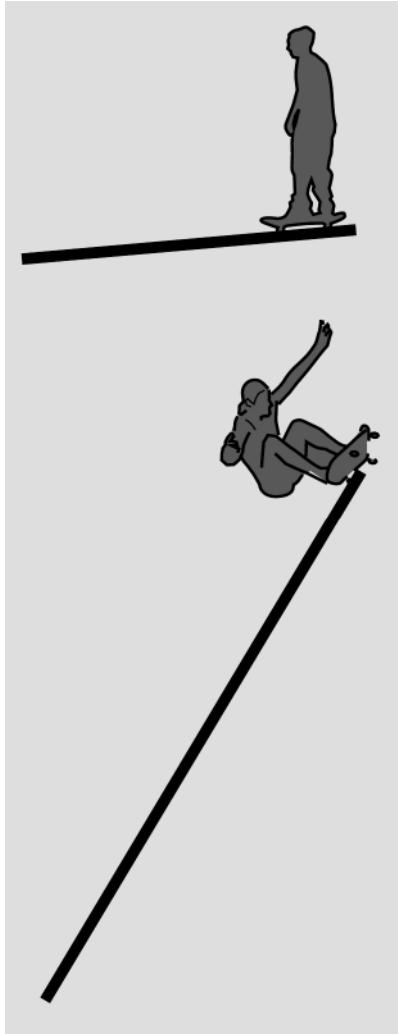
j / Example 2. In 1 and 2, charges that are visible on the front surface of the conductor are shown as solid dots; the others would have to be seen through the conductor, which we imagine is semi-transparent.

4.4 Potential related to field

4.4.1 One dimension

Work equals force times distance, $dW = F dx$, so when the electric field does work on a test charge, the energy lost by the field is $dU = -Eq dx$, and dividing by q and solving for E gives

$$E = -\frac{d\phi}{dx}.$$



k / The potential is analogous to the height, and the electric field to the slope.

The minus sign reflects the plumber's basic aphorism that shit flows downhill: if we release a positive charge at rest in an electric field, the field will accelerate the charge in the direction of lower ϕ — lower "height." In a different gravitational analogy, the two skateboarders in figure k will have very different net forces acting on them. The one on the steep positive slope will be accelerated much more rapidly in the negative direction.

Field generated by an electric eel

example 3

▷ Suppose an electric eel is 1 m long, and generates a voltage difference of 1000 volts between its head and tail. What is the electric field in the water around it?

▷ We are only calculating the amount of field, not its direction, so we ignore positive and negative signs. The field is probably not constant, but we don't have enough information to take that into account, so let's say it's constant. In general, a derivative $d\ldots/d\ldots$ can always be approximated by an expression of the form $\Delta\ldots/\Delta\ldots$ when the value of the derivative is constant, i.e., you don't need calculus to find a rate of change that is constant. Therefore we have

$$|E| \approx \frac{\Delta\phi}{\Delta x} \\ = 1000 \text{ V/m.}$$

Weighing an electron

example 4

J.J. Thomson (p. 76) is considered to have discovered the electron because he measured its charge-to-mass ratio q/m and found it to be much larger than that of an ionized atom, interpreting this as evidence that he was seeing a subatomic particle with a mass much smaller than an atom's. But not only is the electron's q/m relatively large compared to that of an atom, it is simply a huge number ($\sim -10^{11} \text{ C/kg}$) when expressed in SI units. SI units are designed for human scales of experience, so this suggests that in everyday life we should expect it to be very difficult to detect any effect from the weight or inertia of an electron.

As an example, suppose that a metal rod of length L is oriented upright. The conduction electrons are free to move, so they would tend to drop to the bottom of the rod. Electrical forces will however resist this segregation of positive and negative charges. To estimate how hard it would be to observe such an effect, let us imagine connecting the probes of a voltmeter to the ends of the rod. In equilibrium, the electrical and gravitational fields must have effects on an electron that cancel out. Setting the magnitudes of these forces equal to each other, we have $eE = mg$, and since (ignoring signs) $E = \Delta\phi/L$, we predict a potential difference $\Delta\phi = (m/q)gL$. For a one-meter rod, the predicted effect is $\sim 10^{-10} \text{ V}$.

This is quite small, but not impossible to measure, and the theoretical prediction was confirmed for a similar experiment by Tolman and Stewart in a 1916 experiment at Berkeley. This was the first direct evidence that the charge carriers inside a metal wire are in fact electrons. Similarly, we do expect mechanical side-effects in any electrical circuit, e.g., a slight twitching of a flashlight when we turn it on or off, but these will be much too small to notice except with exceptionally delicate and sensitive tools. It is surprising that we can get information about the microscopic structure of a metal merely by measuring its bulk electrical properties in this way.

In a vacuum in one dimension, Gauss's law is that $dE/dx = 0$, and since $E = -d\phi/dx$, we have

$$\frac{d^2 \phi}{dx^2} = 0. \quad [\text{vacuum, one dimension}]$$

This is the simplest form of various equations known as Laplace's equation (or Poisson's equation, in a generalized version where charges are present). The second derivative is a measure of curvature, so this equation states that the graph of the potential as a function of position is a straight line. Another nice ways of saying this is that the potential at a certain point is equal to the average of its values at nearby points on either side.

Figure 1 shows two examples (left and right columns) in which there are essentially the same charges present: a vertical strip of positive charges (black), and a strip of negative ones (white). This is effectively a one-dimensional problem. In both cases, Gauss's law is obeyed, because the field lines begin on positive charges and end on negative ones.

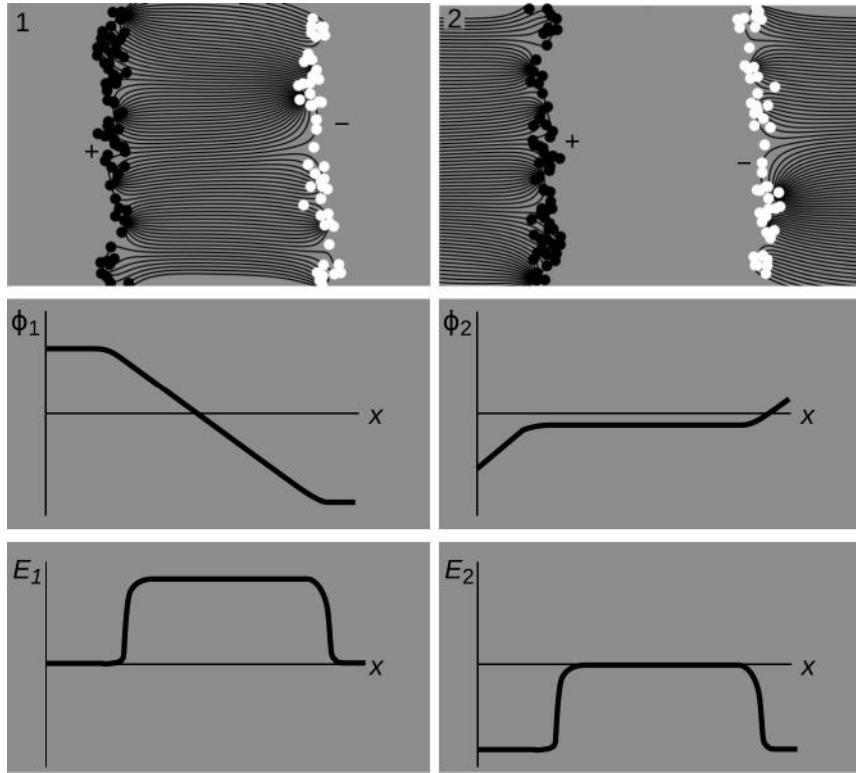
Even though the charges are the same, the fields are different. This shows that problems in electrostatics may not have unique solutions if we are only told about the distribution of charge and nothing else. As an even simpler example, we could have simply used a region of vacuum, with no charges at all, and any uniform field (including a zero field) would have been a solution.

In each example, there are three regions of vacuum, and in each of these we can observe that the function $\phi(x)$ is a straight line. In the regions where there are charges, $\phi(x)$ is a curve, and $d^2 \phi/dx^2 \neq 0$. In terms of the potentials, the difference between the two cases is that $\phi_2 - \phi_1$ is a linear function. The second derivative of a linear function is zero, so both are consistent with Laplace's equation.

In terms of the fields, $E_2 - E_1$ is a uniform field.

To get a unique solution to this type of electrostatics problem, we need to specify some *boundary conditions*. For example, if we required that the field be zero at large distances (as $x \rightarrow \pm\infty$), then we would have uniquely specified E_1 as a solution.

I / Two different fields for essentially the same charge distribution.



If you haven't already had a course in differential equations, you may not have heard of boundary conditions, but they are omnipresent in the physical sciences and engineering, and you have already been using them without knowing it. As a simple example, suppose that a hockey puck is gliding frictionless across some ice. There is a law of physics, Newton's second law, stating that $d^2 x/dt^2 = 0$. Any linear function would be a solution of this equation, since the puck can move at any constant speed, in either direction. However, if we specify the puck's position at two times, then the linear function is uniquely determined. The two given positions are two points on the graph of $x(t)$, and if we connect them with a line segment, they are the segments left and right boundaries.

Boundary conditions can have a variety of forms, and the word "boundary" isn't always literal. For example, we could have given the hockey puck's initial position and velocity rather than its initial and final positions.

In electrostatics, no physically verifiable boundary conditions can ever completely fix the potential, since only differences in potential are measurable. Another way of saying this is that it is the electric field that is measurable, and therefore the indefinite integral $\phi = - \int E dx$ will always have an arbitrary constant of integration. We might wish, for example, to set $\phi = 0$ at some chosen point, and there is nothing wrong with this boundary condition, but neither does it correspond to anything physically observable.

4.4.2 Two or three dimensions

The topographical map in figure m suggests a good way to visualize the relationship between field and potential in two dimensions. Each contour on the map is a line of constant height; some of these are labeled with their elevations in units of feet. Height is related to gravitational energy, so in a gravitational analogy, we can think of height as representing potential. Where the contour lines are far apart, as in the town, the slope is gentle. Lines close together indicate a steep slope.

If we walk along a straight line, say straight east from the town, then height (potential) is a function of the east-west coordinate x . The slope along such a line is $d\phi/dx$ (the rise over the run), and the electric field is minus this derivative.

What if everything isn't confined to a straight line? Water flows downhill. Notice how the streams on the map cut perpendicularly through the lines of constant height.

It is possible to map potentials in the same way, as shown in figure n. Each curve is a line of constant potential, called an equipotential. In the full three-dimensional representation, the equipotentials would be surfaces. Moving along an equipotential is like walking along a hillside without moving up or down. The electric field is strongest where the equipotentials are closest together, and the electric field vectors always point perpendicular to the equipotentials. Figure o shows a representation in this style for the field of a dipole, overlaid with the field lines.

The one-dimensional relationship $E = -d\phi/dx$ generalizes to three dimensions as follows:

$$E_x = -\frac{\partial \phi}{\partial x}$$

$$E_y = -\frac{\partial \phi}{\partial y}$$

$$E_z = -\frac{\partial \phi}{\partial z}$$

This can be notated using the symbol ∇ , called the gradient operator,

$$\mathbf{E} = -\nabla\phi.$$

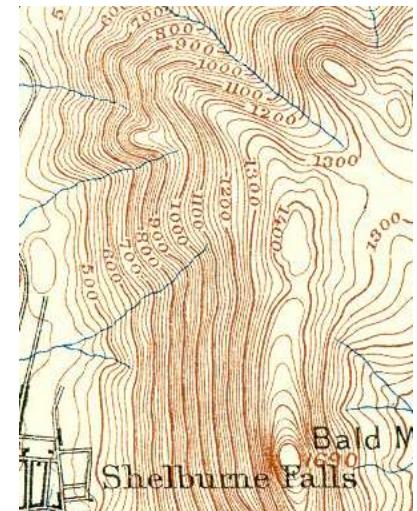
A uniform field

example 5

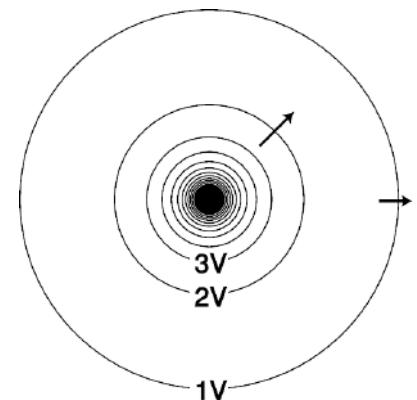
- ▷ Find the electric field corresponding to the potential $\phi = ax + by$, where a and b are constants.
- ▷ The gradient of ϕ has x and y components a and b , so

$$E_x = -a \quad \text{and}$$

$$E_y = -b.$$

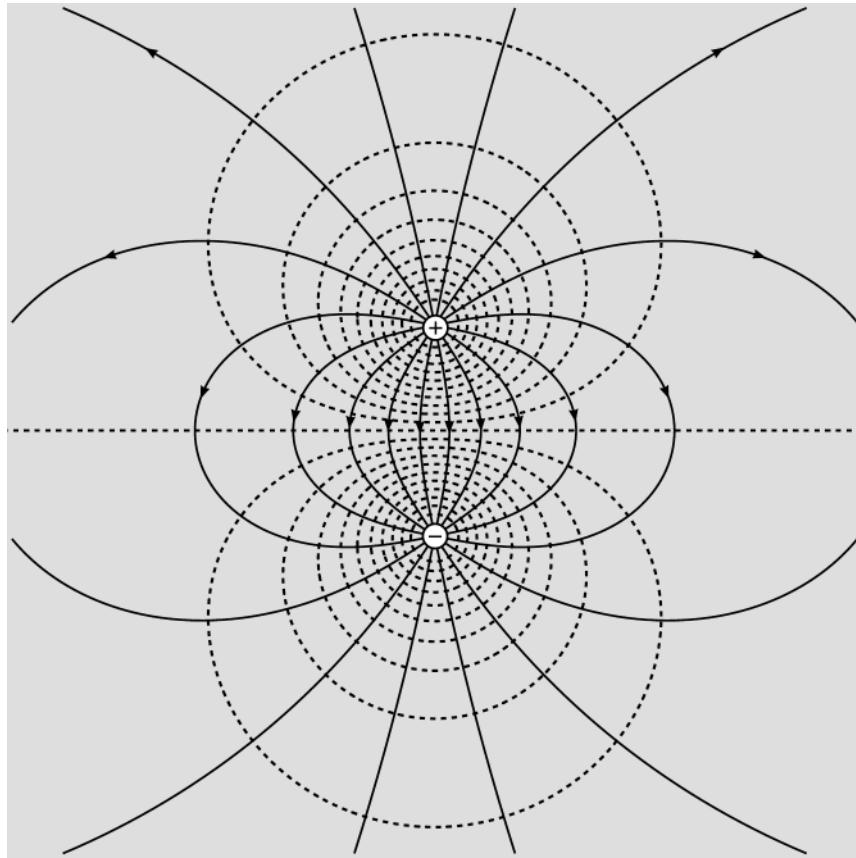


m / A topographical map of Shelburne Falls, Massachusetts.



n / The equipotentials surrounding a point charge. Near the charge, the curves are so closely spaced that they blend together on this drawing due to the finite width with which they were drawn. Some electric fields are shown as arrows.

o / The field lines (solid) and equipotentials (dashed lines) of a dipole.



4.5 Summary of div, grad, and curl

At this point in the text, we have encountered all three of the important derivative operators of vector calculus, described in visual terms. We also know how to calculate two of them explicitly for a given field. This book is designed so that you can follow all the ideas and do all the problems while concurrently taking a one-semester vector calculus course. The details are mostly delayed until the end of this book, so that you shouldn't be overwhelmed by the math even if, as often happens, your vector calculus course doesn't get to the good stuff until the last few weeks. Here is a summary of how the derivative operators differ from each other:

<i>operator</i>	<i>input</i>	<i>output</i>	<i>measures</i>
div	vector field	scalar	how much the field spreads out from a point
curl	vector field	vector	how much the field rotates
gradient, ∇	scalar field	vector	how fast the field changes, and the direction of steepest change

Because they're all derivatives, they also all have some properties in common:

- When you apply them to a constant, they give zero.

- They have the additive or “linear” property that the derivative of a sum is the sum of the derivatives.

These three operators are the only possible spatial derivative operators that are rotationally invariant (sec. 1.3.5, p. 24), meaning that if you rotate your coordinate system, the output will be the same result, just reexpressed in the new coordinate system. Because coordinate systems are just a human choice, any physically meaningful derivative operator has to have this property, and therefore the fundamental laws of electricity and magnetism can only be expressed in terms of these three operators. These laws of physics are called Maxwell’s equations, and are summarized on p. 444. When we restrict to the case of electrostatics, Maxwell’s equations are $\text{div } \mathbf{E} = 4\pi k\rho$ and $\text{curl } \mathbf{E} = 0$.

By the end of this book, we will have fleshed out our picture of vector calculus a little more, and this more complete picture is summarized in sec. 15.1, p. 347.

4.6 Boundary conditions on a conductor

4.6.1 No component of the electric field parallel to the surface

If a charge is located at the surface of a conductor, then moving it by an infinitesimal distance in any direction parallel to the surface leaves it at the same potential, so that no work is done by the electric field. Therefore the electric field never has any component parallel to the surface of a conductor.

A common type of problem in electrostatics is that one is given boundary conditions in which the potential is specified at certain conductors.

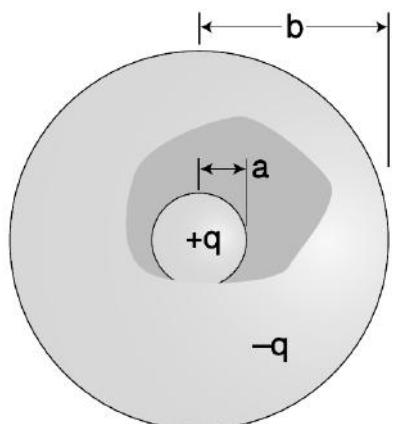
A spherical capacitor

example 6

▷ A spherical capacitor, figure u, consists of concentric, conducting spheres with radii a and b . Find the electric field for $a \leq r \leq b$ when the potential difference between the spheres $\phi_b - \phi_a$ equals $\Delta\phi$.

▷ By symmetry, we expect the field to be purely in the radial direction, and to have constant strength on any spherical surface concentric with the capacitor. If we take such a sphere as a Gaussian surface, then Gauss’s law tells us that the field will be the same as that of a point charge q at the origin, $E = kq/r^2$, where q is the charge on the inner sphere.

We would be done now, except that the problem is not stated in terms of q but in terms of the potential difference, which is what we would actually measure and control in a real-world capacitor.



p / Example 6. Part of the outside sphere has been drawn as if it is transparent, in order to show the inside sphere.

Integrating along a radius, we have

$$\begin{aligned}\Delta\phi &= - \int_a^b E \, dr \\ &= kq \left(\frac{1}{b} - \frac{1}{a} \right).\end{aligned}$$

Eliminating q gives

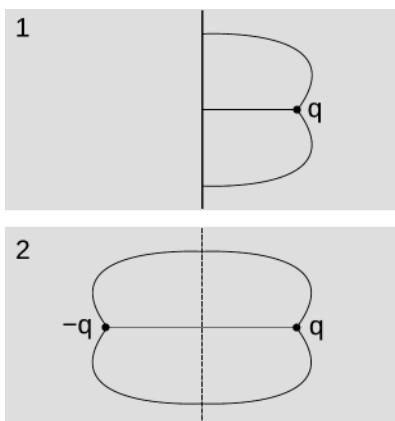
$$E = -\frac{\Delta\phi}{r^2(a^{-1} - b^{-1})},$$

where the minus sign indicates that if $\phi_b > \phi_a$, the field is inward. At $r = a$ and b , this field is perpendicular to both conducting surfaces.

4.6.2 The method of images

A car's radio antenna is usually in the form of a whip sticking up above its metal roof. This is an example involving radio waves, which are time-varying electric and magnetic fields, but a similar, simpler electrostatic example is the following. Suppose that we position a charge $q > 0$ at a distance ℓ from a conducting plane. What is the resulting electric field? The conductor has charges that are free to move, and due to the field of the charge q , we will end up with a net concentration of negative charge in the part of the plane near q . The field in the vacuum surrounding q will be a sum of fields due to q and fields due to these charges in the conducting plane. The problem can be stated as that of finding a solution to Poisson's equation with the boundary condition that $V = 0$ at the conducting plane. Figure q/1 shows the kind of field lines we expect.

This looks like a very complicated problem, but there is trick that allows us to find a simple solution. We can convert the problem into an equivalent one in which the conductor isn't present, but a fictitious *image* charge $-q$ is placed at an equal distance behind the plane, like a reflection in a mirror, as in figure q/2. The field is then simply the sum of the fields of the charges q and $-q$, so we can either add the field vectors or add the potentials. By symmetry, the field lines are perpendicular to the plane, so the plane is an surface of constant potential, as required.



q / The method of images.

Notes for chapter 4

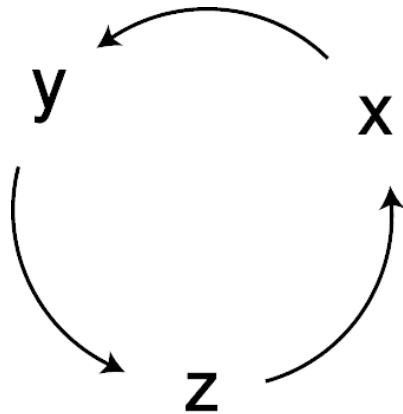
286 The curl operator in terms of the field's components

Fixing a Cartesian coordinate system, we find how to express the curl's components in terms of the derivatives of the field's components with respect to the coordinates.

Because we want the curl to be a kind of first derivative operator, we expect it to depend only on the derivatives of the field's components with respect to the coordinates. The field has three components, and there are three coordinates that we can differentiate with respect to, so there are nine possible partial derivatives that could consider. Because the curl is intended to be a kind of derivative, and derivatives are additive, we can isolate these partial derivatives from each other and add them. E.g., it's conceivable (although not true, as we'll see shortly) that the y component of the curl is $3\partial E_x/\partial x - 7\partial E_y/\partial z$, but we don't have to consider possibilities such as $\sin(\partial E_y/\partial x)$.

The partial derivative $\partial E_x/\partial x$ contributes to the divergence, but it can't contribute to the curl because of symmetry. For example, in the field $x\hat{x}$, this is the only partial derivative that would be nonzero, but by symmetry the curl-meter won't rotate when placed in this field. By rotational invariance, we can immediately conclude that there is no contribution from the other "self" terms, $\partial E_y/\partial y$ and $\partial E_z/\partial z$.

Visualizing the field $y\hat{z}$ with a curl-meter and applying the right-hand rule, we can tell that its curl must have a positive x component. Up until now we have never specified the units of the curl-meter, but now we need to decide. We define the curl so that the result in this case is +1, so that $\partial E_z/\partial y$ occurs in the expression for the x component of the curl. Similar reasoning produces the term $-\partial E_y/\partial z$.



A cyclic permutation of x , y , and z .

The remainder of the result follows from rotational invariance. It is possible to take the x , y , and z axes and rotate them rigidly in the manner shown in the figure, called a cyclic permutation. Therefore if a derivative like $\partial E_z/\partial y$ occurs in the x component of the curl, then we must have the others obtained from it by cyclic permutation, such as $\partial E_x/\partial z$ in the y component. The result is:

$$(\text{curl } \mathbf{E})_x = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}$$

$$(\text{curl } \mathbf{E})_y = \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}$$

$$(\text{curl } \mathbf{E})_z = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}.$$

292 No charge on the interior of a conductor

For a conductor in equilibrium, any charge is on its surface, never in its interior.

The proof of this assertion is essentially Earnshaw's theorem (example 11, p. 62). Suppose that a particle with charge q is in the conductor. For concreteness, let's say q is positive. By assumption, this charge is in a stable equilibrium. If the charge is at the surface, then this equilibrium can be created by both (1) the electric field, and (2) the force that keeps charges from getting out through the surface. But if the charge is in the interior, then only electrical forces can be involved, not other forces of type 2; this is essentially what

we mean by saying that the substance is a conductor. The electric field acting on q would be the field contributed by all the other charges, not by q itself (265). But now everything plays out as in the original argument proving Earnshaw's theorem: a stable equilibrium would require $\operatorname{div} \mathbf{E} < 0$, Gauss's law forbids this negative divergence from being created by external sources.

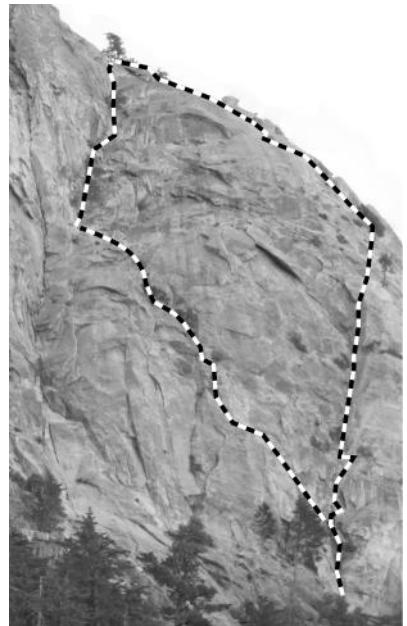
Problems

Key

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

1 The figure shows two climbing routes at Tahquitz Rock, near Idyllwild, California. They begin and end at the same point. Compare the work done against gravity by the same climber climbing the two routes. Why would it be of interest to state these figures instead as the work per kilogram of body mass, in units of J/kg? In the analogy with electrical fields, what would these units be?

▷ Solution, p. 427



Problem 1.

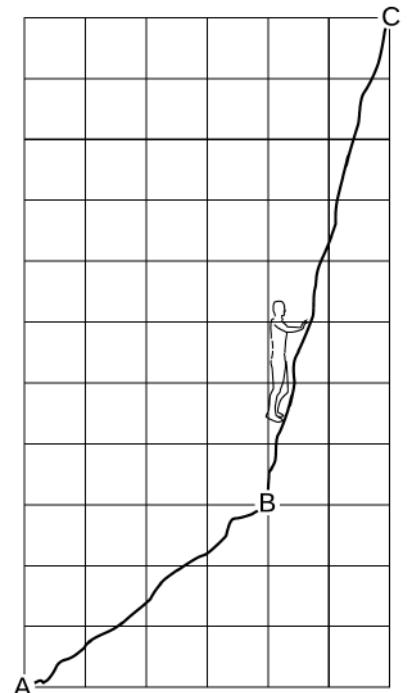
2 The figure shows a rock climbing route lying in a vertical plane. It can be approximated by two line segments. Superimposed on the climb is a grid of squares with sides of length ℓ . The gravitational field is g .

(a) Continuing the gravitational analogy from problem 1, find $\phi_B - \phi_A$ and $\phi_C - \phi_B$, where the “electric potential” ϕ is the gravitational potential energy per unit mass. ✓

(b) In the electrical version of this situation, the “height” is not a physical distance in space at all, so we could say that only the horizontal segments of the squares represent distances, and the situation is effectively one-dimensional. Find the ratio of the “electric fields” E_{BC}/E_{AB} . ✓

(c) In electromagnetism, we can always add an arbitrary constant to the potential while still describing the same physical situation. What would be the analogous statement for the climber in our gravitational analogy?

3 A hydrogen atom is electrically neutral, so at large distances, we expect that it will create essentially zero electric field. This is not true, however, near the atom or inside it. Very close to the proton, for example, the field is very strong. To see this, think of the electron as a spherically symmetric cloud that surrounds the proton, getting thinner and thinner as we get farther away from the proton. (Quantum mechanics tells us that this is a more correct picture than trying to imagine the electron orbiting the proton.) Near the center of the atom, the electron cloud’s field cancels out by symmetry, but the proton’s field is strong, so the total field is very strong. The potential in and around the hydrogen atom can be approximated using an expression of the form $\phi = r^{-1}e^{-r}$. (The units come out wrong, because I’ve left out some constants.) Find the electric field corresponding to this potential, and comment on its behavior at very large and very small r . ▷ Solution, p. 427



Problem 2.

4 Consider the following four potentials, which exist in some region of the positive x axis:

$$\phi_1 = ax$$

$$\phi_2 = ax + b$$

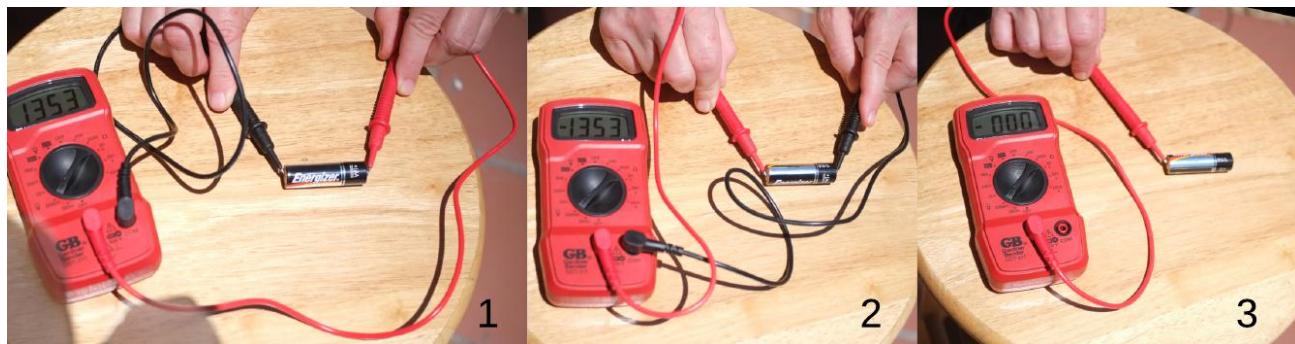
$$\phi_3 = \frac{a}{x}$$

$$\phi_4 = \frac{a}{x} + b.$$

In each case, find the corresponding electric field, and give physical interpretations of what is going on physically and of the constants a and b .

5 The figure shows a voltmeter being connected to a battery in three different ways. In case some of the details are too hard to see, the readings are 1353, -1353, and 0, the rotary dial is on a 2000 mV DC scale, and the banana plug connector on the right is the one labeled COM. Explain why these results are obtained.

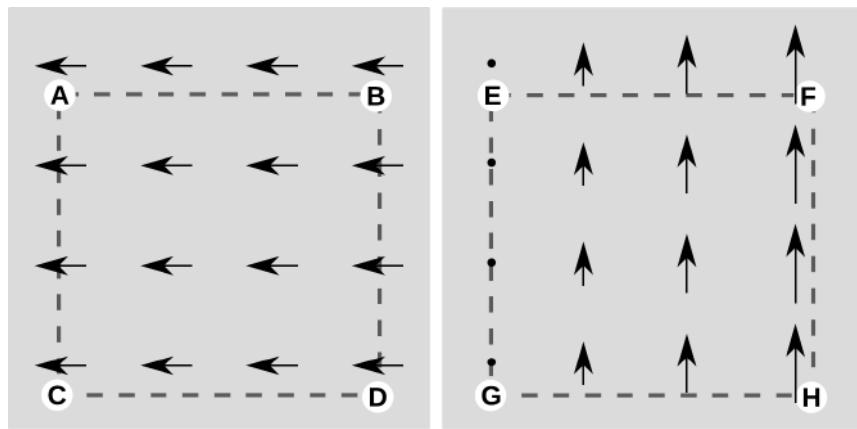
▷ Solution, p. 427



Problem 5.

6 The figure shows two different field patterns.

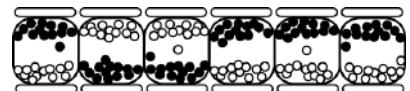
- (a) In the first field pattern, suppose that moving a charge “uphill” along path ABD requires that work be done against the electric field in the amount of one joule per coulomb. How much work would have to be done per unit charge along path ACD?
- (b) In the second figure, the work per unit charge along EFH is 2 J/C. What would it be along EGH?
- (c) Can an electric potential ϕ be defined for field pattern a? For b?



Problem 6.

7 The voltage difference between the ends of a AAA battery is 1.5 V, and the battery’s length is 42 mm. If the electric field inside is constant, find its magnitude.

8 Electronic ink is a technology, used in electronic book readers such as the Amazon Kindle, for displaying images. The figure shows a side view of a cross-section of a small part of the screen, cutting through one row of pixels. Each pixel consists of a tiny capacitor, about 0.1 mm in height, with a capsule inside it. The capsule contains black and white particles of pigments which have opposite charges. When a voltage is applied to the capacitor, the particles sort themselves out in opposite directions. When the white particles are on top, the pixel appears like white paper when viewed from above. When the black particles are on top, it appears like a dot of black ink. An electric field of about 1.5×10^5 V/m is needed in order to make the particles move. Estimate the voltage that has to be applied to the capacitor.



Problem 8.



A spark plug, problem 9.



Problem 12.

9 In example 4 on p. 52, we discussed the spark plugs of a gasoline engine, which need to make an electric field of a certain strength in order to spark. What we didn't discuss then was the very small size of the spark gap, seen in the close-up photo. All other things being equal, the small size of the gap would seem extremely undesirable. Special tools are required in order to measure it, and if it gets crud on the tips, the gap can easily become clogged. Now that you understand the relationship between the field and the potential, can you explain why we would make the gap so small?

▷ Solution, p. 427

10 In example 3, p. 51, we showed that the distant electric field of an electric dipole, in its mid-plane, was proportional to r^{-3} , and in problem 9, p. 69, you showed that this is also true on the dipole's axis. It is in fact true for points along any line passing through an electric dipole that its distant field is $E = br^{-3}$, with the constant factor b differing by a unitless factor depending on the orientation of the line. Find the corresponding electric potential ϕ .

11 A carbon dioxide molecule is structured like O-C-O, with all three atoms along a line. The oxygen atoms grab a little bit of extra negative charge, leaving the carbon positive. The molecule's symmetry, however, means that it has no overall dipole moment, unlike a V-shaped water molecule, for instance. Whereas the potential of a dipole of magnitude D is proportional to D/r^2 , it turns out that the potential of a carbon dioxide molecule at a distant point along the molecule's axis equals b/r^3 , where r is the distance from the molecule and b is a constant. What would be the electric field of a carbon dioxide molecule at a point on the molecule's axis, at a distance r from the molecule?

✓

12 A vacuum tube, in its simplest conceptual form, is a parallel-plate capacitor enclosed in a glass tube so that all the air can be pumped out. A potential difference is applied, and if the negative plate is heated, some electrons will spontaneously pop out of the metal and be accelerated across the gap. The density of the electrons in flight is nonuniform because of their acceleration, and an analysis by Richardson around 1901 showed that the potential would have the form $\phi = cx^{4/3}$, where c is a constant. Find the electric field.

✓

13 The neuron in the figure has been drawn fairly short, but some neurons in your spinal cord have tails (axons) up to a meter long. The inner and outer surfaces of the membrane act as the “plates” of a capacitor. (The fact that it has been rolled up into a cylinder has very little effect.) In order to function, the neuron must create a voltage difference V between the inner and outer surfaces of the membrane. Let the membrane’s thickness, radius, and length be t , r , and L . (a) Calculate the energy that must be stored in the electric field for the neuron to do its job. (In real life, the membrane is made out of a substance called a dielectric, whose electrical properties increase the amount of energy that must be stored. For the sake of this analysis, ignore this fact.) \triangleright Hint, p. 425 ✓

(b) An organism’s evolutionary fitness should be better if it needs less energy to operate its nervous system. Based on your answer to part a, what would you expect evolution to do to the dimensions t and r ? What other constraints would keep these evolutionary trends from going too far?

14 In example 8, p. 58, we found that the field of a uniform, straight line of charge is $E = 2k\lambda/r$, where r is the distance from the line. Find the corresponding electric potential. ✓

15 (a) An electric potential is given by $\phi = ar^2$, where r is the distance from the origin and a is a constant. This can be written as

$$\phi = a(x^2 + y^2 + z^2).$$

Find the components of the corresponding electric field by computing the gradient.

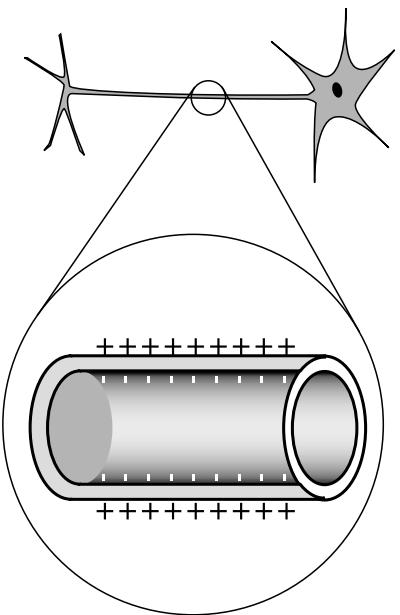
(b) Find the magnitude of the field. ✓

(c) By comparing with the result of example 7, p. 55, show that this is the potential of a uniform sphere of charge, and determine a in terms of the charge density. ✓

(d) Suppose that the potential had instead been

$$\phi = a(x^2 + y^2 + z^2) + b,$$

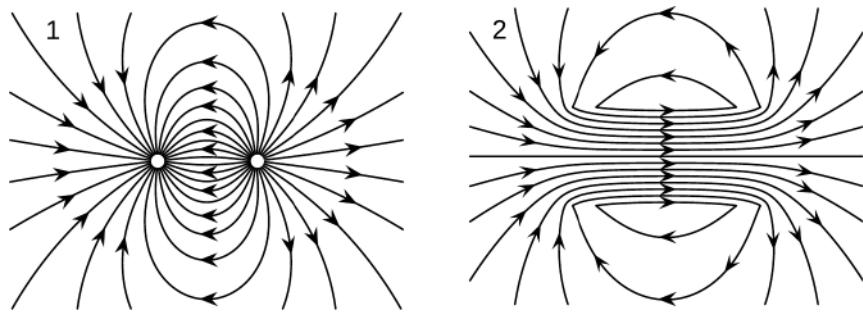
where b is a constant. How would this have affected the results?



Problem 13.

16 The figure shows field patterns 1 and 2. They are similar in some ways, especially at larger distances from the center.

- (a) What can you say about the divergence and curl of these patterns?
- (b) Prove that one of these cannot be a magnetic field pattern.
- (c) Prove that one of them cannot be a static electric field pattern.
- (d) Prove that one of them cannot be expressed as the gradient of a potential.



Problem 16.

17 In example 2 on p. 93, suppose that the larger sphere has radius a , the smaller one b . (a) Show that the ratio of the charges on the two spheres is $q_a/q_b = a/b$. (b) Show that the density of charge (charge per unit area) is the other way around: the charge density on the smaller sphere is *greater* than that on the larger sphere in the ratio a/b .

18 Three charges, each of strength Q ($Q > 0$) form a fixed equilateral triangle with sides of length b . You throw a particle of mass m and positive charge q from far away, with an initial speed v . Your goal is to get the particle to go to the center of the triangle, your aim is perfect, and you are free to throw from any direction you like. What is the minimum possible value of v ?

✓

19 The figure shows a simplified diagram of an electron gun such as the one that creates the electron beam in a TV tube. Electrons that spontaneously emerge from the negative electrode (cathode) are then accelerated to the positive electrode, which has a hole in it. (Once they emerge through the hole, they will slow down. However, if the two electrodes are fairly close together, this slowing down is a small effect, because the attractive and repulsive forces experienced by the electron tend to cancel.)

(a) If the voltage difference between the electrodes is ΔV , what is the velocity of an electron as it emerges at B? Assume that its initial velocity, at A, is negligible. \checkmark

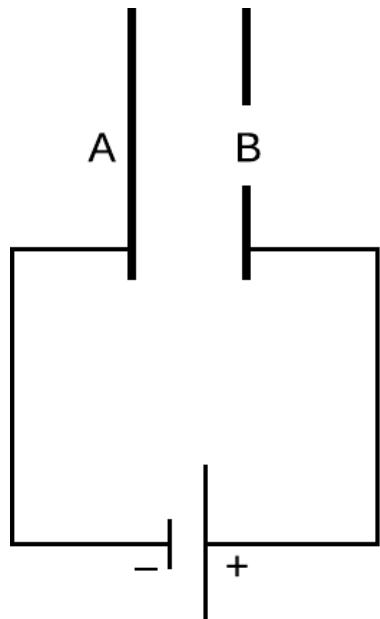
(b) Evaluate your expression numerically for the case where $\Delta V=10$ kV, and compare to the speed of light. \triangleright Solution, p. 427 \checkmark

20 The figure shows a simplified diagram of a device called a tandem accelerator, used for accelerating beams of ions up to speeds on the order of 1-10% of the speed of light. (Since these velocities are not too big compared to c , you can use nonrelativistic physics throughout this problem.) The nuclei of these ions collide with the nuclei of atoms in a target, producing nuclear reactions for experiments studying the structure of nuclei. The outer shell of the accelerator is a conductor at zero voltage (i.e., the same voltage as the Earth). The electrode at the center, known as the “terminal,” is at a high positive voltage, perhaps millions of volts. Negative ions with a charge of -1 unit (i.e., atoms with one extra electron) are produced offstage on the right, typically by chemical reactions with cesium, which is a chemical element that has a strong tendency to give away electrons. Relatively weak electric and magnetic forces are used to transport these -1 ions into the accelerator, where they are attracted to the terminal. Although the center of the terminal has a hole in it to let the ions pass through, there is a very thin carbon foil there that they must physically penetrate. Passing through the foil strips off some number of electrons, changing the atom into a positive ion, with a charge of $+n$ times the fundamental charge. Now that the atom is positive, it is repelled by the terminal, and accelerates some more on its way out of the accelerator.

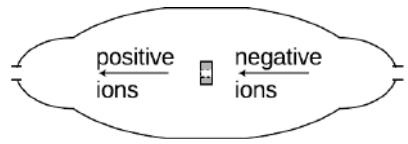
(a) Find the velocity, v , of the emerging beam of positive ions, in terms of n , their mass m , the terminal voltage V , and fundamental constants. Neglect the small change in mass caused by the loss of electrons in the stripper foil. \checkmark

(b) To fuse protons with protons, a minimum beam velocity of about 11% of the speed of light is required. What terminal voltage would be needed in this case? \checkmark

(c) In the setup described in part b, we need a target containing atoms whose nuclei are single protons, i.e., a target made of hydro-



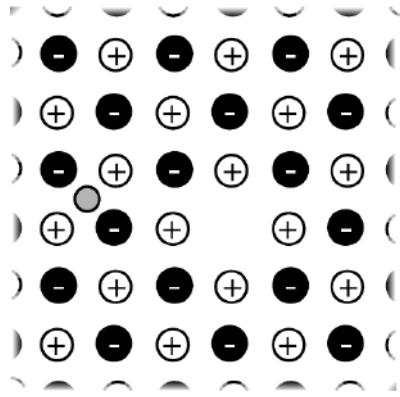
Problem 19.



Problem 20.

gen. Since hydrogen is a gas, and we want a foil for our target, we have to use a hydrogen compound, such as a plastic. Discuss what effect this would have on the experiment.

- 21** (a) A coaxial cable consists of a cylindrical inner conductor with radius a and an outer one with radius b . If the potential difference between the two conductors is $\Delta\phi$, find the electric field at $a \leq r \leq b$. ✓
(b) Show that your answer makes sense when $a = b$ and when $b < a$.



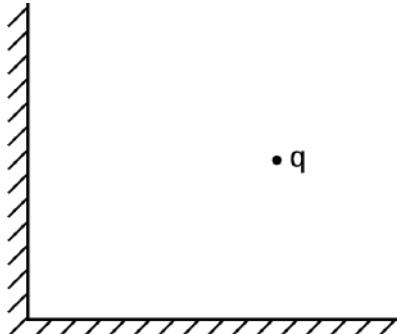
Problem 22.

- 22** Referring back to problem 7, p. 83, about the sodium chloride crystal, suppose the lithium ion is going to jump from the gap it is occupying to one of the four closest neighboring gaps. Which one will it jump to, and if it starts from rest, how fast will it be going by the time it gets there? (It will keep on moving and accelerating after that, but that does not concern us.)

▷ Hint, p. 425 ✓ ★

- 23** A charged particle of mass m and charge q is below a horizontal conducting plane. We wish to find the distance ℓ between the particle and the plane so that the particle will be in equilibrium, with its weight supported by electrostatic forces.

- (a) Determine as much as possible about the form of the answer based on units.
(b) Find the full result for ℓ .
(c) Show that the equilibrium is unstable.



Problem 24.

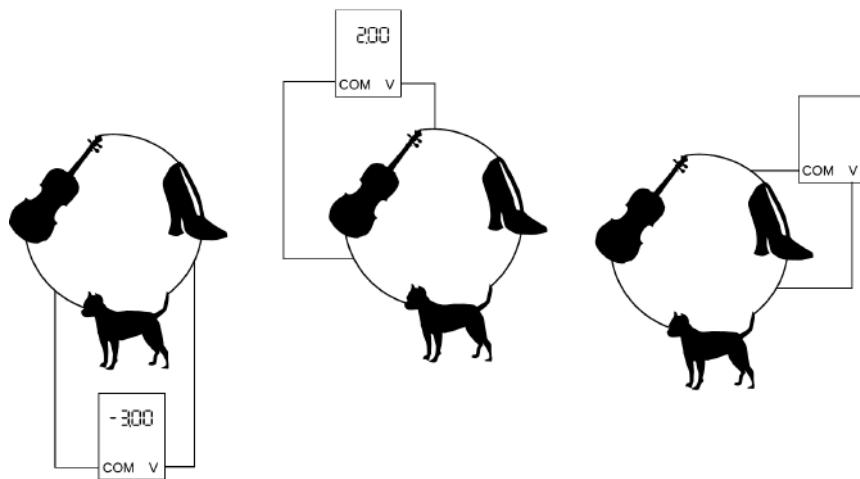
- 24** A point charge q is situated in the empty space inside a corner formed by two perpendicular half-planes made of sheets of metal. Let the sheets lie in the y - z and x - z planes, so that the charge's distances from the planes are x and y . Both x and y are positive. The charge will accelerate due to the electrostatic forces exerted by the sheets. We wish to find the direction θ in which it will accelerate, expressed as an angle counterclockwise from the negative x axis, so that $0 < \theta < \pi/2$.

- (a) Determine as much as possible about the form of the answer based on units.
(b) Find the full result for θ .

★

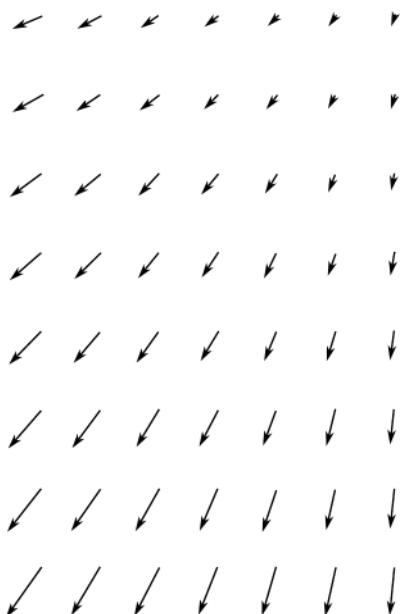
- 25** The figure shows a circuit that contains only static electric fields. Given the first two readings, we want to predict the third one. Is it $+1$, -1 , $+5$, or -5 V? Explain your answer.

Any of the four given possibilities can be obtained by adding the voltages, depending on the signs. It matters how the voltmeter is connected across each component (which side is COM and which is V), and it also matters that one meter reading is positive and the other negative. Without specifically discussing those details, there is no way to determine which answer is correct.



Problem 25.

- 26** Describe the curl and divergence of the field shown in the figure.



Problem 26.

Minilab 4A: Mapping electric fields

Apparatus

board and U-shaped probe

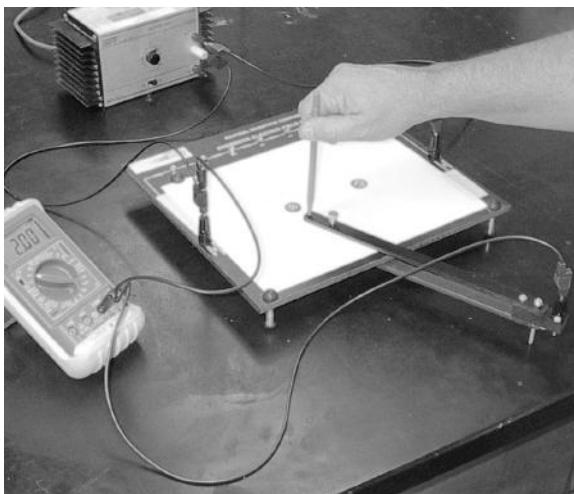
DC power supply

multimeter

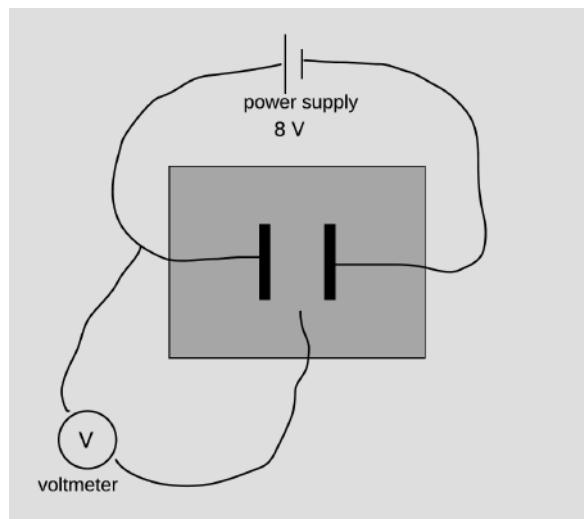
stencils for drawing electrode shapes on paper

Goal: Visualize electric fields using equipotential curves.

Test Maxwell's equations in the special case of electrostatics.



b / Photo of the apparatus, showing a different electrode pattern (two point charges).



a / Circuit diagram.

Figure a shows a circuit diagram of the apparatus. The power supply provides an 8 V potential difference between the two metal electrodes, drawn in black. A voltmeter measures the potential difference between an arbitrary reference voltage and a point of interest in the gray area around the electrodes. The result will be somewhere between 0 and 8 V.

The photo in figure b shows the actual apparatus. The electrodes are painted with silver paint on a detachable board, which goes underneath the big board. What you actually

see on top is just a piece of paper on which you'll trace the equipotentials with a pen. The voltmeter is connected to a U-shaped probe with a metal contact that slides underneath the board, and a hole in the top piece for your pen.

Turn your large board upside down. Find the small detachable board with the parallel-plate capacitor pattern on it, and screw it to the underside of the equipotential board, with the silver-painted side facing down toward the tabletop. Use the washers to protect the silver paint so that it doesn't get scraped off when you tighten the screws. Now connect the voltage source (using the provided wires) to the two large screws on either side of the board. Connect the multimeter so that you can measure the voltage difference across the terminals of the voltage source. Adjust the voltage source to give 8 volts.

If you press down on the board, you can slip the paper between the board and the four buttons you see at the corners of the board. Tape the paper to your board, because the buttons aren't very dependable. There are plastic stencils in some of the envelopes, and you can use these to draw the electrodes accurately onto your paper so you know where they are.

The photo, for example, shows a pattern with two point charges traced onto the paper.

Each group will be assigned to trace one equipotential curve. Let's say yours is 1.0 V. Now put the U-probe in place so that the top is above the equipotential board and the bottom of it is below the board. You will first be looking for places on the pattern board where the voltage is one volt — look for places where the meter reads 1.0 and mark them through the hole on the top of your U-probe with a pencil or pen. You should find a whole bunch of places there the voltage equals one volt, so that you can draw a nice constant-voltage curve connecting them. (If the line goes very far or curves strangely, you may have to do more.)

If you're using the PRO-100 meters, they will try to outsmart you by automatically choosing a range. Most people find this annoying. To defeat this misfeature, press the RANGE button, and you'll see the AUTO indicator on the screen turn off.

Analysis

Once you have your field pattern, make copies for your whole group, and then use a pencil to superimpose a sea-of-arrows representation of the field.

In addition to getting some direct experience with the electric field, the other goal of this lab is to test Maxwell's equations. In the case of electrostatics, Maxwell's equations are $\text{div } \mathbf{E} = 4\pi k\rho$ and $\text{curl } \mathbf{E} = 0$. The pattern of equipotentials that you measure in this lab cannot serve as a test of the zero curl, since we assumed in constructing it that a potential existed, which is equivalent to an assumption of zero curl. However, we do have several ways in which we can test Maxwell's equations with our data. With boundary conditions that $\mathbf{E} = 0$ at infinity, and the potential set to fixed values on the electrodes, Maxwell's equations have a unique solution, which has certain properties.

1. There is no way for charge to go any-

where on the apparatus except the electrodes, so we should have $\text{div } \mathbf{E} = 0$ everywhere else, i.e., no field lines should begin or end in empty space.

2. The potential should be constant on each electrode (which is one of our boundary conditions).
3. The field should have the same two symmetries as the electrodes under reflection across the two axes.
4. Because the charges within each plate repel one another, the equilibrium state should be one in which the charges are more dense near the ends of each plate.
5. As discussed in example 2 on p. 93 and problem 17, p. 108, the charge density should be high near the sharp corners of the electrodes.
6. The voltmeter measures the work done per unit charge on a trickle of charge that it allows to flow through *itself*. If the electric field had a nonzero curl, then this work would not necessarily equal the work done along any particular across the surface of the board, and it could depend on the details of how we positioned the voltmeter and the wires running to it. If, on the other hand, we do see consistent readings on the meter, this supports the assumption that a well-defined potential exists.

Exercise 4B: A preview of the electric potential and measurement of voltages

This exercise is meant to be done before reading ch. 4.

For certain types of electric fields, including fields that don't change over time, we can define an electric potential, ϕ , which is a measure of electrical potential energy per unit charge. It has units of volts, $1 \text{ V} = 1 \text{ J/C}$. In a circuit, we can make an analogy with the behavior of a fluid flowing through pipes: the flow is the *current*, while the pressure causing the flow is the *voltage*.

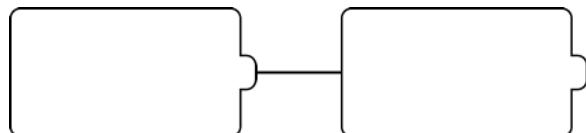
See figure g, p. 90, for instructions on how to use a meter to measure the *difference* in potential between two points, $\Delta\phi = \phi_2 - \phi_1$. Details differ, but meters are basically standardized.

(1) Use the meter to measure the voltage of a battery. The battery is like a pump that creates a pressure difference.

(2) What happens when you do the following things?

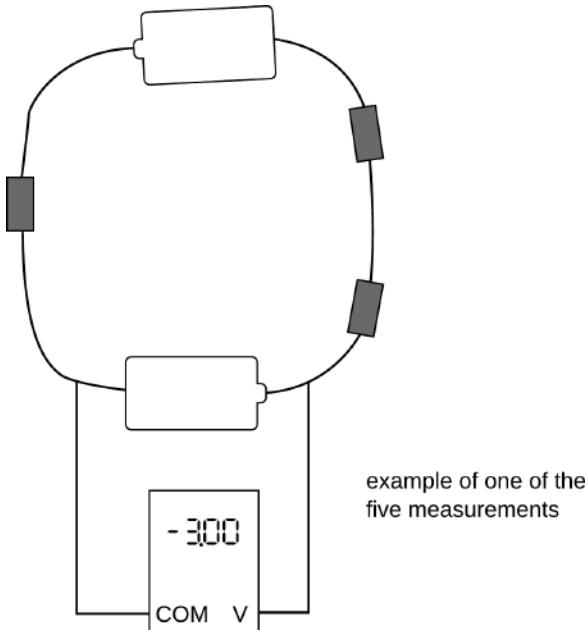
- Reverse the connections.
- Connect both leads to the same terminal of the battery.
- Only connect one lead.

(3) Put two batteries in series:



This is most easily done using battery holders, alligator clips, and banana-plug cables. Measure the $\Delta\phi$ between the ends. The two cables from the voltmeter can simply be touched to the two ends, using them like probes; it is not necessary to make hard connections every time you measure a voltage.

(4) Obtain three resistors with different values. A resistor is like a narrow pipe that resists the flow of a fluid when a pressure difference is applied. This part of the exercise works best if you use three resistors whose values are all different, but not wildly different. Build this circuit and measure all 5 of the potential differences across the individual components.

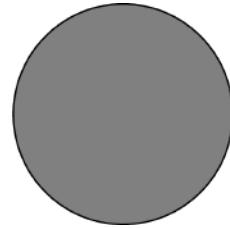
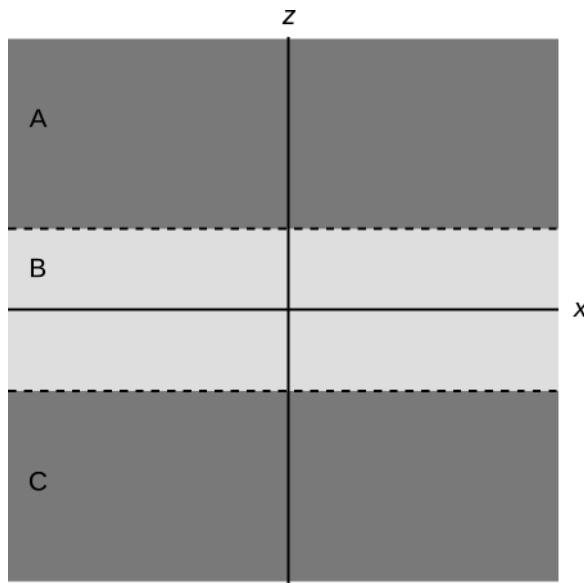


Walk the meter around the circuit so that the V plug is always counterclockwise from the COM plug. Can you find any pattern or relationship in the five numbers?

Exercise 4C: Charge density, field, and potential

(1) A charge distribution is given by

$$\rho = \begin{cases} -a, & z > b \\ +a, & -b < z < b \\ -a, & z < -b \end{cases}$$



(2) A sphere of radius R has charge q distributed uniformly throughout its volume.

(a) Find the electric field. (The boundary condition is that $\mathbf{E} = 0$ at infinity.)

(b) Find the potential.

(a) Use Gauss's law to find the electric field in region B as a function of z (not just at some special location such as $z = b$). As a boundary condition, we let $\mathbf{E} = 0$ at $z = 0$, as suggested by the symmetry of the charge distribution. Without this boundary condition, we could add any uniform field and get a different solution.

Similarly, find the field in region A. To save time, it is not necessary to explicitly write out the field in region C, because it is related by symmetry to the field in A.

(b) Check your result by verifying that $\text{div } \mathbf{E} = 4\pi k\rho$.

(c) Find the potential. You will have three constants of integration, two of which can be determined by requiring that ϕ be a continuous function.

(d) Sketch ρ , E_z , and ϕ as functions of z . Classify these functions as even or odd.

Minilab 4D: Testing the curliness of the electric field

Apparatus

1.5 V battery
battery holder
multimeter
replacement fuses for multimeter
banana-plug cables
capacitors (bipolar electrolytic, 50-330 μF)

Goal: Test whether the electric field in a static electric circuit is curly.

In this lab you will act out the concept presented on p. 91 of testing whether the sum of the voltage differences around a closed loop in a circuit is equal to zero.

As a practical matter, we use capacitors for this experiment that have plates with large surface areas, and with a very small distance between the plates. Such a capacitor is said to have a large *capacitance*. The detailed definition of capacitance and its units of measurement are presented later in this course. These definitions don't matter very much here — in fact, this lab doesn't even need to be done using capacitors. It could be done with other electrical components such as resistors. All that matters is whether the fields are static. The only practical reason for using large capacitance values is that if we had used small capacitances, connecting the voltmeter to the circuit would have tended to allow the charge on the capacitor plates to leak off rapidly through the meter, ruining the measurements. For the capacitance values used here, this process of leaking is quite slow; when I tested it, I found that it took about half an hour for the readings to change by only 15%. So as long as you don't leave the meter connected to the circuit for long periods of time, there won't be a problem.

Another practical issue is that most high-value capacitors are *polar*, meaning that you can apply a voltage difference across them only in one direction. If you do it the wrong way,

smoke may start coming out! The capacitors we will use here are not polar ("bipolar"). High-value, bipolar capacitors are somewhat expensive. The ones we're using cost about \$4 each.

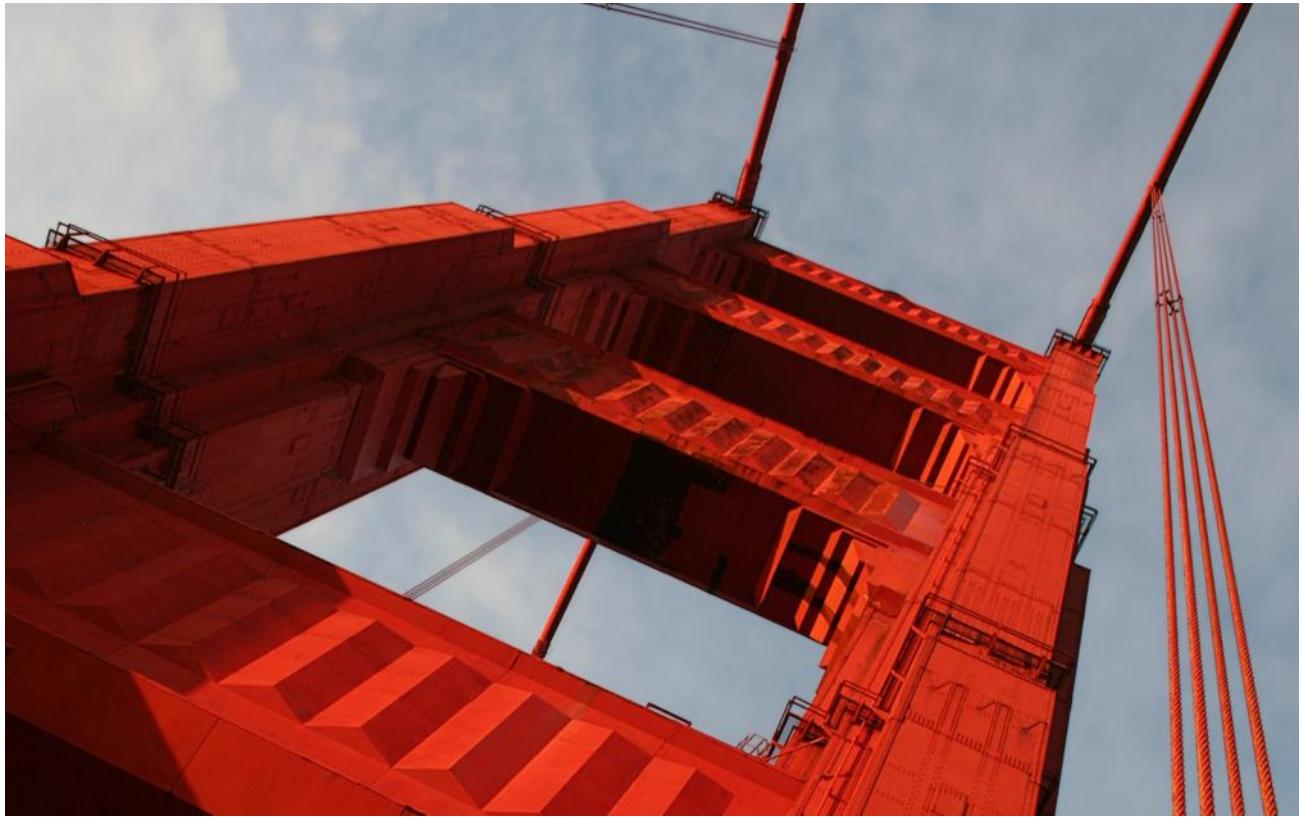
Use your four capacitors, plus the battery, to build a circuit. To make this fun, don't make your circuit all series (like a single chain of a necklace) or all parallel (with no islands of wire between two capacitors). Make it some combination of series and parallel parts. Make sure that all of your capacitors are connected at both ends — if they have one end dangling in the air, then that branch is an open circuit, which might as well have not come to the party.

If you draw the schematic for this circuit, you should find that it's possible to lay it out in such a way that no wires cross over other wires.³ We can then single out loops that are the innermost loops, as in the example on p. 91. For each inner loop, use the meter to check whether the sum of voltage drops around the loop is zero as predicted by theory.

This circuit can exist in more than one state of equilibrium. There is a possible state in which every "island" of metal between capacitors has zero total charge, as well as other possible states in which this is not true. In theory, this makes no difference as to whether the field is conservative. However, it would be nicer to get the circuit into a known state so that, for example, if you need to rebuild it later to get a missing piece of data, you can do so. To do this, use the following procedure. (1) Build the circuit. (2) Take the battery out of the holder. (3) Use a wire to connect one side of one of the capacitors to the other. (You can just touch it to the wires, you don't have to shove in the banana plugs.) This is called shorting across the capacitor. This allows charge to flow freely between the capacitor's two plates, so that it isn't blocked by the existence of the gap between the plates. Do this for each capacitor in

³See https://en.wikipedia.org/wiki/Planar_graph.

turn, and then go back and repeat the process a few more times (or short all the capacitors simultaneously, if you have enough hands and wires). (4) Put the battery back in.



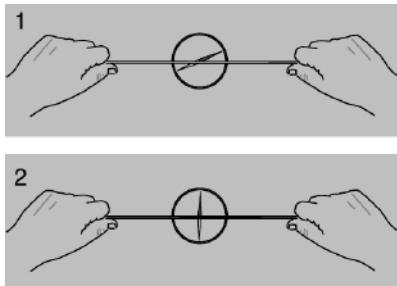
A tower and cables of the Golden Gate Bridge. The pressure and tension in the tower and cables are properties of the electric and magnetic fields at the atomic level.

Chapter 5

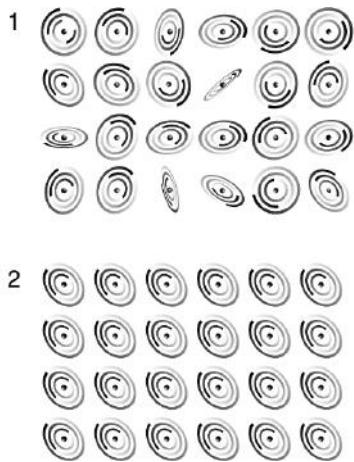
Electromagnetism

5.1 Current and magnetic fields

At this stage, you understand roughly as much about the classification of interactions as physicists understood around the year 1800. There appear to be three fundamentally different types of interactions: gravitational, electrical, and magnetic. Many types of interactions that appear superficially to be distinct — stickiness, chemical interactions, the energy an archer stores in a bow — are really the same: they're manifestations of electrical interactions between atoms. Is there any way to shorten the list any further? The prospects seem dim at first. For instance, we find that if we rub a piece of fur on a rubber rod, the fur does not attract or repel a magnet. The fur has an electric field, and the magnet has a magnetic field. The two are completely separate, and don't seem to affect one another. Likewise we can test whether magnetizing a piece of iron changes its weight. The weight doesn't seem to change by any



a / 1. When no charge flows through the wire, and the magnet is unaffected. It points in the direction of the Earth's magnetic field. 2. Charge flows through the wire. There is a strong effect on the magnet, which turns almost perpendicular to it. If the earth's field could be removed entirely, the compass would point exactly perpendicular to the wire; this is the direction of the wire's field.



b / A schematic representation of an unmagnetized material, 1, and a magnetized one, 2.

measurable amount, so magnetism and gravity seem to be unrelated.

That was where things stood until 1820, when the Danish physicist Hans Christian Oersted was delivering a lecture at the University of Copenhagen, and he wanted to give his students a demonstration that would illustrate the cutting edge of research. He used a battery to make charge flow through a wire, and held the wire near a magnetic compass. The idea was to give an example of how one could search for a previously undiscovered link between electricity (the charge flowing in the wire) and magnetism. One never knows how much to believe from these dramatic legends, but the story is that the experiment he'd expected to turn out negative instead turned out positive: when he held the wire near the compass, the charge flowing through the wire caused the compass to twist!

Oersted was led to the conclusion that when matter creates magnetic fields, it happens because the matter contains moving charges. A permanent magnet, he inferred, contained moving charges on a microscopic scale, but their motion simply wasn't practical to detect using human-scale measuring devices in the lab. Today this seems natural to us based on the planetary model of the atom. As shown in figure b, a magnetized piece of iron is different from an unmagnetized piece because the atoms in the unmagnetized piece are jumbled in random orientations, whereas the atoms in the magnetized piece are at least partially organized to face in a certain direction.

Not until later in this book will we get into the mathematical and geometrical details of the magnetic fields created by moving charges. However, we can immediately make some far-reaching conclusions. Figure c/1 shows, in a cartoonish way, the fact that a line of positive charges, at rest, makes an electric field in the surrounding space. If we approximate the charges as a continuous line with no gaps, then the electric field is one we have already studied in sec. 2.7, p. 58: it points outward, and its magnitude is proportional to $1/r$, where r is the distance from the line.

But now let's switch to a different frame of reference, as in figure c/2. In this frame, the charges are moving, so there is both an electric field and a magnetic field. We are led to the following important conclusion:

Electromagnetism

A certain mixture of electric and magnetic fields will be measured as a different mixture by an observer in a different state of motion.

This shows that electric and magnetic fields are not entirely separate things, but are instead two sides of the same coin. The conclusion holds regardless of whether matter is present. For these reasons, the entire subject of electricity and magnetism is often re-

ferred to by the term electromagnetism. We often refer loosely to “the electromagnetic field,” meaning both **E** and **B** collectively.

In the language of section 2.6.2, p. 56, the fields are not invariant quantities. They depend on the frame of reference, as do other more familiar quantities such as velocity. When we want to convert a velocity vector from one frame of reference to another, we have a rule for doing so, which is simple vector addition. In fancy language, this is called “transforming” the velocity vector. The rules for transforming electric and magnetic fields are more complicated, and we will postpone discussing them until sec. 7.1, p. 175. However, one thing we can say about the transformations is that they must be *additive*, because the laws of physics obey the principle of superposition, and we want the laws of physics to be equally valid in all frames of reference.

By the way, the concepts that we are describing using the words “superposition” and “additive” are usually referred to by mathematicians using the terms “linearity” and “linear.” For example, a mathematician would describe the derivative rule $(f + g)' = f' + g'$ from freshman calculus as the linearity of the derivative.

Throwing salt doesn't make a magnet

example 1

Figure d shows a salt crystal. The smaller, darker spheres are sodium atoms. Bigger, lighter ones are chlorine. When these disparate atoms assemble themselves into a solid, some charge is transferred from the chlorines to the sodiums. This is essentially the same thing that is going on in static electricity examples such as the sticky tape in figure a, p. 41: different substances “like electrons” to different extents, so when they are put in contact, one steals from the other.

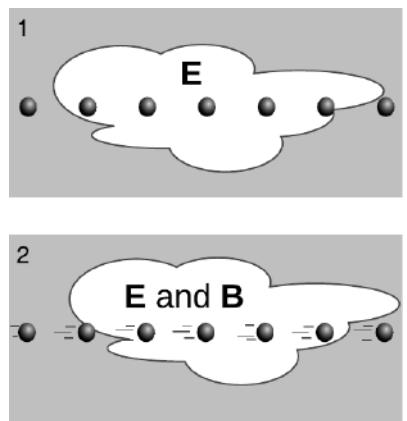
In the frame of reference where the crystal is at rest, it has an outward electric field \mathbf{E}_1 due to the sodiums and an inward field \mathbf{E}_2 from the chlorines. These two fields are theoretically incredibly intense, but due to superposition, they almost exactly cancel at the macroscopic scale, and cannot be detected outside the crystal. (At the microscopic scale, where it is evident that the positive and negative charge distributions aren't exactly the same, the cancellation fails, and there are intense fields. These fields are what hold the crystal together.)

If we throw the salt crystal, then the field \mathbf{E}_1 of the sodiums transforms to some mixture of an electric field plus a magnetic field \mathbf{B}_1 , and similarly we have a \mathbf{B}_2 from the chlorines. But because $\mathbf{E}_1 + \mathbf{E}_2 = 0$, and because the transformation is additive, we have $\mathbf{B}_1 + \mathbf{B}_2 = 0$ as well, and no observable magnetic field is produced.

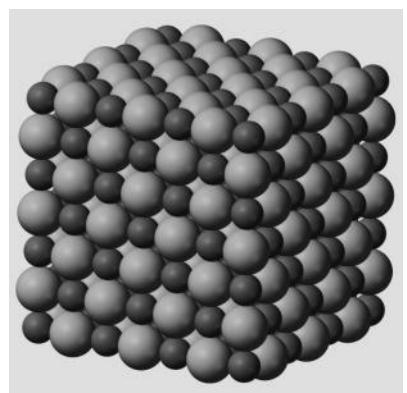
Magnetic field of a wire

example 2

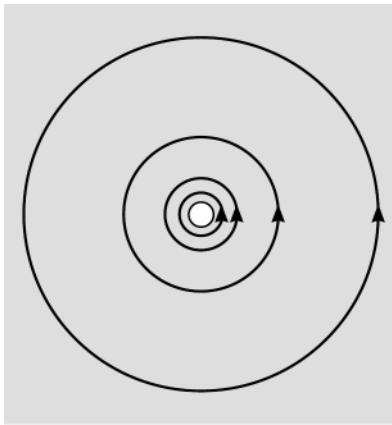
Since the electric field in c/1 is proportional to $1/r$ (sec. 2.7 and



c / 1. A line of positive charges is at rest. There is an electric field. 2. In a different frame of reference, the same charges are seen as moving to the right. There is both an electric field and a magnetic field.



d / A salt crystal, example 1.



e / Example 2, the magnetic field of a wire with charge flowing through it. The field pattern is shown in the plane perpendicular to the wire. The orientation is rotated by 90 degrees relative to figure c. The white circle at the center is the cross-section of the wire.

example 8, p. 58), the electric and magnetic fields in c/2 are also proportional to $1/r$.

Now consider a wire such as the one in figure a/2. Such a wire is electrically neutral, containing equal numbers of positive and negative charges. In the condition where charge is flowing, the charges of one sign are standing still while the ones with the other sign move. (In a metal wire, the moving charges are the electrons and the stationary ones are the nuclei.) This is essentially the situation in figure c/2, except that the electric fields cancel out, leaving a purely magnetic field. We conclude that when charges flow through a wire, the magnetic field surrounding the wire is proportional to $1/r$. We'll work out the constants of proportionality in example 6 on p. 181.

What is the direction of this field? It is not hard to show based on symmetry arguments that the radial component of the field must be zero ([2141](#)), and this is also suggested by Oersted's experimental evidence (figure a, p. 120). Rather, the magnetic field circulates around the wire as shown in figure e.

In cases such as example 2, we find that the magnetic field depends only ([2141](#)) on an electrical quantity called the *current*, defined as the number of coulombs per second that flow past a given point. As with a river or a pipe carrying water, the same current could be created by a large amount of charge moving slowly, or a small amount with a high velocity. Current is denoted I , and is defined formally as

$$I = \frac{dq}{dt}.$$

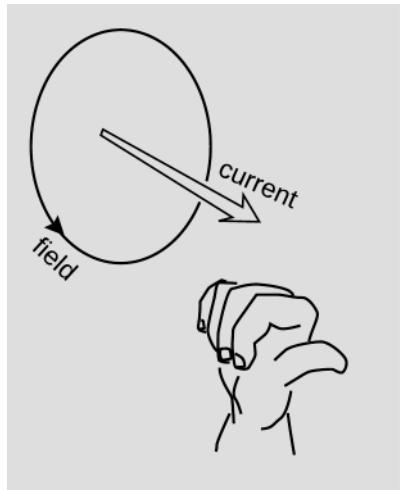
Its units are normally abbreviated as amperes ("amps"), $1 \text{ A} = 1 \text{ C/s}$.

self-check A

Why are the field lines in figure e unevenly spaced? \triangleright Answer, p. 431

In figure e, there was no obvious reason why the magnetic field should have been counterclockwise rather than clockwise. This depends on the direction of the current, as shown in figure f. The fact that we use a right-hand rule for this is clearly nothing basic about physics — the right hand was chosen because humans are mostly right-handed. There are actually two arbitrary conventions behind this, which we encountered in sec. 2.1.1, p. 41. First, Ben Franklin arbitrarily choose one type of charge q to call positive and one to call negative, and this also implies a definition of the sign of the current dq/dt . Second, someone had to resolve the ambiguity between defining the magnetic field to be our civilization's \mathbf{B} , or instead $-\mathbf{B}$. Given these two arbitrary choices, we get the right-hand rule illustrated in the figure.

A straight wire like the one in figure e is not usually a very efficient or compact way of making a strong magnetic field. A more



f / Right-hand rule for the direction of the magnetic field created by the current in a straight wire.

practical device is a solenoid, figure g. When very strong fields are required, the field can be increased by an additional large factor by adding a core made of a magnetic material such as iron. The mathematics for calculating such fields will be deferred to later in this course.

The creation of magnetic fields is not the only effect associated with currents. For example, a current normally causes heating, as in an electric heater or an old-fashioned incandescent lightbulb. Currents can carry information. Your nervous system operates using electric currents carried by ions in your nerve cells. So current is of great practical interest in electrical circuits, which we will study more intensively later in this course. To measure current, there is a device called an ammeter. (Today, people usually use a device called a multimeter, which has multiple functions including working as an ammeter.) We will discuss the ammeter and its use in more detail in sec. 8.1.2, p. 191, when we begin our study of electric circuits.

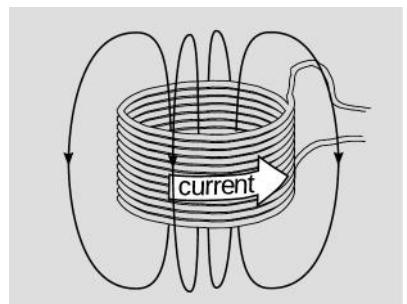
Currents are formed by matter, and are *not* the same thing as electromagnetic fields, which can exist in a vacuum. For example, the current in a copper wire is carried by electrons, whereas sunlight, an electromagnetic wave, travels to the earth through outer space, which is a vacuum.

When the current in an electric circuit is constant, so that charge is flowing at a constant rate, this is referred to as direct current, or DC. Household wall sockets are alternating current, AC, meaning that the current oscillates as a function of time. In the US, electrical power is transmitted as a sine wave with a frequency of 60 Hz. Radio transmitting antennas use AC currents to create an electromagnetic wave, and receiving antennas change the radio signal back to a current. The frequency used in a cell phone is typically about 10^9 Hz.

5.2 Energy, pressure, tension, and momentum in fields

5.2.1 Momentum

Material objects can have energy, momentum, pressure, and tension. We have already seen that electric and magnetic fields have energy, and we will soon see that they carry the other properties on this list as well. Indeed, for an object like the suspension bridge in the photo on p. 119, it is an illusion that the pressure in the towers and the tension in the cables are provided by the steel. Matter is essentially empty space with a sprinkling of pointlike subatomic particles, and the pressure and tension in the bridge are in fact properties of the electric and magnetic fields that exist in this empty space. (Try not to think about this the next time you drive across a bridge.)



g / A solenoid, and its magnetic field pattern. Connections on the right, not shown, are required in order to complete the circuit and provide a source of energy such as a battery.

Summary of energy and momentum densities

$$dU_E = \frac{1}{8\pi k} E^2 dv$$

$$dU_B = \frac{c^2}{8\pi k} B^2 dv$$

$$dp = \frac{1}{4\pi k} \mathbf{E} \times \mathbf{B} dv$$

There is a reason that we discussed the energy density of fields way back in chapter 1, but are only now getting to the momentum density. This is because, for straightforward mathematical reasons, momentum can never be carried by a pure electric or pure magnetic field, but only by a combination of both. That is, the momentum density is an electromagnetic property. To see this, consider the fact that we have only two ways of multiplying vectors: the dot product, which is a scalar, and the cross product, which is a vector. (These operations were reviewed in sec. 1.3.6, p. 25.) Energy is a scalar, so it makes sense that the energy in the fields goes like $\mathbf{E} \cdot \mathbf{E}$ and $\mathbf{B} \cdot \mathbf{B}$, i.e., like the squared magnitudes of the fields. Momentum is a vector, so the momentum density must be a cross product of the fields. But the cross product of parallel vectors is always zero, so expressions like $\mathbf{E} \times \mathbf{E}$ and $\mathbf{B} \times \mathbf{B}$ vanish identically and are useless for our purposes. The momentum density must therefore be proportional to $\mathbf{E} \times \mathbf{B}$, so it is a joint property of the two fields. Filling in the correct constant of proportionality, which we'll come back to in a moment, it turns out that the momentum is given by

$$dp = \frac{1}{4\pi k} \mathbf{E} \times \mathbf{B} dv.$$

The argument given above only demonstrated that the momentum density had to be proportional to $\mathbf{E} \times \mathbf{B}$, but it didn't fix the proportionality constant of $c^2/4\pi k$. The c^2/k has to be there because of units, but we might wonder why the unitless factor isn't simply zero rather than $1/4\pi$. After all, we don't notice this momentum in everyday life; for example, when we turn on a flashlight, it doesn't recoil like a gun. The answer is that the physical quantities on the list we've been discussing — energy, momentum, pressure, and tension — are not independent things. They're all intimately related. For example, if we want to change the kinetic energy of a car, we have to change its momentum as well ([Z141](#)). If a radio signal comes along and pumps kinetic energy into the electrons in the antenna of your phone, then it's also transferring momentum to them, and therefore it must have some momentum itself.

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 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. 6.6.1, p. 161), 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 3

Halley's comet, shown in figure h, 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 4

Figure i 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 $0.001 \mu\text{N}$. For comparison, a short clipping of a human hair weighs $\sim 1 \mu\text{N}$.

The hydrogen bomb

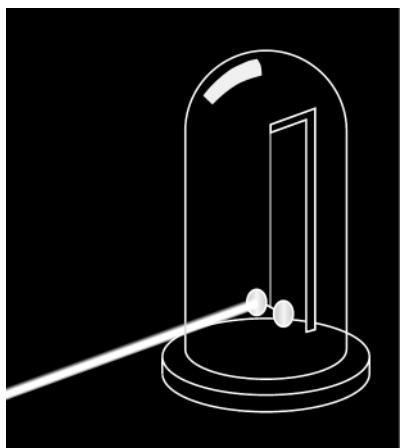
example 5

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 3 and 4, one might wonder why we should even bother dis-



h / Halley's comet, example 3.



i / Example 4.

cussing 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 disturbances in the electric and magnetic fields are never purely electric waves or magnetic waves. They carry momentum, and therefore they must contain an oscillation of both fields. This is why phenomena like light and radio waves are referred to as electromagnetic waves.

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.

5.2.2 Pressure and tension

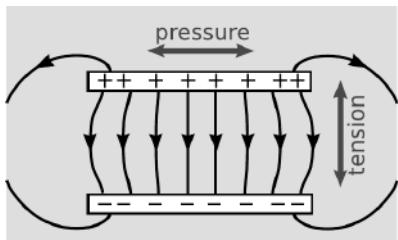
Pressure is defined as the force per unit area, applied perpendicular to the area. It has units of newtons per square meter, which can also be abbreviated as pascals, $1 \text{ Pa} = 1 \text{ N/m}^2$. The earth's atmospheric pressure is about 100 kPa, and car tires are usually inflated to about 250 kPa over atmospheric pressure.

Whenever momentum is transferred from one region of space to another, pressure is involved, and in fact this is an equally valid definition of pressure. For example, when a bat hits a baseball, the rate at which momentum is transferred into the ball, per unit area, is the pressure.

Some materials can sustain tension, which is negative pressure. For example, a rod can sustain either tension or pressure, but a rope can only sustain tension — you can't push with a rope.

In all of the examples above involving material objects, the pressure and tension are in fact properties not of the objects but of the electric and magnetic fields, which are the glue holding the atoms together. Although it is possible to give formulas for the pressure and tension in the electromagnetic field, we will find it more useful in this course to develop some visual rules for making accurate inferences based on pictures of the fields.

As a simple first example of these visual rules, consider the arrangement in figure j, in which positive and negative charges are spread across two parallel metal plates. This is called a capacitor. The field is drawn by hand to have roughly the features we would find if we added up the fields of a large number of point charges on the two plates. The field lines begin on the positive charges, which are the sources of the electric field, and end on the negative ones, which are the sinks. The field is nearly uniform on the interior



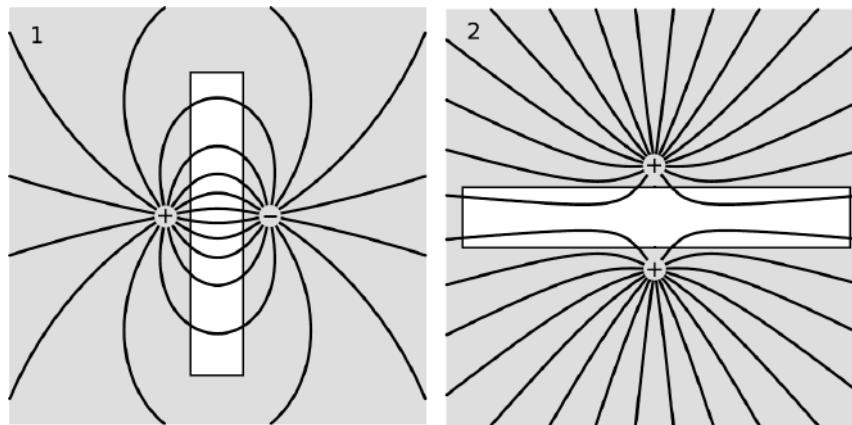
j / Pressure and tension in the field of a parallel-plate capacitor. Side view.

and nearly zero on the exterior. To understand the reason for this behavior, recall that for an infinite plane of charge, the field is constant on both sides (sec. 2.7, p. 58). Since the two plates are finite, their individual contributions to the total field are nearly constant on either side of each one. If they were infinite, then these fields would be exactly constant, and their sum would exactly cancel on the exterior, while reinforcing on the interior.

We know from Coulomb's law that the positive and negative charges are attracting one another, so the plates are being pulled toward each other. The capacitor wants to collapse, and must be prevented from doing so by some external structure (not shown). This is a tension in the vertical direction. The field lines act like taut ropes.

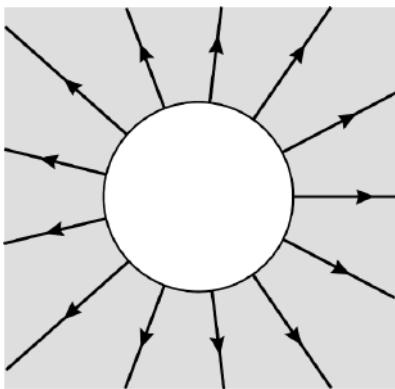
We also know that within each plate, the like charges are all repelling each other, and this would cause the plate to explode in the horizontal direction, except that again there must be something else holding it together. This is a kind of pressure. We see that pressure is exerted in the direction perpendicular to the field lines, as if they were physical objects trying to stay away from one another.

These are general rules, which apply to both the electric and the magnetic fields: field lines sustain tension parallel to themselves, and pressure in the perpendicular direction. Pressure and tension only produce mechanical effects when there is a *difference* in pressure or tension. For example, a car with a flat tire has air pressure inside the tire, but there is no difference capable of creating a net outward mechanical force on the rubber. In figure j, the mechanical stresses on the capacitor exist because there is no field on the outside.

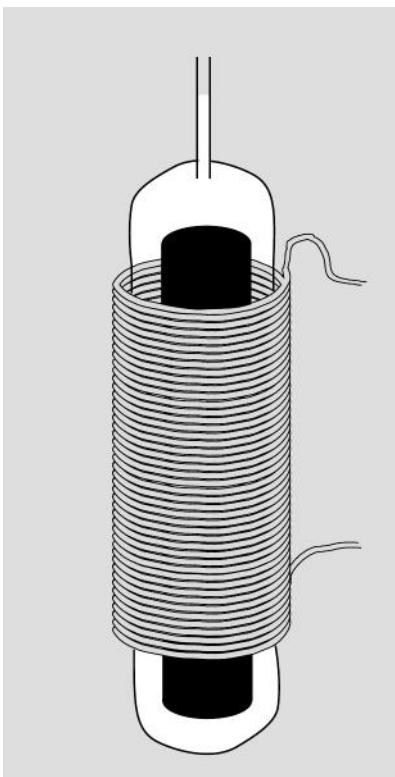


k / Tension and pressure, for two-charge systems.

Figure k shows two examples. In k/1, the highlighted area between the opposite charges contains field lines oriented horizontally, which create tension in that direction. The charges attract one another. There is also pressure in the vertical direction, but this is in the wrong direction to produce any effect on the charges, which are



I / Tension in the electric field surrounding a charged sphere.



m / Example 6. The apparatus was about half a meter tall.

separated from each other horizontally.

In $k/2$, the highlighted area between the two positive charges contains field lines that exhibit vertical pressure, causing a repulsion between the charges, along with a horizontal tension which has no effect on them.

In both examples in figure k, we have highlighted an area between the charges, but it also matters that there is a difference between the pressure or tension in this area and that on the outside.

To see the advantage of this mode of reasoning, compare with the much more complicated logic that we had to do on p. 42 in order to analyze the situation k/1 by talking about the change in the fields' energy that would have occurred if they were moved closer together or farther apart. (Cf. also the complicated handling of the limits of integration in note [264](#).)

Figure 1 shows the example of a charged, conducting sphere. The charge distributes itself uniformly on the surface, and the field is zero on the inside. The surface of the sphere wants to explode because of the electrical repulsion. We confirm this because there is tension on the outside of the sphere, but none on the inside. This imbalance causes any piece of the surface to feel a net outward force.

Magnetostriction

example 6

The examples discussed above involved pressure and tension made by electric fields, but an exactly analogous effect occurs for magnetism. If you've ever heard a transformer buzzing, you've observed it.

Figure m shows part of a sensitive experiment carried out by James Joule in 1842 which discovered the effect. Joule writes, "About the close of the year 1841, Mr. F.D. Arstall, an ingenious machinist of this town, suggested to me a new form of electromagnetic engine. He was of the opinion that a bar of iron experienced an increase of bulk by receiving the magnetic condition." Joule first constructed a delicate experiment to look for a change in the length of an iron bar when it was used as the core of a solenoid. In an iron bar about half a meter tall, he found a *decrease* in length of a little more than one part per million, with the strongest field he could make. This is called magnetostriction.

This would suggest a decrease in volume as well, but there was no sufficiently accurate way to measure the small diameter of the bar. Therefore Joule created the apparatus in figure m, in which the bar was submerged in water, with a thin capillary tube at the top to take the volume of any small displacement of water and amplify it into a visible change in height. The result was no detectable volume change.

These results seem natural in view of our picture of fields as carrying tension in the longitudinal direction and pressure in the transverse direction. The interior field of the solenoid, as shown in figure g, runs in the vertical direction. The lengthwise tension reduced the length of the bar, but the transverse pressure expanded its diameter, and the result was that the volume stayed the same.

In the more familiar modern example of the buzzing transformer, an AC current at 60 Hz creates oscillations at a frequency of 120 Hz. This doubling of the frequency occurs because the energy, pressure, and tension of a field all depend on the square of the field, so that for a magnetic field, \mathbf{B} produces the same effect as $-\mathbf{B}$. This means that the maximum magnetostriction occurs twice per cycle, at both the peak and the trough of the sine wave.



n / A mechanical example of shear forces. If the plate of jello is disturbed horizontally, internal shear forces bring it back to equilibrium.

It is beyond the scope of this book to work out fully general equations for the pressure and tension in the fields. In the most general case, when both an electric and a magnetic field are present, this gets a little complicated, and one needs to consider not just pressure and tension but also *shear* forces, figure n. However, when only one field is present, we can make a few simple observations. Let's say this is an electric field. Symmetry prevents the existence of any shear forces, so we can only have tension in the direction parallel to the field and pressure along the two axes perpendicular to it. Symmetry requires that the pressures along these two axes be equal. Furthermore, if the only quantities we have handy are k and E , then units require that the pressure and tension be proportional E^2 (problem 21, p. 149), as we would have expected, since rotational invariance requires that the pressure remain unchanged if we replace \mathbf{E} with $-\mathbf{E}$. It also turns out that the pressures are equal to the tension in absolute value ([Z142](#)).

Summarizing, we find that when the field is purely electric or purely magnetic:

- There are only pressure P and tension T , not shear, and both are proportional to the square of the field.
- $|P| = |T|$

5.3 Force of a magnetic field on a charge

Figure o shows the superposition of two magnetic fields. One is a uniform field pointing to the left, and the other is the field of a wire carrying a current that flows into the page. (The circle with a cross is a standard way of indicating a direction into the page, meant to evoke the tail feathers of an arrow. A circle with a dot would mean a direction out of the page.) To construct this diagram, we only needed vector addition plus the facts about the magnetic field of a wire proved in example 2 on p. 121. Since that example did not include any derivation of the constant of proportionality in the relationship $B \propto 1/r$, we can't say, in any particular set of units, what is the current in the wire that would produce exactly this field pattern.

Since the pressure is greater below the wire than above it, there is an upward force on the wire. Summarizing the geometrical relationship, we have:

current into the page
B to the left
force up.

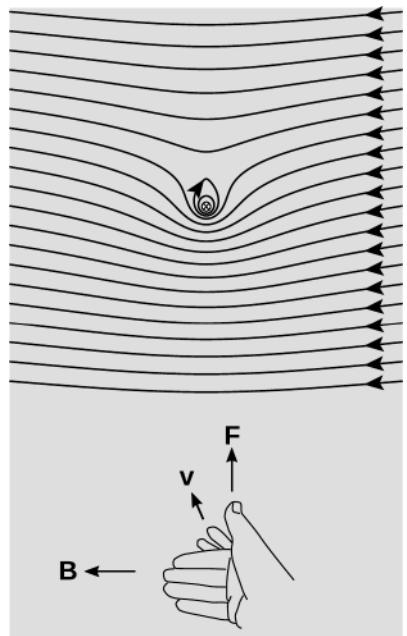
This is a right-hand rule, which smells like a cross product. Since the cross product is the only rotationally invariant way to multiply two vectors in order to get a third vector, the only possible force law consistent with these observations is that when a charged particle moves with velocity \mathbf{v} through a magnetic field \mathbf{B} , the force acting on the particle is proportional to $q\mathbf{v} \times \mathbf{B}$. The constant of proportionality has to be unitless, and although we haven't yet proved it, this constant equals 1. Taking into account the possibility that the particle is also acted on by an electric field, we have

$$\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B},$$

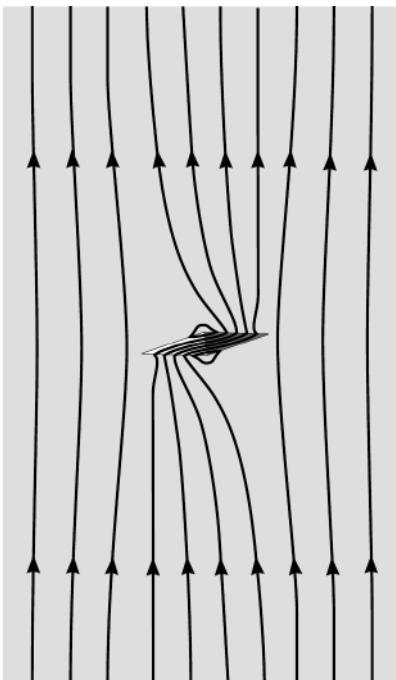
which is known as the Lorentz force law.

As an application, we can now describe how Thomson was able to measure the velocity of the electrons in the experiment in which he discovered the electron. As described in sec. 3.2, p. 76, he measured the deflection of the electrons in an electric field, but he also determined how much magnetic field, applied perpendicular to the beam, was necessary in order to produce the same deflection. He then had $v = E/B$.

Thomson was using a beam of electrons in a vacuum tube, but in applications it is more common for the moving charges to be the ones in a current-carrying wire. When the wire is perpendicular to the field, as in figure o, the force per unit length acting on the wire is $F/\ell = IB$ (problem 13, p. 146).



o / A uniform magnetic field superposed on the field of a wire carrying current into the page.



p / A compass needle in the earth's field.

5.4 The dipole

Figure p shows the superposition of a compass needle's magnetic field (essentially the field of figure m, p. 26) with the earth's magnetic field. The compass is not aligned with the earth's field, and from what we know about magnetic compasses we expect that there will now be a torque that will tend to bring it into alignment. It could in principle be very difficult to find this torque, since the magnetic field pattern is so complicated. But by using the visual techniques of section 5.2.2, it's pretty easy to get the right answer for its sign. There is tension at the top right and bottom left of the needle because of the densely packed field lines. This tension will create a counterclockwise torque.

A real-world hiking compass is usually filled with water in order to create friction. The needle vibrates about an equilibrium, but this friction rapidly kills the oscillations, and it ends up in a state of lowest energy, in which it is aligned with the earth's field ([Z141](#)). If we were to redraw figure p for the equilibrium condition, it would have a left-right mirror symmetry, so the torque must vanish, and this is consistent with equilibrium.

The reason we can treat the earth's field as uniform in this example is that the compass is so small compared the earth. In this approximation, which is an excellent one in this example, we can think of the compass as an idealized, very small object which feels no net force from the field, only a torque.¹ Such an idealized object is called a *dipole*. There can be both electric and magnetic dipoles. An electric dipole tends to align itself with the electric field.

If we change the orientation of a dipole, say a magnetic one, then its own contribution to the field, \mathbf{B}_1 rotates along with it. Meanwhile, the external field \mathbf{B}_2 remains constant. For a given orientation, the superposition of the two fields may be something complicated like figure p. But regardless of how complicated it is, the total energy of this field will always vary with orientation in a very simple way: it is proportional to the cosine of the angle between the external field and an axis, such as the long axis of the compass needle, determined by the structure of the dipole (proof [Z142](#)). This is the behavior of a vector dot product, so we have for the energy

$$U = -\mathbf{m} \cdot \mathbf{B},$$

where \mathbf{m} is a vector called the dipole moment, and the minus sign is a matter of convention.

A mechanical analogy is that if a pendulum has length ℓ and mass M at the end, then we can define a vector with magnitude ℓM

¹To get a nonvanishing force from the earth's magnetic field, we would have to have an isolated "magnetic charge," and the apparent nonexistence of such things is expressed by Gauss's law for magnetism.

and direction pointing from the axis to the mass, and an identical expression holds for the energy of the pendulum in the earth's gravitational field, since the height of the mass is just $-\ell \cos \theta$, where θ is the displacement from equilibrium. But there is no such thing as a pure gravitational dipole, because any material object in a gravitational field will experience a net force as well as a torque.

We could try to be whimsical and get around this by putting a mass at one end of a stick and a helium balloon at the other, tuning up the setup so that the whole thing was neutrally buoyant. Less whimsically, this is similar to the simplest way of making an *electric* dipole: attach charges q and $-q$ to opposite ends of a stick of length ℓ . The result is an electric dipole moment, notated \mathbf{D} , with magnitude ℓq and energy

$$U = -\mathbf{D} \cdot \mathbf{E}$$

in an external electric field.

Dipole moment of a molecule of NaCl gas example 7

- ▷ In a molecule of NaCl gas (cf. example 1, p. 121), the center-to-center distance between the two atoms is about 0.24 nm. Assuming that the chlorine completely steals one of the sodium's electrons, compute the magnitude of this molecule's dipole moment.
- ▷ The total charge is zero, so it doesn't matter where we choose the origin of our coordinate system. For convenience, let's choose it to be at one of the atoms, so that the charge on that atom doesn't contribute to the dipole moment. The magnitude of the dipole moment is then

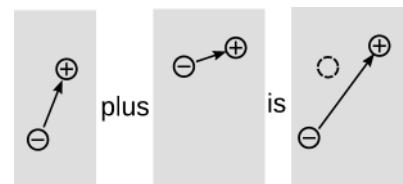
$$\begin{aligned} D &= (2.4 \times 10^{-10} \text{ m})(e) \\ &= (2.4 \times 10^{-10} \text{ m})(1.6 \times 10^{-19} \text{ C}) \\ &\approx 4 \times 10^{-29} \text{ C} \cdot \text{m}. \end{aligned}$$

The experimentally measured value is $3.0 \times 10^{-29} \text{ C} \cdot \text{m}$, which shows that the electron is not completely "stolen."

Because the dot product has the linear property $(\mathbf{D}_1 + \mathbf{D}_2) \cdot \mathbf{E} = \mathbf{D}_1 \cdot \mathbf{E} + \mathbf{D}_2 \cdot \mathbf{E}$, it follows from our definition of the dipole moment that dipole moments add like vectors (and otherwise it would not have been legitimate to call them vectors).² Figure q shows a less abstract justification for this, using electric dipoles made of charges $\pm q$. For the remainder of this section, we'll focus on electric dipoles, returning to magnetic dipoles in sec. 11.4, p. 267.

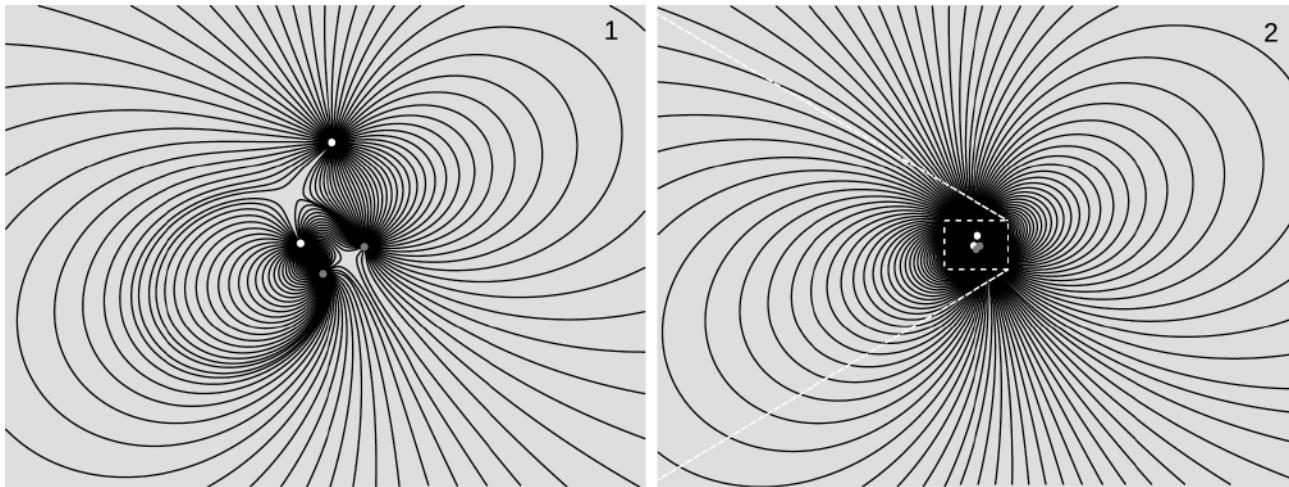
Of course a complicated set of charges really is complicated, when you see it up close. Figure r/1 shows the field pattern made

²This does not necessarily hold for cases with infinitely many charges, as in example 2, p. 75.



q / Dipoles add like vectors. When a two-charge dipole is superposed on another two-charge dipole, we can cancel two of the charges, and the result is that the displacement vectors add.

by two positive charges and two negative ones. It's very complicated, and it's not the same as the field pattern of any two-charge dipole. But when we see it from far away, $r/2$, it looks very simple. It's common sense that you can't easily see the internal structure of an object when you only look at it from far away.



r / 1. The field pattern made by two positive and two negative charges. 2. Seen from 10 times farther away, the pattern looks much simpler. This is what the generic, distant field of any dipole looks like. The dashed rectangle shows the field of view from the left-hand panel.

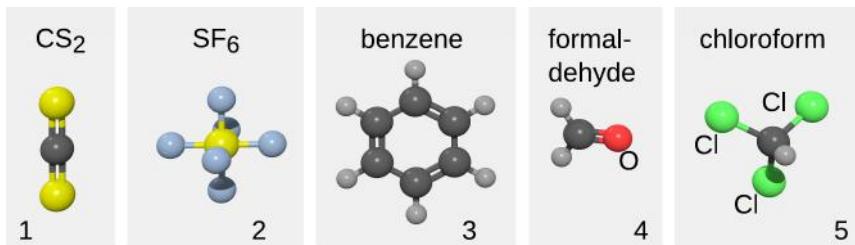
So it's believable that all electric dipoles should have a similar field pattern at large distances (taking into account their dipole moment and orientation), but what is that field pattern? We've already seen in example 3 on p. 51 and problem 9 on p. 69 that the field of a two-charge dipole varies like $1/r^3$ at large distances, both in the mid-plane and on the axis. This also holds true along any line through the dipole, not just lines like these that are oriented in symmetrical ways. (To see this, express the dipole moment as a sum of components perpendicular and parallel to the line.)

So in general, if we throw a bunch of charged particles in a sack, their field at large distances will be proportional to the total charge, and will fall off like $1/r^2$, as we expect from Coulomb's law. There will also be a dipole field, but it falls off more quickly at large distances, so the Coulomb part dominates. But if we happen to throw particles in the sack whose total charge is zero, then there is no $1/r^2$ part, and the dominant field at large distance is the $1/r^3$ dipole field. This is what happens, for example, with molecules that have nonzero dipole moments.

The universal form of electric dipole fields at large distances also holds for magnetic dipoles, as discussed on p. 88. In mini-lab 1 you probably found that the field went approximately like $1/r^3$ at the larger distances.

Molecules with zero and nonzero dipole moments example 8

It can be useful to know whether or not a molecule is polar, i.e., has a nonzero dipole moment. A polar molecule such as water is readily heated in a microwave oven, while a nonpolar one is not. Polar molecules are attracted to one another, so polar substances dissolve in other polar substances, but not in nonpolar substances, i.e., “like dissolves like.”



s / Example 8. The positive x axis is to the right, y is up, and z is out of the page. Dark gray atoms are carbon, and the small light gray ones are hydrogen. Some other elements are labeled when their identity would otherwise not be clear.

In a symmetric molecule such as carbon disulfide, figure s/1, the dipole moment vanishes. For if we rotate the molecule by 180 degrees about any one of the three coordinate axes defined in the caption of the figure, the molecule is unchanged, which means that its dipole moment is unchanged. The only vector that has these properties is the zero vector.

Similar symmetry arguments show that sulfur hexafluoride, s/2, and benzene s/3, have vanishing dipole moments.

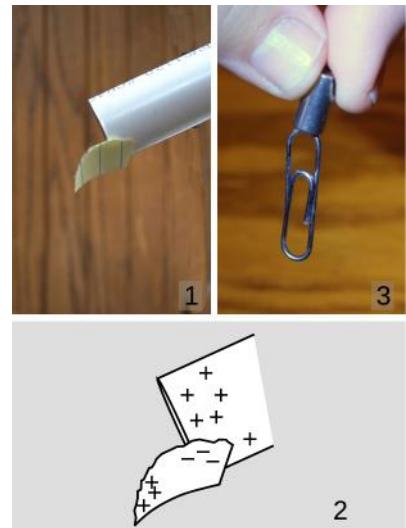
The formaldehyde molecule, s/4, does not have enough symmetry to guarantee that its dipole moment must vanish, but it does have enough to dictate it must be equivalent to a two-charge dipole lying along the left-right axis. Chloroform, s/5, is a front-to-back dipole for the orientation drawn in the figure.

From these considerations we can tell, for example, that carbon disulfide will be soluble in benzene, but chloroform will not.

Symmetry arguments are not enough to determine, for example, whether formaldehyde’s dipole moment points to the left or to the right in the figure. That requires some knowledge of chemistry and the periodic table.

Often we observe static electrical interactions between a charged object and an object that has zero total charge, t/1. The mechanism is shown in figure t/2. Although the scrap of paper has zero total charge, it does contain charges that can move. The positive charges in the pipe attract the negative charges and repel the positive one, turning the paper into a dipole. This is called an induced dipole moment. The attraction is stronger than the repulsion, due to the shorter distance, so the net force is attractive.

A fancier mathematical treatment of the effect is as follows. If



t / 1. A charged piece of plastic pipe attracts an uncharged piece of paper. 2. The mechanism for the effect. 3. The analogous magnetic effect.

a dipole with zero total charge is placed in a uniform field, it may experience a torque, but it will not experience any *force*. The situation changes if the field is nonuniform. The force in the x direction is

$$F_x = -\frac{\partial U}{\partial x}.$$

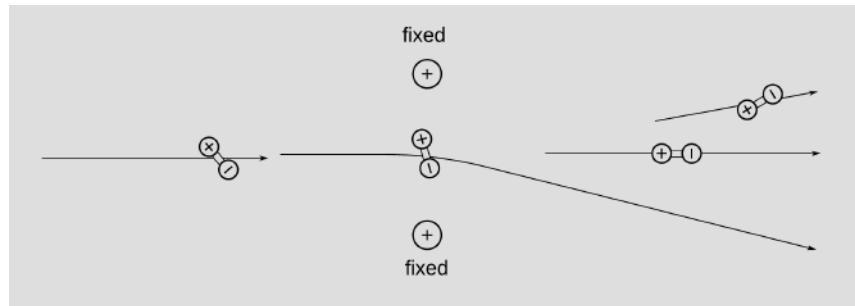
We then have

$$F_x = -\frac{\partial}{\partial x}(\mathbf{D} \cdot \mathbf{E}) = \mathbf{D} \cdot \frac{\partial \mathbf{E}}{\partial x},$$

which depends on the dipole's properties only through \mathbf{D} . Similar expressions apply for the y and z components.

This principle can be used as a way of measuring the unknown dipole moments of a beam of particles, as in figure u. A magnetic version of this device was used in the historic Stern-Gerlach experiment that discovered the spin of the electron.

u / An electric dipole spectrometer. A beam of randomly oriented dipoles is shot through a "croquet hoop" consisting of two fixed positive charges. Although the field along the central axis of symmetry equals zero, the field is nonuniform, and therefore the dipoles feel a nonvanishing force, and are sorted out according to their orientations.



self-check B

In figure u, one of the dipoles is shown emerging from the spectrometer after being deflected upward. Why does it make sense, given its orientation, that this particular dipole was deflected in this direction? ▷

Answer, p. 432

5.5 $E=mc^2$

5.5.1 Fields carry inertia

Suppose you're given a black box, figure v. 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, $m = p/v$.

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

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.³ 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, 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, p. 121), 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.

5.5.2 Equivalence of mass and energy

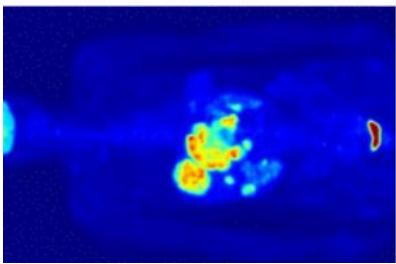
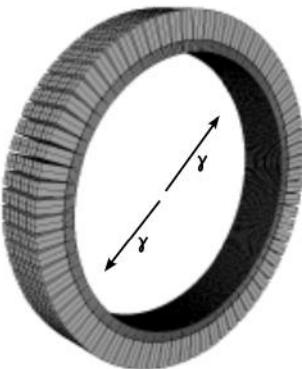
In this example, energy and mass are *equivalent*. Based on units, the relation must be of the form $E = (\text{constant})mc^2$, where the constant is unitless. Einstein showed that the unitless constant was equal to 1, and was the same for any system, regardless of what type or types of energy are involved.⁴ 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 18), and this is why nobody discovered mass-energy equivalence experimentally hundreds of years before Einstein.

A full treatment of this topic would be outside the scope of

³We 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.

⁴Here 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.



w / 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.

this book, but 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 we can have processes that convert one to the other.

A rusting nail

example 9

▷ 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 10

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.

5.5.3 A naughty infinity

In this subsection we discuss a theoretical issue that ends up having far-reaching implications. The reader who doesn't find this side-trip interesting can skip it.

Your mental picture of an electron may be either a point or a small sphere. Which is right? Suppose that we think of it as a sphere of finite radius r , containing a uniformly distributed charge q . Its electric field contains some amount of energy E , and it follows from units that this energy equals kq^2/r , multiplied by some unitless constant. The unitless constant is calculated in problem 9, p. 252, but all that matters for our present purposes is that it's positive, since the energy density is everywhere positive, and the integral of something positive must be positive.

Now clearly bad things will happen if we set $r = 0$. The energy of the electric field will blow up to infinity. Before Einstein, we could have said that this was not necessarily a cosmic catastrophe. A physicist of the 19th century would probably have sniggered and told us patronizingly that the energy E would be of no consequence in any case, because almost all of it was locked up in the part of the field very close to the electron, and was therefore basically a constant and fixed contribution to the total energy of the universe. Adding any constant, no matter how large, to the total energy of the universe has no observable consequences. Worrying about whether r is truly *exactly* zero, our supercilious friend tells us, is something that can never be tested experimentally, and is therefore a meaningless thing to discuss, like how many angels can dance on the head of a pin.

But Einstein's $E = mc^2$ tells us that the issue cannot be waved away so easily. The field's energy is equivalent to a certain amount of mass. It should contribute to the electron's inertia. Now we know what the mass is, so the energy in the field must be $\lesssim mc^2$. This

results in $r \gtrsim r_0 = ke^2/mc^2$, where r_0 is called the classical electron radius, although it doesn't just apply to electrons. Furthermore, there are processes that create electron-antielectron pairs, and these processes don't require infinite energy. The energy required to create such a pair should only be $2mc^2$, which is finite. In general, classical electromagnetism becomes an inconsistent theory if you consider point particles with $r \lesssim r_0$.

Thus relativity seems to suggest that charged particles can't be pointlike, and that there should be a lower bound on their sizes. For an electron, r_0 is on the order of 10^{-15} meters. Particle physics experiments became good enough decades ago to search for internal structure in the electron at this scale, and they tell us that it doesn't exist, in the sense that the electron cannot be a composite particle at this scale. (For comparison, they tell us that a proton *is* composite — it turns out to be made of quarks.)

One could imagine that the electron had a finite size without being composed of smaller particles, which would be consistent with the particle physics experiments, because they basically test whether things can be broken up or internally manipulated, like taking a mechanical pocketwatch and hitting it with a hammer. Maybe there is size without any dynamical internal structure, just a rigid charge distribution. Physicists around 1900 did put great effort into constructing such models of the electron, but they never succeeded. With hindsight, we can see that this approach, too, was doomed because of relativity. Relativity doesn't allow perfectly rigid objects. In such an object, vibrations would propagate as sound waves at infinite speed, but relativity doesn't allow speeds greater than c .

So in the end, if we want to describe the charge and electric field of an electron at scales below r_0 , we need some other theory of nature than classical electricity and magnetism. That theory is quantum mechanics. In nonrigorous language, quantum mechanics describes the scene at this scale in terms of rapid, random quantum fluctuations, with particle-antiparticle pairs springing into existence and then reannihilating.

5.5.4 Summary of relativity

Summarizing what we know so far about relativity, we have the following:

- Time is relative. The rate at which a clock runs is greatest in the frame where the clock is at rest.
- For consistency with item 1, nothing can go faster than c .
- Mass and energy are equivalent, $E = mc^2$.

Notes for chapter 5

2122 Direction of the magnetic field of a wire

The magnetic field of a current-carrying wire has no radial component.

Let a line of charge, with constant positive charge density, lie along the x axis, at rest in frame of reference 1. In frame 1, at a certain point on the y axis, there is a radial electric field E_y and no magnetic field.

Now let frame 2 be moving to the right with speed u , and suppose that the magnetic field at this point did have a nonzero radial component B_y in this frame.

Or, reversing the direction of motion, let frame 3 be moving with velocity $-u$ in the x direction. Because the magnetic field is proportional to the current, the magnetic field at our point of interest must be $-B_y$, as described by an observer in this frame.

But this is untenable, because the transformation of the fields must take as its inputs only the vectors \mathbf{E} and \mathbf{u} , the velocity of the frame of reference relative to the original frame. The vector \mathbf{u} could have any direction in the x - z plane, and there is not enough information in these inputs to determine, by any rotationally invariant function, the sign of B_y .

2122 Proportionality of magnetic field to current

The magnetic field created by a current is exactly proportional to the current.

It is not surprising that the magnetic field of a current-carrying wire, as in example 2, p. 121, is approximately proportional to the current. This must be so based on the fact that the transformation of the fields is additive and is smooth as a function of the velocity. What is more remarkable is that this proportionality to the current is *exact* at all velocities. We will see later that there is theoretical justification for this fact.

2124 Newtonian relation between momentum and kinetic energy

As an example of the intimate relationship between momentum and kinetic energy, the momentum of material objects must be conserved in Newtonian mechanics if conservation of energy is to hold in all frames of reference.

The basic insight can be extracted from the special case where there are only two particles interacting, and they only move in one dimension. Conservation of energy says

$$K_{1i} + K_{2i} + U_i = K_{1f} + K_{2f} + U_f.$$

For simplicity, let's assume that the interactions start after the time we're calling initial, and end before the instant we choose as final. Then there is no change in potential energy, $U_i = U_f$, and we can subtract the potential energies from both sides, giving,

$$\begin{aligned} K_{1i} + K_{2i} &= K_{1f} + K_{2f} \\ \frac{1}{2}m_1v_{1i}^2 + \frac{1}{2}m_2v_{2i}^2 &= \frac{1}{2}m_1v_{1f}^2 + \frac{1}{2}m_2v_{2f}^2. \end{aligned}$$

In a frame of reference moving at velocity u relative to the first one, the velocities all have u added onto them:

$$\begin{aligned} \frac{1}{2}m_1(v_{1i} + u)^2 + \frac{1}{2}m_2(v_{2i} + u)^2 \\ = \frac{1}{2}m_1(v_{1f} + u)^2 + \frac{1}{2}m_2(v_{2f} + u)^2 \end{aligned}$$

When we square a quantity like $(v_{1i} + u)^2$, we get the same v_{1i}^2 that occurred in the original frame of reference, plus two u -dependent terms, $2v_{1i}u + u^2$. Subtracting the original conservation of energy equation from the version in the new frame of reference, we have

$$m_1v_{1i}u + m_2v_{2i}u = m_1v_{1f}u + m_2v_{2f}u,$$

or, dividing by u ,

$$m_1v_{1i} + m_2v_{2i} = m_1v_{1f} + m_2v_{2f}.$$

This is a statement of conservation of momentum.

2132 Minimizing the energy of a compass needle

We discuss the minimization of the energy in the magnetic field of a compass needle.

The result of our analysis of figure p on p. 132 tells us that the compass needle will tend to align itself so that its own internal field is in the same direction as the ambient field. This is a little surprising, since the energy stored in the magnetic field on the interior is then increased. However, the needle also contributes to the field *outside* itself, and this field is predominantly in the opposite direction compared to the interior field. It's true but not obvious that this smaller exterior field, integrated over the large exterior volume, ends up being the dominant effect.

Z132 Dipole's energy proportional to cosine of an angle

We prove that the energy of a dipole in an external field is proportional to the cosine of the angle between the field and some internal axis determined by the structure of the dipole.

In the notation introduced in the main text, \mathbf{B}_1 is the field of the dipole, \mathbf{B}_2 the uniform external field. The energy density of the total magnetic field at any given point is proportional to $(\mathbf{B}_1 + \mathbf{B}_2) \cdot (\mathbf{B}_1 + \mathbf{B}_2)$, with the only variation occurring because the whole field pattern represented by \mathbf{B}_1 can be rigidly rotated. Multiplying out the factors of the dot product, and discarding constant terms, we have $2\mathbf{B}_1 \cdot \mathbf{B}_2$. The variable part of the total energy is found by integrating this expression over all space, but \mathbf{B}_2 is a constant, so we can take it outside the integral, the result being $\mathbf{B}_2 \int \dots dv$, where \dots represents whatever field is going on inside the dipole. This has the form of a vector dot product, which is proportional to the cosine of the angle between the two factors.

Z130 In a pure electric field, $|P| = |T|$

We prove that in a pure electric field, the pressure and tension at a given point are equal in absolute value. Although we won't prove it here, the same holds for the magnetic field.

In the case of a purely electric field, we expect

the pressure and tension to depend only on the square of the field. Therefore, we only need one concrete example in which we know the answer in order to show that the two constants of proportionality are equal.

Physically, the simplest example of this type would be the interaction between two point charges. However, the math turns out a little simpler if we instead consider two parallel lines of charge.

First consider the attractive case, with the two lines of charge having opposite densities λ and $-\lambda$ of charge per unit length. These are parallel to the y axis and located at equal distances above and below the x - y plane, at $z = \pm h$. Their attraction per unit length is proportional to the integral of the tension over the plane that is equidistant from the two lines. Each one contributes a field proportional to $1/r$, and the vector sum of the two fields is perpendicular to the plane and has a magnitude proportional to $r^{-1} \cos \theta$, where θ is the angle between the the z axis and the radial line to the point being considered. Since the tension is proportional to the square of the field, the attractive force per unit length is proportional to the integral $\int_{-\infty}^{\infty} (r^{-1} \cos \theta)^2 dx$. Doing a change of variable to a unitless $u = x/h$, and expressing r , θ , and x in terms of u , we end up with the definite integral

$$\int_{-\infty}^{\infty} \frac{du}{(1+u^2)^2} = \frac{\pi}{2}.$$

Now we consider two lines of charge λ and λ , with the same sign, so that they repel one another. The combined field is now proportional to $r^{-1} \sin \theta$. Discarding all the same constant factors in the same way, but preserving the factors that actually depend on u , we get the integral

$$\int_{-\infty}^{\infty} \frac{u^2 du}{(1+u^2)^2} = \frac{\pi}{2}.$$

The forces are supposed to be equal, and the two definite integrals are equal. We therefore find we have equal constants of proportionality in the relation for the pressure and tension in terms of the field.

Problems

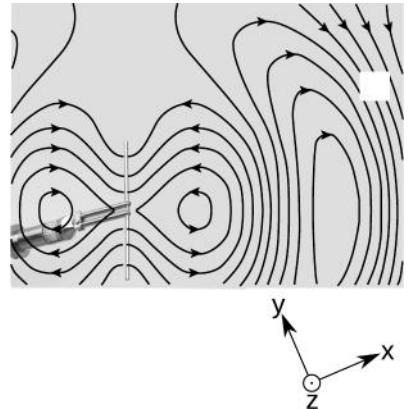
Key

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

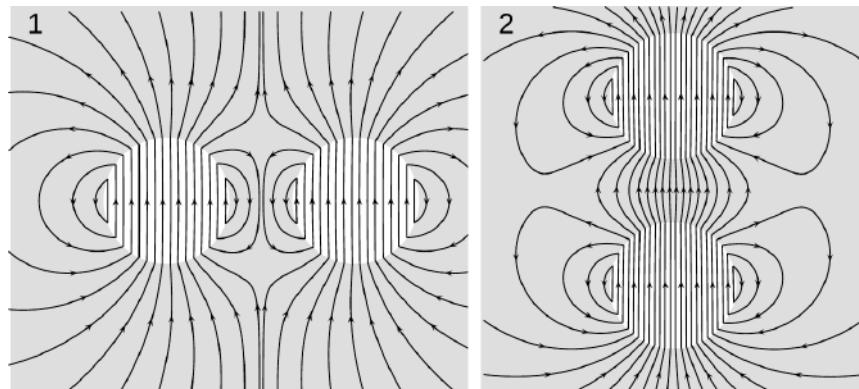
1 The figure depicts the electric fields in the radiation pattern of a certain type of radio antenna, shown in the photo superimposed in the background. Consider the small region of space indicated by the white square. As the waves pass through this area, spreading out like ripples from the antenna, 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. 126, 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?

2 (a) Figure 1 shows the magnetic field patterns of two steel ball bearings, each of which has been magnetized so that its interior field is uniform and points upward. Employ reasoning similar to that used on p. 127 in analyzing figure k to determine the direction of the forces that the spheres exert on each other.
(b) Do the same for figure 2, where the balls are oriented as before, but with their axes collinear.

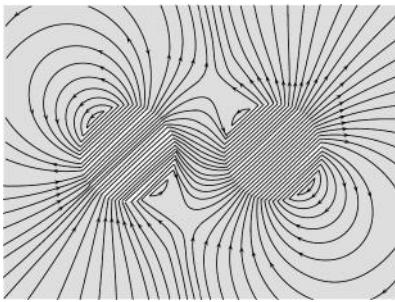
▷ Solution, p. 428



Problem 1.



Problem 2.



Problem 3.

3 The magnetized spheres in the figure are like the examples in problem 2, but tilted at 45 degrees, so that there is a lower level of symmetry. Use visual reasoning to find the approximate direction of the force that the left-hand magnet makes on the right-hand magnet. You will need the fact, discussed on p. 130, that the pressure and tension are equal in magnitude.

4 A particle with a charge of 1.0 C and a mass of 1.0 kg is observed moving past point P with a velocity $(1.0 \text{ m/s})\hat{x}$. The electric field at point P is $(1.0 \text{ V/m})\hat{y}$, and the magnetic field is $(2.0 \text{ T})\hat{y}$. Find the force experienced by the particle. ✓

5 Suppose a charged particle is moving through a region of space in which there is an electric field perpendicular to its velocity vector, and also a magnetic field perpendicular to both the particle's velocity vector and the electric field. Show that there will be one particular velocity at which the particle can be moving that results in a total force of zero on it; this requires that you analyze both the magnitudes and the directions of the forces compared to one another. Relate this velocity to the magnitudes of the electric and magnetic fields. (Such an arrangement, called a velocity filter, is one way of determining the speed of an unknown particle.)

6 A charged particle is released from rest. We see it start to move, and as it gets going, we notice that its path starts to curve. Can we tell whether this region of space has $\mathbf{E} \neq 0$, or $\mathbf{B} \neq 0$, or both? Assume that no other forces are present besides the possible electrical and magnetic ones, and that the fields, if they are present, are uniform.

7 A charged particle is in a region of space in which there is a uniform magnetic field $\mathbf{B} = B\hat{z}$. There is no electric field, and no other forces act on the particle. In each case, describe the future motion of the particle, given its initial velocity.

- (a) $\mathbf{v}_o = 0$
- (b) $\mathbf{v}_o = (1 \text{ m/s})\hat{z}$
- (c) $\mathbf{v}_o = (1 \text{ m/s})\hat{y}$

8 The following data give the results of two experiments in which charged particles were released from the same point in space, and the forces on them were measured:

$$\begin{aligned} q_1 &= 1 \mu\text{C}, & \mathbf{v}_1 &= (1 \text{ m/s})\hat{x}, & \mathbf{F}_1 &= (-1 \text{ mN})\hat{y} \\ q_2 &= -2 \mu\text{C}, & \mathbf{v}_2 &= (-1 \text{ m/s})\hat{x}, & \mathbf{F}_2 &= (-2 \text{ mN})\hat{y} \end{aligned}$$

The data are insufficient to determine the magnetic field vector; demonstrate this by giving two different magnetic field vectors, both of which are consistent with the data.

9 The following data give the results of two experiments in which charged particles were released from the same point in space, and the forces on them were measured:

$$\begin{aligned}q_1 &= 1 \text{ nC}, & \mathbf{v}_1 &= (1 \text{ m/s})\hat{\mathbf{z}}, & \mathbf{F}_1 &= (5 \text{ pN})\hat{\mathbf{x}} + (2 \text{ pN})\hat{\mathbf{y}} \\q_2 &= 1 \text{ nC}, & \mathbf{v}_2 &= (3 \text{ m/s})\hat{\mathbf{z}}, & \mathbf{F}_2 &= (10 \text{ pN})\hat{\mathbf{x}} + (4 \text{ pN})\hat{\mathbf{y}}\end{aligned}$$

Is there a nonzero electric field at this point? A nonzero magnetic field?

10 The following data give the results of three experiments in which charged particles were released from the same point in space, and the forces on them were measured:

$$\begin{aligned}q_1 &= 1 \text{ C}, & \mathbf{v}_1 &= 0, & \mathbf{F}_1 &= (1 \text{ N})\hat{\mathbf{y}} \\q_2 &= 1 \text{ C}, & \mathbf{v}_2 &= (1 \text{ m/s})\hat{\mathbf{x}}, & \mathbf{F}_2 &= (1 \text{ N})\hat{\mathbf{y}} \\q_3 &= 1 \text{ C}, & \mathbf{v}_3 &= (1 \text{ m/s})\hat{\mathbf{z}}, & \mathbf{F}_3 &= 0\end{aligned}$$

Determine the electric and magnetic fields. ✓

11 In problem 10, the three experiments gave enough information to determine both fields. Is it possible to design a procedure so that, using only two such experiments, we can always find \mathbf{E} and \mathbf{B} ? If so, design it. If not, why not?

12 A charged particle of mass m and charge q moves in a circle due to a uniform magnetic field of magnitude B , which points perpendicular to the plane of the circle.

(a) Assume the particle is positively charged. Make a sketch showing the direction of motion and the direction of the field, and show that the resulting force is in the right direction to produce circular motion.

(b) Find the radius, r , of the circle, in terms of m , q , v , and B . ✓

(c) Show that your result from part b has the right units.

(d) Discuss all four variables occurring on the right-hand side of your answer from part b. Do they make sense? For instance, what should happen to the radius when the magnetic field is made stronger? Does your equation behave this way?

(e) Restate your result so that it gives the particle's angular frequency, ω , in terms of the other variables, and show that v drops out. ✓

Remark: A charged particle can be accelerated in a circular device called a cyclotron, in which a magnetic field is what keeps them from going off straight. This frequency is therefore known as the cyclotron frequency. The particles are accelerated by other forces (electric forces), which are AC. As long as the electric field is operated at the correct cyclotron frequency for the type of particles being manipulated, it will stay in sync with the particles, giving them a shove in the right direction each time they pass by. The particles are speeding up, so this only works because the cyclotron frequency is independent of velocity.

13 (a) A line of charge with density λ is moving at velocity v in the direction parallel to its own length. Find the current.

(b) Suppose that there is an externally imposed magnetic field B perpendicular to the line of charge, which we can think of as a wire. Consider a portion of the wire of length ℓ , and show that the force per unit length is $F/\ell = IB$, as claimed on p. 131.



14 Compare the two dipole moments.

Problem 14.

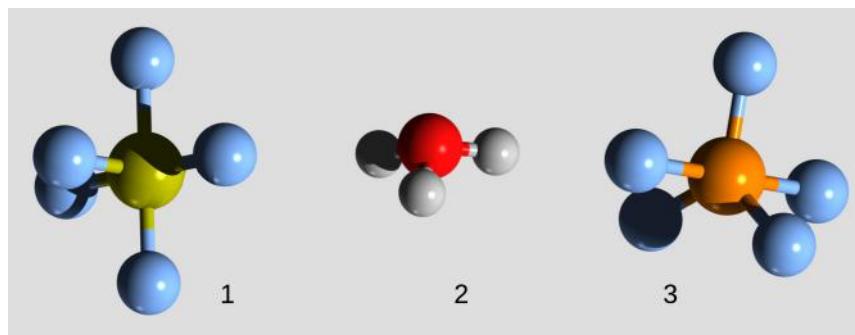
15 A dipole consists of two point charges lying on the x axis, a charge $-q$ at the origin, and a $+q$ at $x = \ell$. The dipole is immersed in an externally imposed, nonuniform electric field with $E_x = bx$, where b is a constant. Add the forces acting on the dipole. Verify that the total force depends only on the dipole moment, not on q or ℓ individually, and that the result is the same as the one found by a fancier method using a derivative on p. 136. \triangleright Solution, p. 428

16 An electric dipole is composed of charges $\pm q$ ($q > 0$), each with mass m , at the ends of a massless rod of length ℓ . The dipole is initially released in an orientation perpendicular to the ambient electric field of magnitude E . (a) Find its maximum angular velocity ω . \checkmark

- (b) Comment on the interpretation of the sign of your result.
(c) Show that the units of your answer make sense.
(d) Discuss how your answer depends on all four variables, and show that it makes sense. That is, for each variable, discuss what would happen to the result if you changed it while keeping the other two variables constant. Would a bigger value give a smaller result, or a bigger result? Once you've figured out this *mathematical* relationship, show that it makes sense *physically*.

17 An electron has a nonzero magnetic dipole moment, which is aligned with its spin. Suppose that at distance r , at a certain point in space relative to an electron, the magnetic field has magnitude B . Find the magnitude of the magnetic field at a point lying at distance $10r$, along the same line. ✓

18 The figure shows three molecules. You don't actually need to know any chemistry to do this problem, but in case you're interested, they're PCl_5^{2-} , H_3O^+ , and PF_5 , respectively. If you look carefully at the figure, you should be able to see that molecule 1 is a pyramid, 2 is not planar, and in 3 there are three atoms in a line plus three atoms arranged to form an equilateral triangle in a perpendicular plane. Use symmetry arguments to determine which of these molecules have zero electric dipole moments.



Problem 18.

- 19** (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:

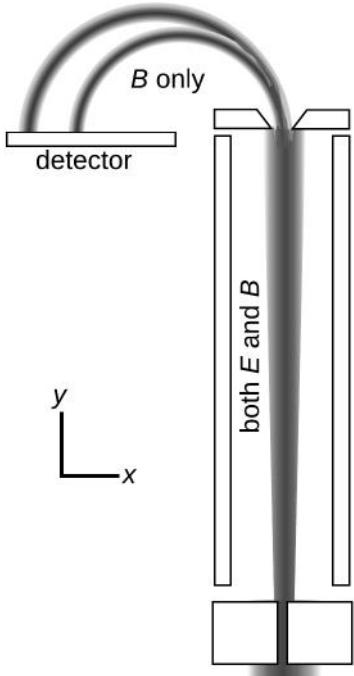
neutron	1.67495×10^{-27} kg
proton	1.67265×10^{-27} kg
electron	0.00091×10^{-27} kg
antineutrino	$< 10^{-35}$ kg

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!)



Problem 20.

- 20** The figure shows a simplified example of a device called a sector mass spectrometer. In an oven near the bottom, positively ionized atoms are produced. For simplicity, we assume that the atoms are all singly ionized. They may have different masses, however, and the goal is to separate them according to these masses. In the example shown in the figure, there are two different masses present. The reason this is called a “sector” mass spectrometer is that it contains two regions of uniform fields.

In the first sector, between the two long capacitor plates, there is an electric field E in the x direction. Superimposed on this is a uniform magnetic field B in the negative z direction (into the page). As analyzed in problem 5, these fields are chosen so that ions at a certain velocity v are not deflected. You will need the result of that problem in order to do this problem. Only the ions with the correct velocity make it out through the slits at the upper end of the capacitor.

In the second sector, at the top, there is no electric field, only a magnetic field, which we assume for simplicity to have the same magnitude and direction as in the first sector. This causes the beam to bend into a semicircular arc and hit a detector. In the first such spec-

trometers, this detector was simply some photographic film, whereas in modern ones it would probably be a silicon chip similar to the sensor of a camera.

The diameter h of the semicircle depends on the mass m of the ion. The quantity $\Delta h/\Delta m$ tells us how good the spectrometer is at separating similar masses.

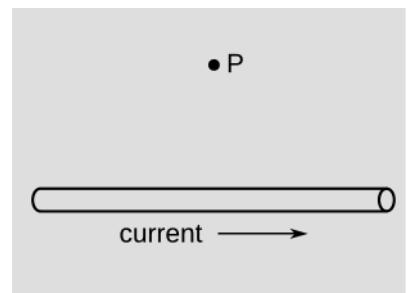
- (a) Express $\Delta h/\Delta m$ in terms of E , B , and e , eliminating v (which we can neither control nor measure directly). ✓
- (b) Show that the units of your answer make sense.
- (c) You will have found that increasing E makes the spectrometer more sensitive, while increasing B makes it less so. Explain physically why this is so. What stops us from getting an arbitrarily large sensitivity simply by making B small enough?

Remark: This design makes inefficient use of the ion source's intensity, because any ions with the wrong velocity are wasted. For this reason, real-world spectrometers of this type include complicated focusing elements.

- 21** (a) Show that the units of energy density are the same as the units of pressure.
(b) Verify, as claimed on p. 130, that if the only unitful quantities we have available are k and E , then the pressure and tension in the electric field must be proportional to E^2 .

- 22** Use the right-hand rule illustrated in figure f, p. 122, to find the direction of the magnetic field produced at point P by the current in the wire. ▷ Solution, p. 428

- 23** An electric dipole \mathbf{D} immersed in an electric field \mathbf{E} has energy $U = -\mathbf{D} \cdot \mathbf{E}$. Suppose that the magnitudes of both vectors are fixed, but the dipole is free to rotate. Find the orientation that would minimize the energy.



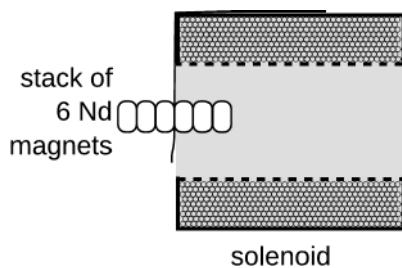
Problem 22.

Minilab 5: The dipole and superposition

Apparatus

6 Nd magnets
2 solenoids
1.5 V battery
battery holder
stopwatch
string
scissors
alligator clips
 $270\ \Omega$ resistor
compass

Goal: Measure the change in the period of oscillation of a magnetic dipole as the external field is doubled, and compare with theory.



a / The stack of neodymium magnets hangs by a thread in the end-plane of the solenoid. Cross-sectional side view.

The figure shows a cross-sectional side view of a solenoid. The central axis runs from left to right. The idea of this lab is to hang a permanent magnet, which acts as a magnetic dipole, from a string, precisely located in the end-plane of the solenoid. The oscillation of the magnet about the solenoid's axis is exactly mathematically equivalent to the motion of a pendulum, so the period T and the magnetic field B supplied by the solenoid are related by $T \propto B^{-1/2}$. If superposition holds, then the field can be precisely doubled by bringing in a second solenoid, oriented the same way, so that one of its end-planes coincides with the dipole as well.

In order to create the magnetic field of the

solenoid, a current is needed. When this current is established, and the magnetic field pops up, the field takes in some energy. This energy comes from a 1.5 V battery. In addition, once the circuit is hooked up, there is a continuous slow dissipation of energy to heat as the current flows through the copper wires, and the battery supplies this energy as well. The rate of heat dissipation is only about a milliwatt, so there is no danger that the battery will die. To make this work, you need to make a *complete circuit*, meaning that there is a closed loop of wire going from one terminal of the battery to one terminal of the solenoid, through the solenoid, back through another wire, and ending up at the other terminal of the battery.

In order to keep the current exactly constant, it is necessary to hook up both solenoids to the battery throughout the whole lab, not just when the second solenoid is needed. The cables are long enough so that it can be put far enough away (half a meter or so) so that its small external field will have no effect on the results when only the original solenoid's field is desired.

The solenoids should be wired in *series*, so that any charge that passes through one must also pass through the other. A series circuit is like a necklace with beads on it. The opposite of a series circuit is a parallel circuit, which has junctions giving the electrons choices about which way to go at certain points.

Furthermore, the oscillations of the permanent magnet are rather fast and rapidly damped unless the current is reduced a little. We will do this by adding a component called a resistor to the circuit. The resistor dissipates additional energy as heat.

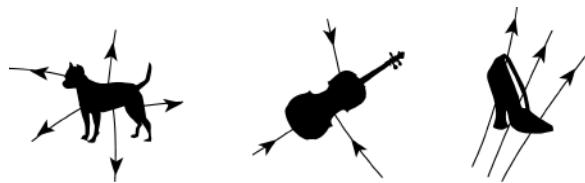
So, including all the parts, your series circuit will look like a necklace with the following "beads" around its circumference: battery, resistor, solenoid, solenoid, and back to the battery again. The order of the circuit elements in the loop actually has no effect on the results; because charge is conserved, it is not possible for charge to get lost or used up on its way

around the circuit.

A complication comes from the ambient field in the room, which is a combination of the earth's field and the field of the building materials. This field is a fraction of that of the solenoids, but still significant enough to affect the results. To deal with this, we will intentionally align the solenoids with the horizontal component of the ambient field, so that the vector addition of the horizontal fields is equivalent to addition of real numbers. Then we will take data with the current flowing in both directions, so that in one case the ambient field adds to the field of the solenoid(s) and in the other it subtracts. Averaging the periods in these two conditions eliminates the effect of the ambient field.

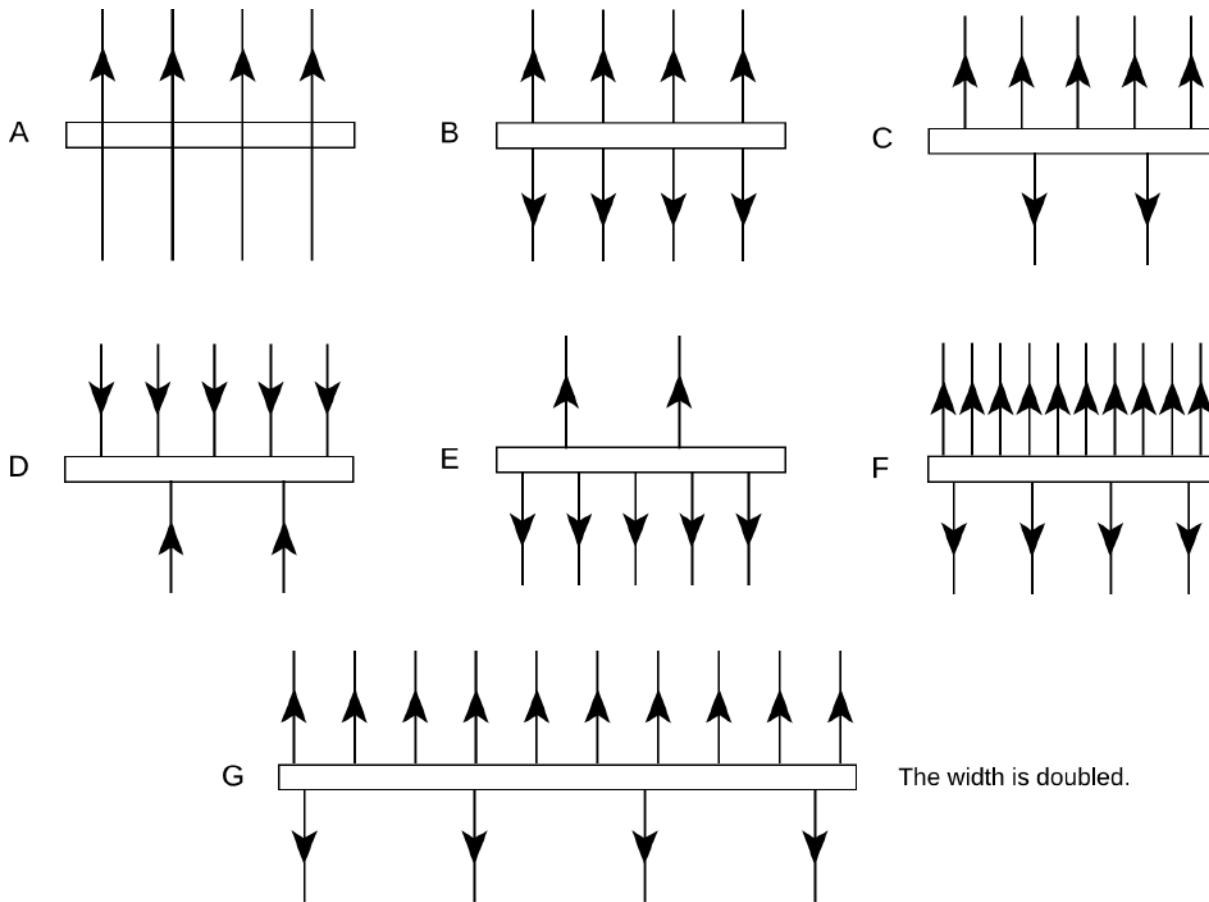
Exercise 5: Tension in the electric field

1. Warm-up: Use Gauss's law to find the charges on the objects.



2. Suppose that a metal plate is immersed in a uniform external electric field E_{ext} . This field superposes with any field due to charge q on the plate itself. The force acting on the plate will depend on the product qE_{ext} .

For the following examples, compute the forces, including a sign to indicate their direction, positive being up. Use units such that the density of field lines in A is an electric field of +4.



3. Can you find a way to express these results in a simpler form, or make a rule that explains the results?



Chapter 6

Radiation

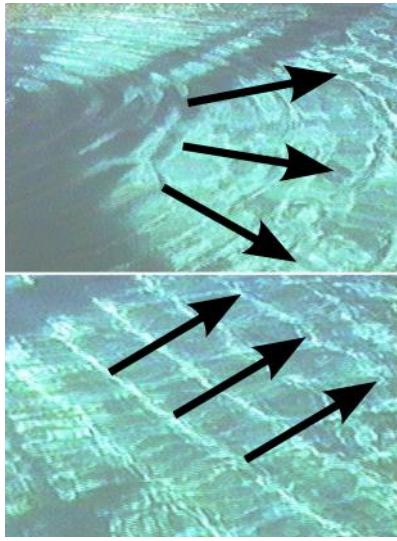
Way back at the beginning of this book, on the second page of the very first chapter, we proved, based only on experiments demonstrating the relative nature of time, that the universe could not operate as envisioned in Newton's picture of instantaneous action at a distance. Electromagnetic forces must propagate as wave disturbances in the fields. Later we saw that the only sensible¹ way of finding the direction of propagation of such a wave was the vector cross product $\mathbf{E} \times \mathbf{B}$, and argued (p. 126) that therefore these wave disturbances must contain oscillations of both the electric and magnetic field. In this chapter we will take up the description of these electromagnetic waves in more detail.

By the way, although I'm offering answers to "why" questions about these waves, it's important to understand that in mathematics and the hard sciences, a "why" question can have more than one answer. It's different in other fields such as religion. If you ask a theologically well-informed Christian why humans have to endure

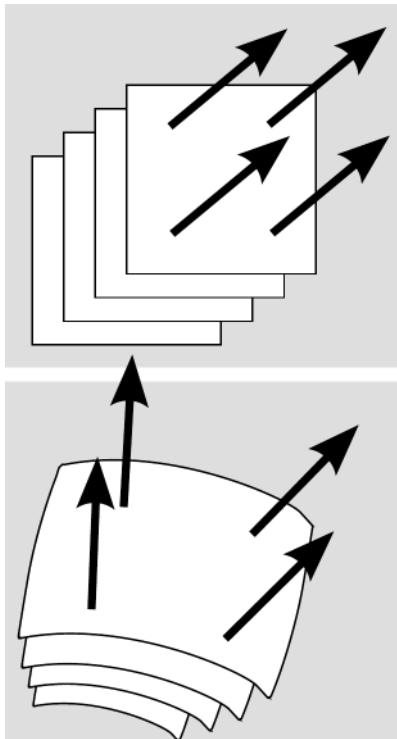


a / "Why?"

¹i.e., rotationally invariant



b / Circular and linear wave patterns.



c / Plane and spherical wave patterns.

pain and suffering, there is one definite answer, which is that it's a punishment for Eve's original sin of obtaining an unauthorized education from a serpent. Done. Explained. But in math and science, we have the problem of how to prove things rigorously without ever being allowed to appeal to some supernatural authority for basic principles to use as starting points. We do this by explicitly stating some set of assumptions, and making it clear that all our later reasoning holds only *conditionally*, if those assumptions are true.

A familiar example is the question of why the distance between two points in space is given by $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. There is more than one answer to this "why" question. One is that if we accept Euclid's postulates, then it follows as a theorem, the Pythagorean theorem. But in the Cartesian approach, this formula for the distance is taken as an assumption, and some of Euclid's postulates become theorems that can be proved. These are two totally different answers to the same "why" question, and they are both valid.

The point of this digression, in the context of our study of electromagnetism, is that if you look in different books, you will find different logical developments of the subject, starting from different sets of assumptions. In this book, we start from some of the assumptions described above, derive facts about electromagnetic waves, and then end up using those facts to infer the full set of physical laws for electromagnetism, known as Maxwell's equations (summarized on p. 444). Other books might take Maxwell's equations as assumptions (perhaps with some experimental or logical justification) and use them to prove the properties of electromagnetic waves. By p. 264, we'll finish making all the connections, and it will become clear there is a system of tightly interlocking logical relationships, very much like the Euclidean/Cartesian example.

6.1 Wave patterns

Waves spread out in all directions from every point on the disturbance that created them, and this is why we refer to electromagnetic "radiation" — the wave radiates out in all directions.

If such a wave 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, b/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, b/2. For a three-dimensional wave such as a sound wave, the analogous patterns would be spherical waves and plane waves, c.

All kinds of wave 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.

Because of the geometrical relationship of the momentum density to $\mathbf{E} \times \mathbf{B}$, we need a full three dimensions in order to describe an electromagnetic wave, so there is no hope of treating anything as effectively two-dimensional, and the simplest wave pattern is therefore the plane wave.

For convenience throughout this chapter, we will always take our waves to be propagating along the z axis. The definition of a plane wave then becomes very simple: it is an electromagnetic field pattern in which the fields are both functions only of z and t , $\mathbf{E} = \mathbf{E}(t, z)$ and $\mathbf{B} = \mathbf{B}(t, z)$.

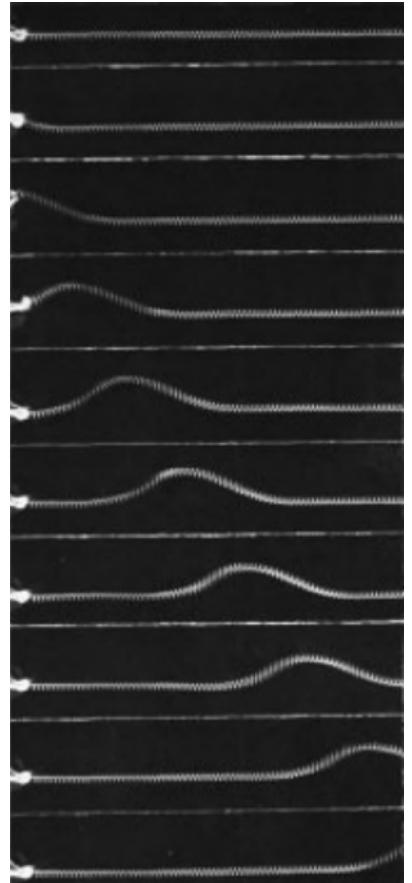
6.2 What it is that waves

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 a sound wave or the waves on a coil spring in figure d, is not a mechanical wave, i.e., it isn't a vibration of any material medium such as the air or the spring. What vibrates is the fields — invisible, intangible, and massless. We will see that electromagnetic waves are transverse, i.e., they vibrate from side to side (like the waves on a spring) rather than along the direction of propagation (like a sound wave). But because they are not vibrations of a material medium, this vibration doesn't mean that anything is actually *traveling* from side to side. A bug sitting on the spring in figure d 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



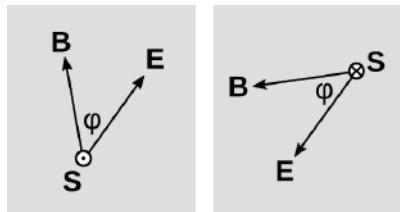
d / A force acts on a coil spring at the left end, producing a mechanical wave pulse that propagates to the right.

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."

6.3 Geometry of a plane wave

6.3.1 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.



e / Are these possible fields for electromagnetic plane waves?

6.3.2 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 $U_{\mathbf{E}} + U_{\mathbf{B}}$, for a total energy of $2U_{\mathbf{E}} + 2U_{\mathbf{B}}$. 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 $U_{\mathbf{B}'}$, 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 $U_{\mathbf{E}'}$, 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 $U_{\mathbf{E}} + U_{\mathbf{B}} = U_{\mathbf{B}'} = U_{\mathbf{E}'}$, 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 $E = cB$. (This was problem 12, p. 34). With a couple of centuries of hindsight, it would have been better if we had con-

structed 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.”

6.3.3 \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

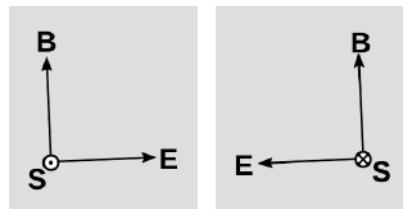
example 1

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.

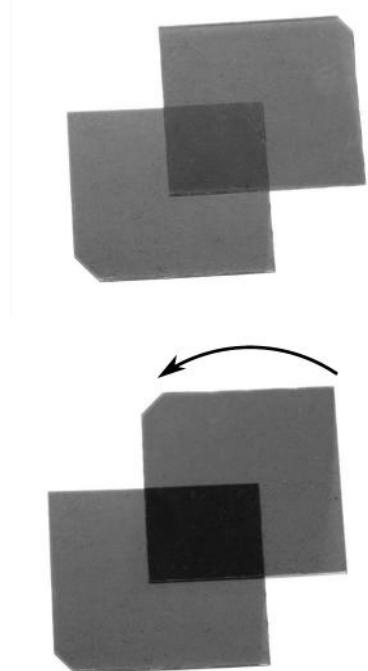
At the first filter, light with the “right” orientation gets through, while light with the “wrong” orientation is blocked. Of course, a randomly chosen angle will not be at exactly 0° or 90° . A wave with an intermediate angle of polarization can be broken down into *components* (say the components of \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.

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.



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



g / Example 1.

Discussion question

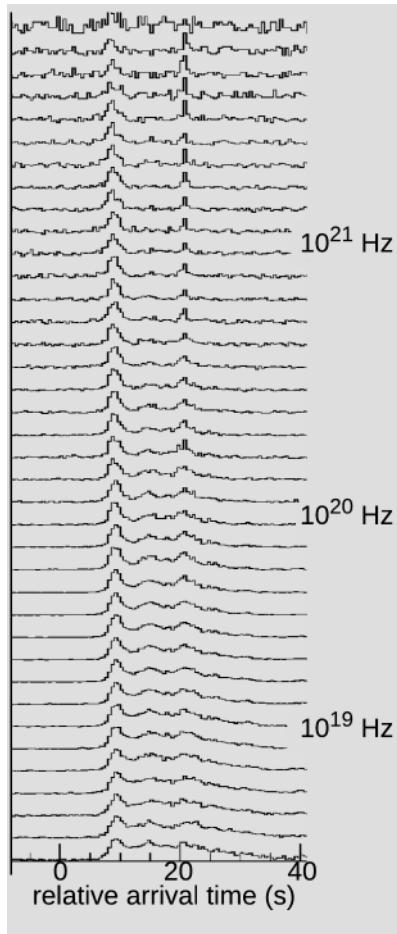
A Suppose someone tells you that a beam of light consists of a stream of electrons moving through space. Use the experiment in example 1 to convince them that they're wrong.

6.4 Propagation at a fixed velocity

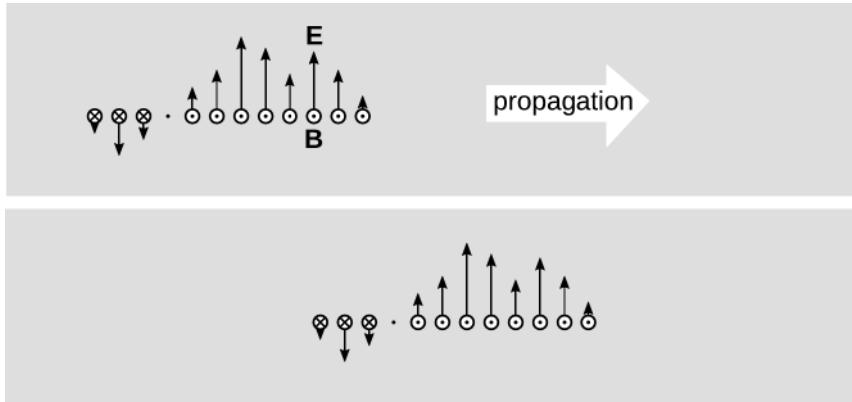
If we look at the speed of waves in general, we see two main types of behavior. In one type, exemplified by sound waves or waves on a coil spring as in figure d, all waves travel at the same speed, regardless of their amplitude or frequency. Water waves are a good example of another type, called a dispersive 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 pulse on the coil spring in figure d is propagating nondispersively, because it maintains its shape as it moves.

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. 124). 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 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, which we will show on p. 178 is c . They are nondispersive, so a plane wave glides along rigidly as in figure i,



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).

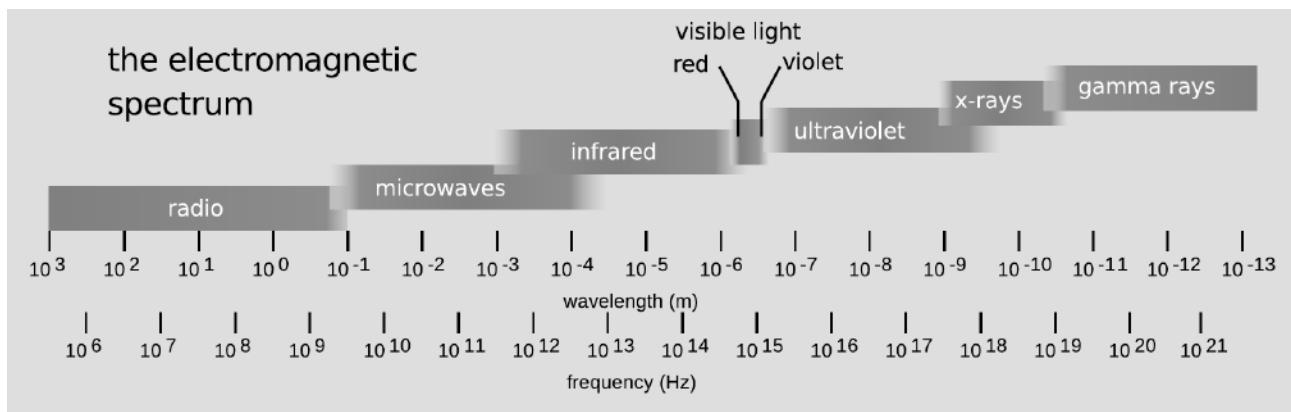
without changing shape.

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.

6.5 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 (sec. 6.7), 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 λ (Greek letter lambda, which makes the “L” sound). On a sinusoidal wave, the wavelength is the distance in space between one crest and the next. These are not two independent parameters. During one cycle of vibration, which takes time $1/f$, one wavelength λ travels past a fixed point in space. We therefore have $c = (\text{distance})/(\text{time}) = \lambda f$. A higher frequency corresponds to a shorter 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!)

6.6 Momentum and rate of energy flow

6.6.1 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.

6.6.2 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

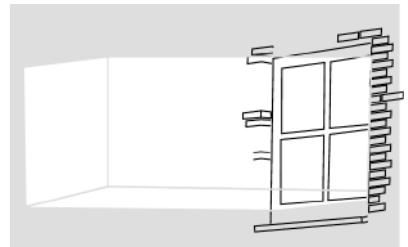
$$\frac{\text{power}}{\text{area}} = \frac{\text{momentum}}{\text{volume}} \times c^2.$$

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.

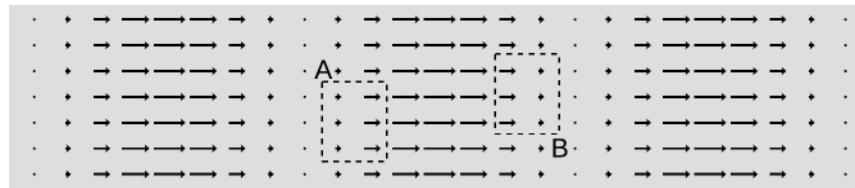
The quantity $(c^2/4\pi k)\mathbf{E} \times \mathbf{B}$, which is just the momentum density multiplied by c^2 , is often notated \mathbf{S} , and is referred to as the Poynting vector, after John Henry Poynting — a wonderful coincidence, because the vector *points* in the direction of the momentum and 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 2. There is a net flow of energy out of region A, and a net flow into B.



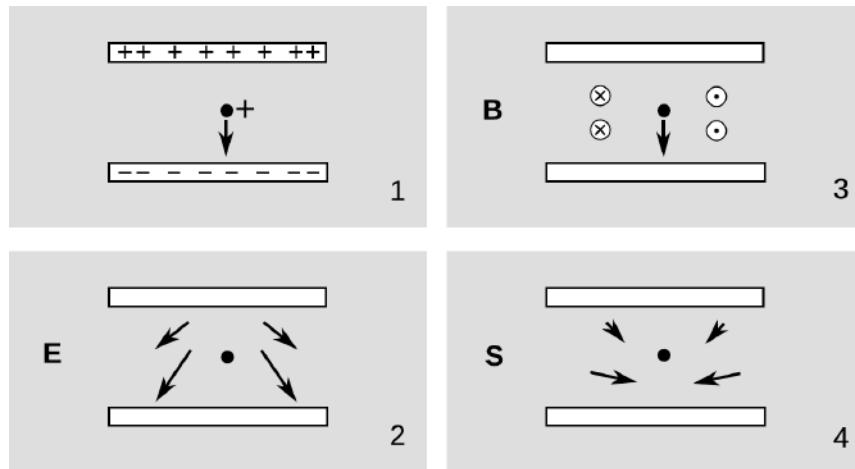
Poynting vector of a plane wave

example 2

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 2, 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 ([2167](#)).

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 3

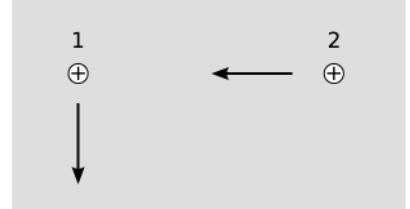
In discussion question A, p. 52, we convinced ourselves that if a charge was released inside a capacitor, the kinetic energy it gained could be properly accounted for by the energy lost from the electric field. We also calculated this quantitatively in note [264](#). Figure I shows how this plays out 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 3 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 ([Z167](#)).

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 illustrated in figure f, p. 122.) What does this imply for conservation of momentum?



Discussion question A.

6.7 Maxwell's equations in a vacuum

The laws of physics are local, so if we zoom in on an electromagnetic wave, as in figure m, and take on the role of the bug, we must be able to make sense of the evolution of the wave pattern as time goes on. The structure of the laws of physics for electromagnetism is that they talk about the curl and divergence of the fields. It's not arbitrary that they have this structure, because the curl and divergence are the only rotationally invariant derivative operators we can define that take a vector-valued field as an input.

What the bug observes at this moment in time is that the electric field has a curl that points into the page. (You should verify this yourself by visualization, using the right-hand rule, if necessary referring to the drawing of the curl-meter in figure b on p. 86.) She also observes that the magnetic field's curl is up. For static fields, we're supposed to have $\text{curl } \mathbf{E} = 0$ and $\text{curl } \mathbf{B} = 0$, so this would be impossible. But this is not a static field, it's a field that is changing over time. The bug needs some laws of physics that relate the time-varying nature of the fields to their curls. With this motivation, we state Maxwell's equations² in a vacuum:

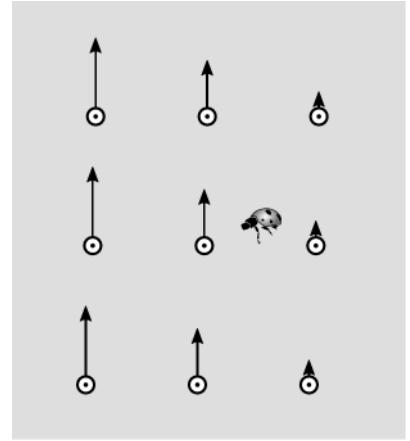
$$\text{div } \mathbf{E} = 0$$

$$\text{div } \mathbf{B} = 0$$

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\text{curl } \mathbf{B} = \frac{1}{c^2} \frac{\partial \mathbf{E}}{\partial t}.$$

There is not much else we could write down that would satisfy the bug's requirements. The signs are the ones that make the geometry of figure m work.³ The factor of $1/c^2$ has to be there because of



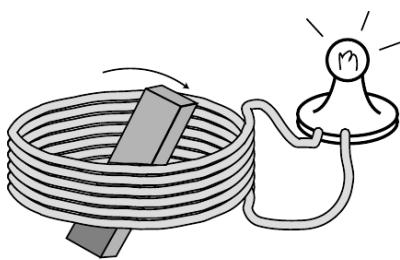
m / A bug makes local observations of the leading edge of the wave from figure i. The wave is moving to the right, and the magnetic field is perpendicular to the page.

²The full version of Maxwell's equations, including matter, are summarized on p. 444.

³The fact that the two signs are opposite is what gives us solutions that are

units. We could throw in other unitless factors and change the $1/c^2$ to $7/c^2$, for example, but $1/c^2$ is the choice that turns out to make the wave propagate at c ([2167](#)).

The physics embodied in these two new terms of Maxwell's equations is referred to as induction (a different usage of the term than the one introduced for static fields on p. 135). A changing magnetic field induces a curly electric field, and a changing electric field induces a curly magnetic field. This has many important technological applications, but since those are not our topic right now, we give only two qualitative examples. Further quantitative material is presented in ch. 15.



n / A generator.

The generator

example 4

A generator, n, consists of a permanent magnet that rotates within a coil of wire. The magnet is turned by a motor or crank, (not shown). As it spins, the nearby magnetic field changes. This changing magnetic field results in an electric field, which has a curly pattern. This electric field pattern creates a current that whips around the coils of wire, and we can tap this current to light the lightbulb.

Radiation from a dipole antenna

example 5

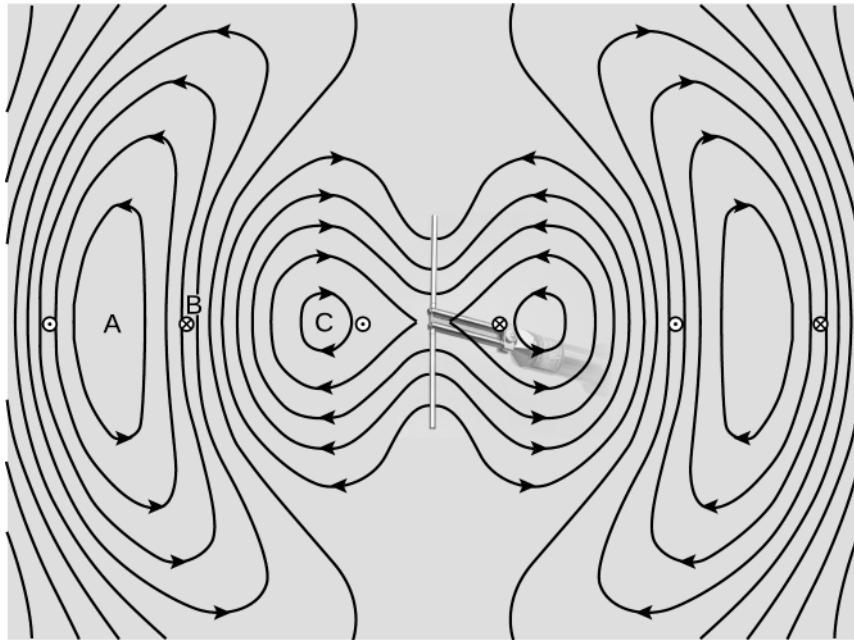
Figure o depicts the electric fields in the radiation pattern of an antenna, shown in the photo superimposed in the background. This type of antenna is called a dipole antenna, meaning that as a source, it acts like an electric dipole whose dipole vector oscillates sinusoidally. The magnetic field is perpendicular to the page and is shown only at six arbitrarily chosen points in the midplane. Let's verify that this pattern appears to satisfy Maxwell's equations in a vacuum.

The electric field lines never begin or end at any point in the air surrounding on the antenna, only at points on the antenna, which means that they satisfy $\text{div } \mathbf{E} = 0$ everywhere that they should. (And where they reach the antenna, which is a conductor, they are perpendicular to the surface.)

Although the magnetic field lines are not shown, they are simply circles centered on the dipole axis and perpendicular to it. Each magnetic field vector shown on the left wraps around and connects to the symmetrically placed field vector on the right. Because these circles close on themselves, there are no magnetic sources or sinks, as required by $\text{div } \mathbf{B} = 0$.

In order to check $\text{curl } \mathbf{E} = -\partial \mathbf{B} / \partial t$, we pick points A, B, and C, of which only B coincides with a point where the magnetic field was drawn. Here is a table, to be justified afterward, describing the

oscillating waves, rather than exponential freak-outs. We could make the signs + and - rather than - and +, but this would be equivalent to redefining the magnetic field like $\mathbf{B} \rightarrow -\mathbf{B}$.



o / The radiation pattern emitted by a dipole antenna, example 5.

behavior of the fields and their derivatives at these points.

	\mathbf{E}	$\text{curl } \mathbf{E}$	\mathbf{B}	$\partial\mathbf{B}/\partial t$
A	0	max (out)	0	max (in)
B	max (up)	0	max (in)	0
C	0	max (in)	0	max (out)

Points A, B, and C are chosen to be ones at which the electric field is at a zero, a maximum, and a minimum, hence the values in the first column. We can tell, for example, that \mathbf{E} is small at A because the field lines are far apart, and we know that $\mathbf{E} = 0$ must occur somewhere near A because as the field pattern has been spreading outward, the field has been changing from down to (soon) up.

The second column comes from visualizing a curl-meter inserted in the figure. The curl is nearly zero at B because the curl is flipping directions near there at this time.

The third column comes from our expectation that the Poynting vector should point outward, which requires that \mathbf{E} and \mathbf{B} reach their maxima at similar times, i.e., be at least approximately in phase, as in example 2 on p. 162 (Poynting vector of a plane wave).⁴

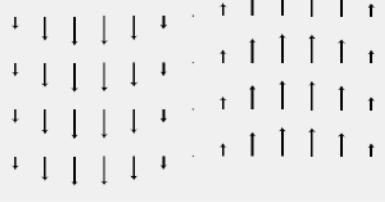
The fourth column follows from the fact that the time variation of the wave is sinusoidal, and the derivative of a sine wave is also a

⁴This expectation cannot be valid everywhere on the figure. Close up to the antenna, we can find places where the electric field points directly inward or outward, so that the Poynting vector cannot be straight outward. However, far from the antenna the pattern begins to look locally more and more like a plane wave, so this kind of thing can't happen.

sine wave, shifted 90 degrees forward in phase (e.g., $\sin' = \cos$).

We can now verify that these estimates are consistent with $\text{curl } \mathbf{E} = -\partial \mathbf{B} / \partial t$ since the second and fourth columns match up except for the flipped direction arising from the minus sign.

Discussion question



A The diagram shows an electric field pattern frozen at one moment in time. Let's imagine that it's the electric part of an electromagnetic wave. Consider four possible directions in which it could be propagating: left, right, up, and down. Determine whether each of these is consistent with Maxwell's equations. If so, infer the direction of the magnetic field.

Discussion question A.

6.8 Einstein's motorcycle

It seems as though we have a nice, tidy picture of electromagnetism all wrapped up. There should only be a few loose ends to tie up, such as how to incorporate charges and currents into Maxwell's equations, and how to connect Maxwell's equations to some practical problem-solving, such as the computation of the field of a solenoid.

But 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 material similar to this, 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? We take up this train of thought again in ch. 7.

Notes for chapter 6

≥162 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.

≥163 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. 52, in example 3, p. 162, and in note [≥64](#). In the quantitative analysis of note [≥64](#), 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. 162, 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.

≥164 Maxwell's equations and propagation at c

Maxwell's equations give electromagnetic waves that propagate at c.

As in the main text, we take a wave of the form

$$E_x = f(z - vt)$$
$$B_y = (1/c)f(z - vt).$$

This is the most general form of a plane wave, with any shape defined by the function f , propagating in the positive z direction at velocity v . To see this consider what happens if we want to increase t by Δt while keeping the input to the function f the same: we must increase z by $v\Delta t$.

The divergence vanishes, as required by Maxwell's equations, since neither component has a nonvanishing partial derivative with respect to x or y .

We now need to evaluate the curl, which we do by using the componentwise expressions from

note [Z101](#).

$$(\text{curl } \mathbf{E})_y = \frac{\partial E_x}{\partial z} = f'$$
$$(\text{curl } \mathbf{B})_x = \frac{\partial B_y}{\partial z} = -(1/c)f'$$

Maxwell's equations require these to equal the time derivatives

$$-\frac{\partial B_y}{\partial t} = (v/c)f' \quad \text{and}$$
$$\frac{1}{c^2} \frac{\partial E_x}{\partial t} = -(v/c^2)f'.$$

Equating these to each other as required by the two curl equations in Maxwell's equations, we find $1 = v/c$ and $-1/c = -v/c^2$. These two equations are both satisfied if and only if $v = c$. (To get propagation with $v = -c$ we would have had to rearrange things so as to create a Poynting vector in that direction.)

Problems

Key

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

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

2 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. ✓

3 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.

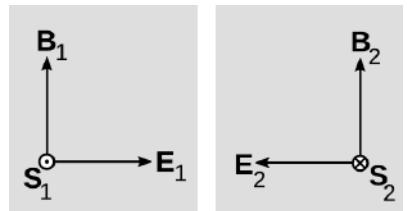
(c) Show that the results from the two methods are consistent with each other, and give a physical interpretation.

▷ Solution, p. 428

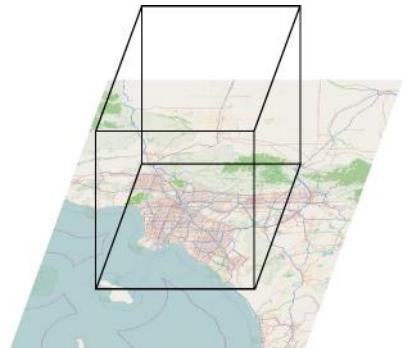
4 The solar constant is defined as the average flux of electromagnetic energy coming from the sun to the earth, per square meter. It equals 1.36 kW/m^2 . Now imagine, as suggested by the figure, a giant cubical volume of space, 100 km on a side. Find the total momentum of the electromagnetic energy inside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

✓

5 A plane electromagnetic wave is supposed to have its electric and magnetic fields perpendicular to each other. Suppose someone claims they can make an electromagnetic wave in which the electric and magnetic fields are perpendicular to the direction of propagation, but parallel to each other. We have already ruled out such a possibility on p. 157 based on conservation of energy. Give a different proof by using Maxwell's equations. ▷ Solution, p. 428



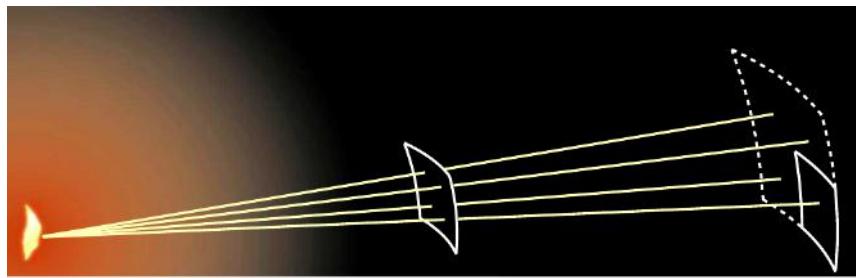
Problem 3.



Problem 4.

6 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. 428



Problem 6.

7 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. ★

Exercise 6A: Polarization

Apparatus

calcite crystal
polarizing film

1. Lay the crystal on a piece of paper that has print on it. You will observe a double image. See what happens if you rotate the crystal.

Evidently the crystal does something to the light that passes through it on the way from the page to your eye. One beam of light enters the crystal from underneath, but two emerge from the top; by conservation of energy the energy of the original beam must be shared between them. Consider the following three possible interpretations of what you have observed:

- (a) The two new beams differ from each other, and from the original beam, only in energy. Their other properties are the same.
- (b) The crystal adds to the light some mysterious new property (not energy), which comes in two flavors, X and Y. Ordinary light doesn't have any of either. One beam that emerges from the crystal has some X added to it, and the other beam has Y.
- (c) There is some mysterious new property that is possessed by all light. It comes in two flavors, X and Y, and most ordinary light sources make an equal mixture of type X and type Y light. The original beam is an even mixture of both types, and this mixture is then split up by the crystal into the two purified forms.

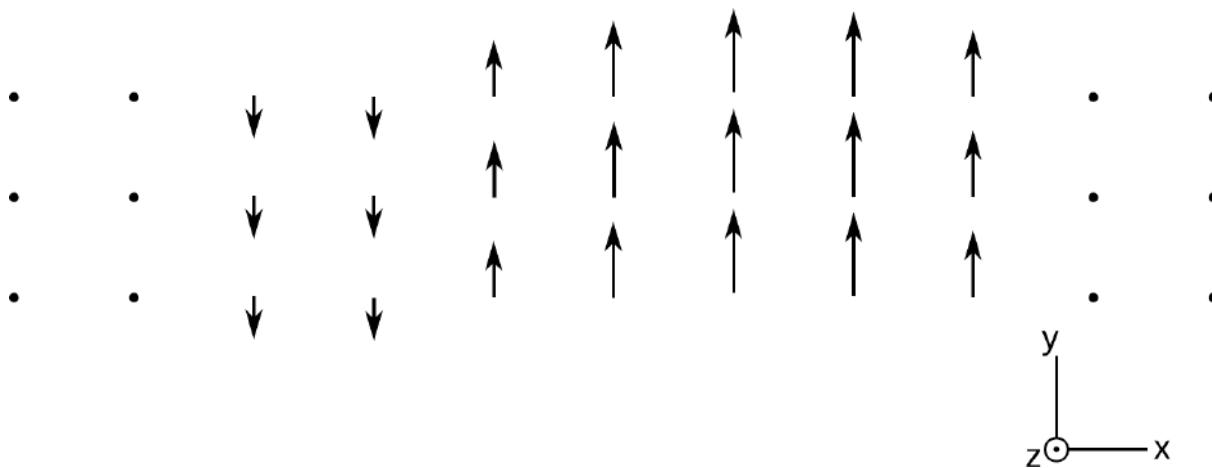
In parts 2 and 3 you'll make observations that will allow you to figure out which of these is correct.

2. Now place a polaroid film over the crystal and see what you observe. What happens when you rotate the film in the horizontal plane? Does this observation allow you to rule out any of the three interpretations?
3. Now put the polaroid film under the crystal and try the same thing. Putting together all your observations, which interpretation do you

think is correct?

4. Look at an overhead light fixture through the polaroid, and try rotating it. What do you observe? What does this tell you about the light emitted by the lightbulb?
5. Now position yourself so that you can observe a glancing reflection of a light source from a shiny surface, such as a glossy tabletop. You'll see the lamp's reflection. Observe this reflection through the polaroid, and try rotating it.

Exercise 6B: Maxwell's equations applied to a plane wave



1. The figure shows a sea-of-arrows representation of a vector field, along with a coordinate system. Suppose that this is the electric field of a plane wave propagating in the positive x direction. Determine the magnetic field.
2. As in discussion question A, p. 88, figure out the direction and approximate relative strength of $\text{curl } \mathbf{E}$ at various points along the x axis. Describe this in terms such as “small $+y$.” Do the same thing for the other functions.

$\text{curl } \mathbf{E}$ _____

$\text{curl } \mathbf{B}$ _____

$\partial \mathbf{E} / \partial t$ _____

$\partial \mathbf{B} / \partial t$ _____

3. Verify qualitatively that

$$\begin{aligned}\text{curl } \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \quad \text{and} \\ \text{curl } \mathbf{B} &= \frac{1}{c^2} \frac{\partial \mathbf{E}}{\partial t}.\end{aligned}$$

Minilab 6: A polarizing filter

Apparatus

- sodium gas discharge tube
- 2 polarizing films
- photovoltaic cell and collimator
- protractor

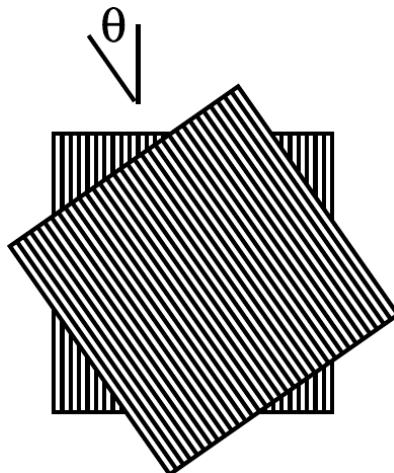
Goal: Test quantitatively the hypothesis that polarization relates to the direction of the field vectors in an electromagnetic wave.

Turn on the sodium gas discharge tube immediately so that it starts warming up. This tube acts as a lamp emitting bright light in little isolated pulses, each of which has a randomly oriented polarization.

The polarizing films are essentially transparent pieces of plastic with microscopic parallel lines on them. If you hold one of the polarizing films up to your eye and look around, you will see that in general it reduces the brightness of what you see. But if you make more detailed observation, you will see that more is going on. For some light sources, such as the sodium discharge tube, the orientation of the film makes no difference, but for other sources the orientation does have an effect. Two good examples are the light from a cell phone's LCD screen and light that has undergone a glancing reflection from a glossy tabletop. These are polarized sources of light, and the film preferentially transmits different polarizations.

It would not make sense for the film simply to throw away any waves that were not perfectly aligned with it, because a field oriented on a slant can be analyzed into two vector components, at 0° and 90° with respect to the film. Even if one component is entirely absorbed, the other component should still be transmitted.

Based on these considerations, now think about what will happen if you look through two polarizing films at an angle to each other, as shown in the figure above. What will happen as you change the angle θ ?



a / Two polarizing films oriented at an angle θ relative to each other.

The idea of this lab is to make numerical measurements of the transmission of initially unpolarized light transmitted through two polarizing films at an angle θ to each other. To measure the intensity of the light that gets through, you will use a photocell, which is a device that converts light energy into a potential difference.

You will use a voltmeter to measure the potential across the photocell when light is shining on it. A photovoltaic cell is a complicated nonlinear device, but I've found empirically that under the conditions we're using in this experiment, the potential is proportional to the power of the light striking the cell: twice as much light results in twice the potential.

This measurement requires a source of light that is unpolarized, constant in intensity, and comes from a specific direction so it can't get to the photocell without going through the polaroids. The ambient light in the room is nearly unpolarized, but varies randomly as people walk in front of the light fixtures, etc. A suitable source of light is the sodium gas discharge tube. Make sure you have allowed it to warm up for at least 15-20 minutes before using it; before it warms up, it makes a reddish light, and the polarizing films do not work very well on that color.

Each group will be assigned two angles θ .

Measure the power $P(\theta)$ of the light transmitted through the two polarizing films, and also measure $P(0)$ for comparison, and tabulate your result for $P(\theta)/P(0)$ on the board. Don't assume that the notches on the plastic housing of the polarizing films are a good indication of the orientation of the films themselves.

Calculate a theoretical value for $P(\theta)/P(0)$ based on a model in which the polaroid passes only the component of the field vectors along a certain axis.

Use a graph to compare the results with theory. On this type of scientific graph, the standard style is to show experiment as dots, and theory as a line or curve (because theory is a function that exists at all values of the independent variable, not just the ones where you have data).

Chapter 7

★More about relativity (optional)

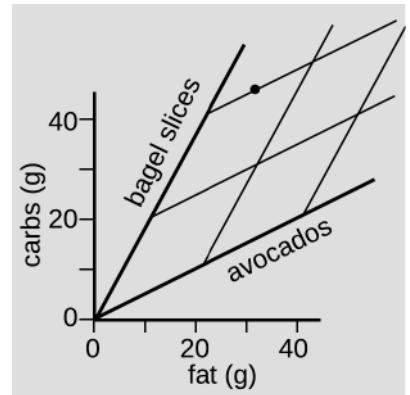
7.1 Einstein's motorcycle: the resolution

When we left our youthful protagonist at the end of chapter 6, he was asking himself what he would observe if he rode a motorcycle inside an electromagnetic wave, chasing the wave at nearly the speed of light. And what if he rode *at* the speed of light?

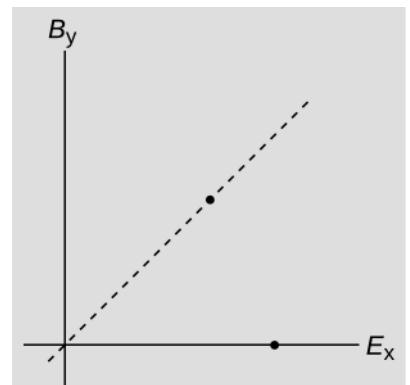
It would seem that in the case where he rode at c , he should be at rest with respect to the wave, so that in his frame of reference the fields don't change with time. But this violates Maxwell's equations. With the benefit of hindsight, let's untangle this paradox that Einstein only finally resolved five years later, with the publication of his special theory of relativity.

We'll approach this by examining how the mix of electric and magnetic fields in one frame of reference compares with what is measured by an observer in a different frame of reference. In case this is all a little too abstract, let's start with some practice in visualizing this kind of thing using a graph. An avocado has about twice as many grams of fat as carbohydrates, whereas if you eat a half-bagel slice with a thin shmear of cream cheese, the ratio is the other way around. Figure a shows a way of converting back and forth between (avocado, bagel) coordinates and (fat, carbs). The dot represents the two halves of a bagel, with half an avocado split between them to make avocado toast. The result is that we get about 30 g of fat and 45 g of carbs.

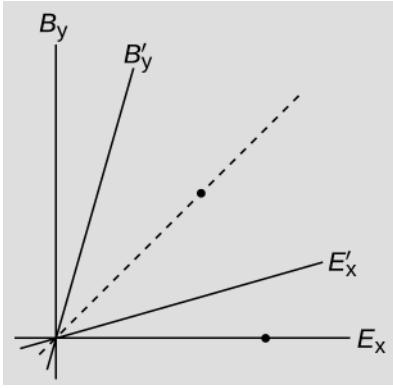
Now let's get back to Einstein's motorcycle paradox. The field that we observe, as we move in the z direction with velocity v , has nonzero components E_x and B_y . On a graph with E_x and cB_y on its axes, figure b, a particular spot in the field pattern of an electromagnetic wave is represented by a point on the diagonal that runs through the origin with a slope of 1. (The factors of c that occur in this discussion are only because of the inconvenience of the SI. Ignore them as you read. I'll say things like "the E field and the B field are equal," with the understanding that this is really only true if the factors of c are filled in.) A pure electric field is a point on the horizontal axis, and pure magnetic field is a point on the vertical axis.



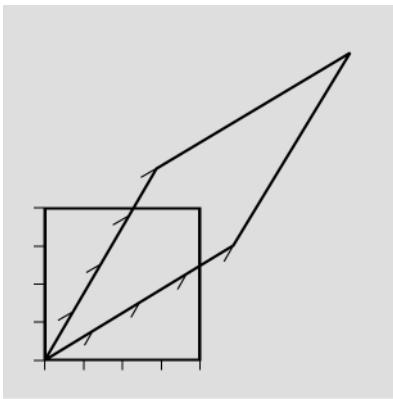
a / A graphical method for visualizing the conversion between (avocado, bagel) and (fat, carbs) coordinates.



b / Two electromagnetic field configurations represented as dots on a graph of B_y versus E_x . (To avoid clutter, factors of c are suppressed; the vertical axis should really be cB_y in SI units.)



c / Changing between two frames of reference.



d / How a square on the (E_x, B_y) graph paper changes as we change frames of reference. The area stays the same.

We already know that if we change our frame of reference, a particular combination of electric and magnetic fields looks like some other set of fields. We know that these transformations are linear, and we also know from examples like figure c on p. 121, involving a line of charges, what happens to a point on the horizontal axis. We could represent this kind of information by drawing pictures of how the points move around on a fixed graph-paper grid, but it turns out to be much easier to keep the points fixed and move the axes, as in figure c. The slope of the new E'_x axis turns out to be v/c , where v is the velocity of the new frame relative to the old one as it moves in the z direction. We'll justify this claim in more detail later, but for now we note that clearly this makes sense when $v = 0$, since then there is no change in our frame of reference, and the axis shouldn't tilt at all.

For practice, let's interpret the two points in figure c. The bottom point is a purely electric field in the original frame of reference. For an observer moving in the positive z direction at velocity v , we use the (E'_x, B'_y) axes to determine the fields. Although we haven't yet marked in tick marks, so we can't tell whether E'_x is bigger or smaller than E_x , we can see that the dot is below the new "horizontal" axis, so $B'_y < 0$, and this is consistent with our earlier analysis of figure c on p. 121.

The other point lies on the diagonal, so it's a combination of fields that we could have in a plane wave propagating in the z direction. An electromagnetic wave that obeys the laws of physics in one frame should also obey the laws of physics in another frame, and therefore it makes sense that in the tilted axes representing (E'_x, B'_y) , it again lies on the diagonal. To make this part of the example come out right, we are forced to make the two axes tilt by equal angles, like the blades of a pair of scissors as they close in from both sides.

We still haven't specified how much the axes stretch or shrink, i.e., how far apart the tick marks should be. It's fairly straightforward to prove that the area of a square on the original graph paper must keep the same area as we distort it into a parallelogram on the new graph paper. The form of the transformation is now completely determined, and algebra shows it to be

$$E'_x = \gamma E_x - \frac{v}{c} \gamma c B_y$$

$$c B'_y = -\frac{v}{c} \gamma E_x + \gamma c B_y,$$

where γ (Greek letter gamma, which gakes the 'g' sound) is shorthand for $1/\sqrt{1 - (v/c)^2}$. The form of these equations is pretty simple if you ignore all the factors of c and realize that there is just an over-all factor of γ . The factors of c occur in front of every magnetic field, and also under every v , so that all velocities are being measured as a fraction of the speed of light.

The trivial transformation

example 1

In the case where $v = 0$, we have $\gamma = 1$, and the transformations become $E'_x = E_x$ and $B'_y = B_y$, i.e., nothing changes. This makes sense, because we haven't actually changed our frame of reference.

The motorcycle at 3% of c

example 2

Suppose that Einstein rides his motorcycle through an electromagnetic plane wave at 3% of c . (He'd better not hit anything at this speed, because he has more kinetic energy than is released in a nuclear explosion.) Because v/c is small, we have $\gamma = 1.0004$, which means that to an excellent approximation we can ignore all the factors of γ . This is an electromagnetic wave he's riding through, so $E_x = cB_y$, i.e., the electric and magnetic fields are "the same" when we ignore factors of c .

The result of the transformation of the fields is straightforward: each field is reduced by almost exactly 3%. This makes sense intuitively by comparison with material objects. For example, running in the same direction as the wind makes the wind feel less intense.

The fact that both fields are reduced by the same 3% is important. They need to obey the laws of physics in the new frame as well as the old one, and the laws of physics require the electric and magnetic fields to be equal to each other in a plane wave.

We can also see that this makes sense graphically. A velocity this small is hard to draw clearly, but figure e shows what happens with a somewhat higher speed.

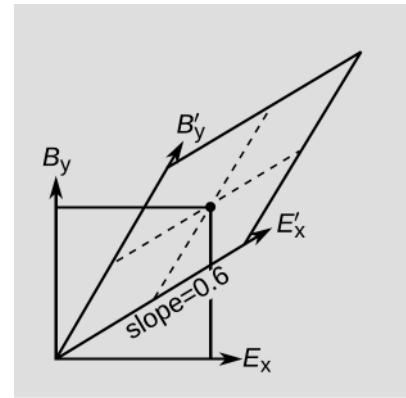
The motorcycle at 99% of c

example 3

The velocity in example 2 was small compared to c , but by extrapolating to higher speeds, we can imagine that this example helps to solve the mystery of the motorcycle in the case where we let the velocity approach c . We expect the strength of the fields to keep going down, and we can guess that they will approach zero and become undetectable, nullifying the paradox.

To check this, let's redo the calculation of example 2, but now with $v/c = 0.99$. It's no longer a good approximation to take $\gamma \approx 1$; we now have $\gamma \approx 7$. The reduction in the intensity of each field is a factor of $(1 - v/c)\gamma = 0.07$, which checks out with our guess (and one can indeed prove that $\lim_{x \rightarrow 1} (1-x)(1-x^2)^{-1/2} = 0$.)

Earlier I claimed that we could think of v/c as being the same thing as the slope of the tilted horizontal axis on the graph, and again we see that this makes sense in the case of the trivial transformation, example 1. This claim is justified in more detail in note [≥185](#).



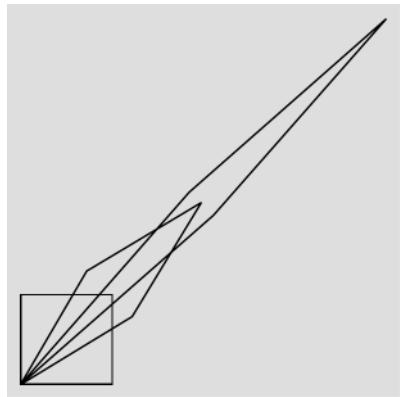
e / The dot represents the fields in an electromagnetic plane wave. Changing to a frame of reference moving at 60% of c cuts the fields in half, but they are still equal to one another, as required for a plane wave. This is similar to, but easier to see, than the 3% velocity of example 2.

7.2 Implications for the structure of space and time

7.2.1 Combination of velocities

These apparently innocuous ideas now lead us to some extremely subversive conclusions about time and space. We are now led back, full circle, to the starting point of this book, in which we began by considering the nature of time.

Figure f shows the result of doing two transformations in a row, each represented by its effect on a square of the original graph paper. For convenience, each of these is chosen to be a transformation that lengthens one diagonal by a factor of two, while cutting the other in half. With a little arithmetic, it can be shown that each of these corresponds to a velocity change v equal to $3/5$ of c . That is, we have our original frame A, a frame B that moves at v relative to A and another frame C that moves at v relative to B. According to the conception of space and time created by Galileo and Newton, velocities should add, so the result should be $3/5 + 3/5 = 6/5$, i.e., a speed 20% greater than the speed of light. But this isn't what we actually see in the diagram. In each of the three parallelograms, the slope of the bottom edge represents that frame's velocity in units of c . The first slope is 0, and the second is $3/5$, but the third is clearly less than 1.

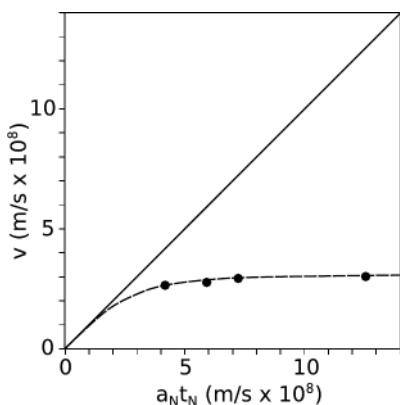


f / Transforming twice in a row by $3/5$ of c . The slope of the bottom edge of each parallelogram is its velocity, in units of c .

Einstein recalled that this was the point that had prevented him from putting together the special theory of relativity for several years. He had been assuming that velocities had to add. It turns out that they don't. Addition is just an approximation, which happens to work at velocities that are small compared to c .

This was also the resolution of the motorcycle paradox that had been bothering him for years. Any continuous process of acceleration acts like figure f. The motorcycle can speed up and speed up, but it will never reach the speed of light. Let's not detour into a detailed description right now, but the rider sees the wave reduced in amplitude, as in example 2, p. 177. He also sees the wave's oscillations slowed down in time, an effect called the Doppler shift (example 5).

Slopes on these diagrams represent velocities, in units of c , but the only slopes that stay the same when we change frames of reference are the ± 1 slopes of the diagonals. We therefore find that all observers agree on c . Since we've already shown on p. 158 that electromagnetic waves travel at a fixed speed, it follows that this speed must be c , and this is why c is often referred to as the speed of light.



g / Example 15.

Accelerating electrons

example 4

Figure ao shows the results of a 1964 experiment by Bertozzi in

which electrons were accelerated by the static electric field E of an accelerator of length ℓ_1 . They were then allowed to fly down a beamline of length $\ell_2 = 8.4$ m without being acted on by any force. The time of flight t_2 was used to find the final velocity $v = \ell_2/t_2$ to which they had been accelerated. (To make the low-energy portion of the graph legible, Bertozzi's highest-energy data point is omitted.)

If we believed in Newton's laws, then the electrons would have an acceleration $a_N = Ee/m$, which would be constant if, as we pretend for the moment, the field E were uniform. (The electric field inside this type of accelerator is not really quite uniform, but this will turn out not to matter.) The Newtonian prediction for the time over which this acceleration occurs is $t_N = \sqrt{2m\ell_1/eE}$. An acceleration a_N acting for a time t_N should produce a final velocity $a_N t_N = \sqrt{2e\Delta\phi/m}$, where $\Delta\phi = E\ell_1$ is the potential difference. (By conservation of energy, this equation holds even if the field is not constant.) 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.

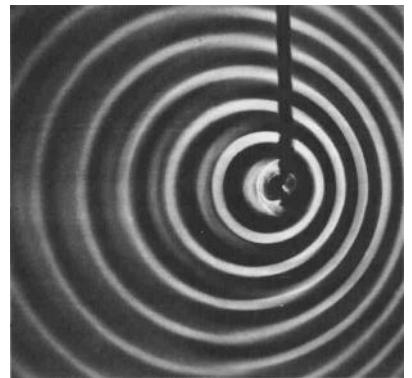
The experimental data, shown as black dots, clearly tell a different story. The velocity asymptotically approaches a limit, which we identify as c . The dashed line shows the predictions of special relativity, which we are not yet ready to calculate because we haven't yet seen how kinetic energy depends on velocity at speeds comparable to c .

An ultra-high-precision test of relativity

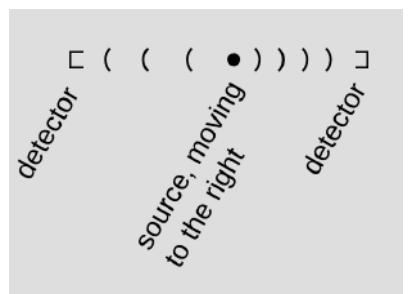
example 5

I briefly mentioned above that an observer chasing an electromagnetic wave observes an increase in the wave's period, which is called the Doppler effect. Let's notate the factor by which the period increases as D . The Doppler effect also occurs for other waves such as water waves (figure h) and sound waves, but the math in these cases actually works out to be more complicated because there are two velocities involved: the wave's velocity relative to the medium and the observer's velocity relative to the medium. Because electromagnetic waves are not vibrations of a medium, their Doppler shifts can depend on only a single parameter v , which is the velocity of one observer relative to another observer. The Doppler shift factor D turns out to be exactly the factor by which the long diagonal is stretched in figures like d and f. When we change frames of reference more than once, as in figure f, the stretch factors just multiply, so it is a firm prediction of relativity that Doppler shifts must multiply in this way, and this is different from the predictions of nonrelativistic theories such as ones in which light is a vibration of the ether.

This has led to one of the most high-precision tests to which a sci-



h / The pattern of waves made by a point source moving to the right across the water. Because the wave crests on the right side are closer together, they pass over a fixed point more frequently than if the source had been at rest. This increase in frequency is the Doppler effect.



i / A schematic drawing of the experiment in example 5.

entific theory has ever been subjected. Suppose that a source, moving at velocity v to the right relative to the laboratory, emits electromagnetic waves in both the forward and backward directions. Relativity predicts that the forward and backward Doppler shifts must be such that $D_1 D_2 = 1$. The first test of this type was carried out by Ives and Stilwell in 1938, and a particularly exquisite higher-precision update was carried out in 2003 by Saathoff *et al.* The electromagnetic waves were provided by positively charged lithium ions accelerated to $v/c = 0.064$ in a circular accelerator. The result was $D_1 D_2 = 0.999999999$, with error bars of about ± 1 in the final decimal place. The spectacular agreement with theory has made this experiment a lightning rod for anti-relativity kooks.

Ether theories predict, in the case where the lab is at rest relative to the ether, that $D_1 D_2$ differs from one. This is easiest to see conceptually in the case where the source moves at c , which is not prohibited in ether theories. In this situation, the forward-going wave crests are all superimposed on top of each other, giving $D_1 = \infty$, while D_2 is finite. For smaller velocities, such as the ones used in the Saathoff experiment, a calculation shows that $D_1 D_2$ differs from 1 by $-(v/c)^2$, which is grossly inconsistent with the data.

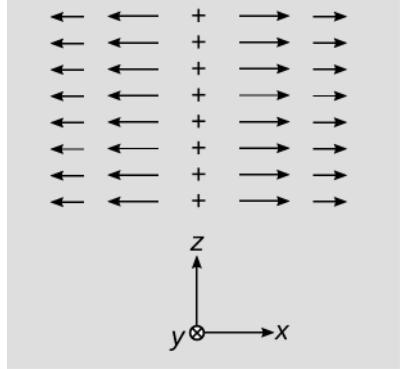
7.2.2 Length contraction

The nonlinear way that velocities combine is not compatible with our ordinary assumptions about how space and time work, but a more direct and obvious violation of our intuition arises if we consider figure j. A line of charges parallel to the z axis creates an electric field pattern. We pick a point on the right side of the line, where the electric field E_x has some positive value. If we now change to a frame of reference moving at velocity v in the positive z direction, then the equations for the transformation of the fields give us $E'_x = \gamma E_x$ and $B'_y = -(v\gamma/c^2)E_x$. The magnetic field is no surprise — a current exists in the new frame, and it makes a magnetic field (see example 8 below).

But the electric field has been intensified by a factor γ , which is greater than 1. What is going on here? Based on Gauss's law, this is the field that we expect if the line of charge had a greater number of coulombs per meter. But charge is invariant when we change frames of reference (sec. 2.6.2, p. 56). The only possible conclusion is that to the observer who is moving relative to the line of charges, the distances along the line have all been contracted by a factor of γ .

Length contraction

A measuring stick has the greatest length according to an observer at rest relative to the stick. An observer in motion



j / A line of charge creates an electric field.

relative to the stick, in the direction parallel to it, finds its length to be reduced by a factor of γ .

Figure t shows how γ depends on v/c . Because it is a smooth, even function of v , its derivative vanishes at $v = 0$, and the lowest-order varying term in its Taylor series is one proportional to $(v/c)^2$ (problem 7, p. 186). This explains why we don't normally notice length contraction in everyday life. The velocities we normally experience are small compared to c , so v/c is small, and $(v/c)^2$ is even smaller.

Magnetic field of a straight wire

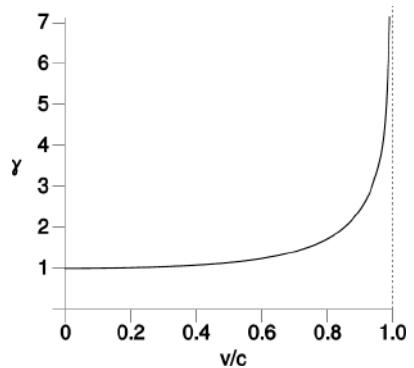
example 6

Although it's not our main topic right now, we now know enough physics to determine a fairly useful result for the magnetic field of a long, straight wire. In example 8 on p. 58, we worked out the electric field of a long line of charge with a density of λ in units of Coulombs per meter. The result was $E = 2k\lambda/r$ in the radial direction. In example 2 on p. 121, we discussed such a line of charge in a frame of reference moving along its length. In this frame there is an electric current, which makes a magnetic field. In that example we inferred that since $E \propto 1/r$ for the line of charge, $B \propto 1/r$ for the magnetic field of a current-carrying wire, and we also found the geometry of the magnetic field, which is recapitulated in figure l.

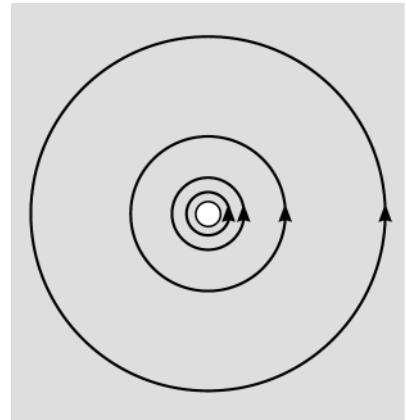
We can now find the full equation for the magnetic field of the current-carrying wire, which we model as a moving line of charge. In the frame moving relative to the charges at velocity v into the page, they are length-contracted, so their density becomes $\gamma\lambda$. In a time interval dt , the charges move a distance $v dt$, and the charge contained in this length is $dq = \gamma\lambda v dt$. The current is therefore $I = dq/dt = \gamma\lambda v$.

The magnitude of the magnetic field, as discussed at the beginning of this section, is $B = (v/c^2)\gamma E$, which works out to be $B = (v/c^2)\gamma(2k\lambda/r) = (2k/c^2)I/r$. All of the factors except for the 2 have to be there because of units. The fact that no γ occurs in the result is an example of a more general fact that when moving charges create a magnetic field, the field depends only on the current, not on any other details of the motion of the charges.

Real wires in electric circuits are not moving lines of charge. They are electrically neutral over all. In a copper wire, for example, there are positively charged copper nuclei, each with charge $+29e$, and also 29 electrons per atom, giving a charge of $-29e$. We could worry about whether this means that our calculation above is wrong, but again, it never ends up mattering how a current is created when we calculate a magnetic field. The electric field discussed above is wrong for a real wire — it should be zero — but the magnetic field is still right. If we wished, we could redo the calculation of B using two superposed lines of charge, a pos-



k / A graph of γ as a function of v/c .

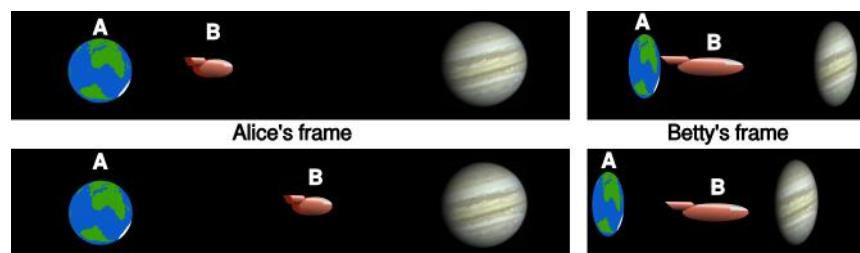


l / Example 6, the magnetic field of a current-carrying wire, with the current coming out of the page. The field pattern is shown in the plane perpendicular to the wire. The white circle at the center is the cross-section of the wire. The orientation differs by 90 degrees from that in figure j.

itive one and a negative one, but B would be the same.

7.2.3 Time dilation

Alice stays on earth while her twin Betty heads off in a spaceship for Tau Ceti, a nearby star. Tau Ceti is 12 light-years away, so even though Betty travels at 87% of the speed of light, it will take her a long time to get there: 14 years, according to Alice.



m / Betty's int

Betty is moving relative to the stars, so in her frame of reference, the length of the trip is length-contracted. At this speed, her y is 2.0, so that the voyage will seem half as long, and to her it will only last 7 years.

This example shows that the relativistic effect on length must be accompanied by a similar effect on time.

Time dilation

A clock seems to run fastest according to an observer at rest relative to the clock. An observer in motion relative to the clock finds its speed to be reduced by a factor of γ .

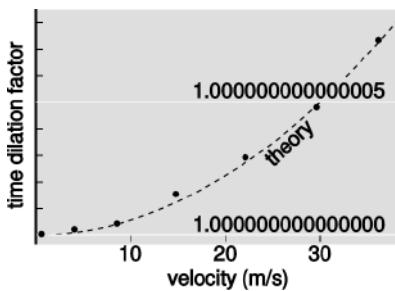
We began this book with evidence that time was not absolute, and then used this as evidence that electromagnetism should be a theory of fields, not instantaneous interaction at a distance. We have now circled back and used our understanding of fields to infer a more detailed and quantitative description of the relative nature of time.

Before passing to a description of two laboratory tests of time dilation in examples 7 and 5, we note that relativistic time dilation is these days a part of everyday life. The GPS system used in phones would not work at all if it didn't take into account the time dilation of the GPS satellites, which are moving at $v/c \sim 10^{-10}$. GPS was originally a military system, and legend has it that the general in charge of the project in the 1980's demanded that the scientists include an "off" switch for the relativistic effects, just in case Einstein was wrong.

A moving atomic clock

example 7

When v is small, relativistic effects are approximately proportional to v^2 , so it is very difficult to observe them at low speeds. For ex-



n / Example 7, time dilation measured with an atomic clock at low speeds. The theoretical curve, shown with a dashed line, is calculated from $\gamma = 1/\sqrt{1 - (v/c)^2}$. This graph corresponds to an extreme close-up view of the lower left corner of figure t. The error bars on the experimental points are about the same size as the dots.

ample, a car on the freeway travels at about 1/10 the speed of a passenger jet, so the resulting time dilation is only 1/100 as much. For this reason, it was not until four decades after Hafele and Keating that anyone did a conceptually simple atomic clock experiment in which the only effect was motion, not gravity; it is difficult to move a clock at a high enough velocity without putting it in some kind of aircraft, which then has to fly at some altitude. In 2010, however, Chou *et al.*¹ succeeded in building an atomic clock accurate enough to detect time dilation at speeds as low as 10 m/s. Figure n shows their results. Since it was not practical to move the entire clock, the experimenters only moved the aluminum atoms inside the clock that actually made it “tick.”

Large time dilation

example 8

The time dilation effect in the Hafele-Keating experiment was very small. If we want to see a large time dilation effect, we can't do it with something the size of the atomic clocks they used; the kinetic energy would be greater than the total megatonnage of all the world's nuclear arsenals. We can, however, accelerate subatomic particles to speeds at which γ is large. For experimental particle physicists, relativity is something you do all day before heading home and stopping off at the store for milk. An early, low-precision experiment of this kind was performed by Rossi and Hall in 1941, using naturally occurring cosmic rays.

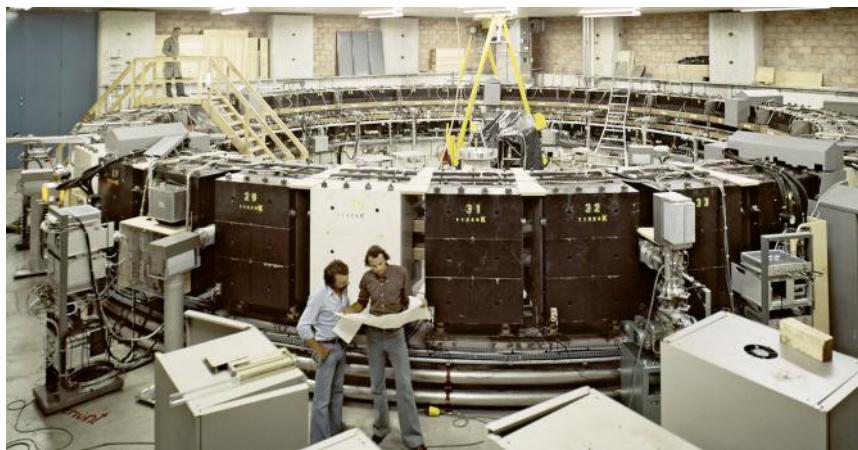
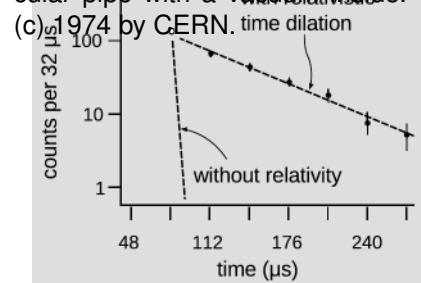


Figure w shows a 1974 experiment² of a similar type which verified the time dilation predicted by relativity to a precision of about one part per thousand. Particles called muons (named after the Greek letter μ , “myoo”) were produced by an accelerator at CERN, near Geneva. A muon is essentially a heavier version of the electron. Muons undergo radioactive decay, lasting an average of only

¹Science 329 (2010) 1630

²Bailey at al., Nucl. Phys. B150(1979) 1

o / Apparatus used for the test of relativistic time dilation described in example 5. The prominent black and white blocks are large magnets surrounding a circular pipe with a vacuum inside.



p / Example 5: Muons accelerated to nearly c undergo radioactive decay much more slowly than they would according to an observer at rest with respect to the muons. The first two data-points (unfilled circles) were subject to large systematic errors.

$2.197 \mu\text{s}$ before they evaporate into an electron and two neutrinos. The 1974 experiment was actually built in order to measure the magnetic properties of muons, but it produced a high-precision test of time dilation as a byproduct. Because muons have the same electric charge as electrons, they can be trapped using magnetic fields. Muons were injected into the ring shown in figure w, circling around it until they underwent radioactive decay. At the speed at which these muons were traveling, they had $\gamma = 29.33$, so on the average they lasted 29.33 times longer than the normal lifetime. In other words, they were like tiny alarm clocks that self-destructed at a randomly selected time. Figure v shows the number of radioactive decays counted, as a function of the time elapsed after a given stream of muons was injected into the storage ring. The two dashed lines show the rates of decay predicted with and without relativity. The relativistic line is the one that agrees with experiment.

Notes for chapter 7

2177 Slope on (E_x, B_y) plot related to v/c

When we change frames of reference, the horizontal axis on a graph of B_y versus E_x is to give it a slope equal to v/c , where v is the velocity with which the new frame moves in the z direction.

Let's call the slope u and the velocity v . Our goal is to prove that $u = v/c$. Clearly this equation holds when $v = 0$, because there is then no difference between the two frames of reference, so the axis shouldn't tilt, which means $u = 0$.

We can also tell several things from the example of figure c on p. 121: (1) If $v \neq 0$, then $u \neq 0$, i.e., if we change to a frame of reference that really is different, then we really do see a different combination of fields. (2) By symmetry, u should be an odd function of v . (3) For small velocities, the current in this example is proportional to v , and therefore u must be proportional to v , or at least approximately so for small velocities. The smallness of the velocities matters at this point because at larger velocities, γ could differ significantly from 1, and this would make all of our arguments quite a bit more complicated. Although $u = v/c$ is really exactly true at all velocities, we'll content ourselves here with proving it in the low-velocity approximation. (4) Based on units, the constant of proportionality between u and v has to be $1/c$ multiplied by a unitless constant. We want to show that this unitless constant is 1.

Having figured out roughly what we expect, we now proceed to fix the relationship between u and v by requiring that if Maxwell's equations held in the original frame, they hold as well in the new frame. It's sufficient just to consider the equation $\text{curl } \mathbf{E} = -\partial \mathbf{B} / \partial t$.

Suppose Einstein is chasing the electromagnetic wave at a v that is small, e.g., 3% of the speed of light, as in example 2 on p. 177. Then $\partial \mathbf{B} / \partial t$ will be multiplied by approxi-

mately $1 - v/c$, i.e., reduced by just about 3% in this example, because the amount of the wave that is passing over him in a given time is reduced by that amount. (If he could travel at $v = c$, then this factor would be $1 - v/c = 0$, i.e., none of the wave would pass over him.)

At the same time, the electric field is reduced according to $E'_x = \gamma E_x - (u/c)\gamma c B_y$, which becomes $E'_x \approx E_x - (u/c)c B_y$ for small velocities. But this is an electromagnetic wave, so $E_x = c B_y$, and therefore E_x is multiplied by the factor $1 - u$. This means that $\text{curl } \mathbf{E}$ is also cut down by this amount.

If Maxwell's equation $\text{curl } \mathbf{E} = -\partial \mathbf{B} / \partial t$ is to hold in the new frame, then the factors by which the two sides are reduced must be the same, giving $1 - u = 1 - v/c$, which is the desired result.

Problems

Key

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

1 Can a field that is purely electrical in one frame of reference be purely magnetic in some other frame?

2 Astronauts in three different spaceships are communicating with each other. Those aboard ships A and B agree on the rate at which time is passing, but they disagree with the ones on ship C.

- (a) Alice is aboard ship A. How does she describe the motion of her own ship, in its frame of reference?
- (b) Describe the motion of the other two ships according to Alice.
- (c) Give the description according to Betty, whose frame of reference is ship B.
- (d) Do the same for Cathy, aboard ship C.

3 What happens in the equation for γ when you put in a negative number for v ? Explain what this means physically, and why it makes sense.

4 The Voyager 1 space probe, launched in 1977, is moving faster relative to the earth than any other human-made object, at 17,000 meters per second.

- (a) Calculate the probe's γ .
- (b) Over the course of one year on earth, slightly less than one year passes on the probe. How much less? (There are 31 million seconds in a year.) ✓

5 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. ✓

6 (a) Show that for $v = (3/5)c$, γ comes out to be a simple fraction.

- (b) Find another value of v for which γ is a simple fraction.

7 The relativistic factor

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

introduced in the text, gives the amount of length contraction and time dilation. We express it here in units in which $c = 1$. Find the first two nonvanishing terms in its Taylor series, and show that, as claimed on p. 181, the first non-constant term is of order v^2 .

8 Example 5, p. 179, discussed the Doppler effect. The function D discussed in that example is given by

$$D(v) = \sqrt{\frac{1+v}{1-v}},$$

where the velocity v is expressed as a fraction of c . Expand this function in a Taylor series, and find the first two nonvanishing terms. Show that these two terms agree with the nonrelativistic expression $1 + v$, so that any relativistic effect is of higher order in v .

DC circuits

Chapter 8

Electrical resistance

8.1 Circuits

8.1.1 Complete circuits, open circuits

How can we put electric currents to work? Figure a/1 shows an attempt to use static electricity to light a lightbulb. We take two substances, e.g., glass and fur, rub them together, and then touch them to wires connected to a lightbulb. This method is unsatisfactory. True, current will flow through the bulb, since electrons can move through metal wires, and the excess electrons on the glass rod will therefore come through the wires and bulb due to the attraction of the positively charged fur and the repulsion of the other electrons. The problem is that after a zillionth of a second of current, the rod and fur will both have run out of charge. No more current will flow, and the lightbulb will go out.

Figure a/2 shows a setup that works. The battery pushes charge through the circuit, and recycles it over and over again. (We will have more to say later in this chapter about how batteries work.) This is called a *complete circuit*. Today, the electrical use of the word “circuit” is the only one that springs to mind for most people, but the original meaning was to travel around and make a round trip, as when a circuit court judge would ride around the boondocks, dispensing justice in each town on a certain date.

Note that an example like a/3 does not work. The wire will quickly begin acquiring a net charge, because it has no way to get rid of the charge flowing into it. The repulsion of this charge will make it more and more difficult to send any more charge in, and soon the electrical forces exerted by the battery will be canceled out completely. The whole process would be over so quickly that the filament would not even have enough time to get hot and glow. This is known as an *open circuit*. Exactly the same thing would happen if the complete circuit of figure a/2 was cut somewhere with a pair of scissors, and in fact that is essentially how an ordinary light switch works: by opening up a gap in the circuit.

8.1.2 Measuring the current in a circuit

The measurement of potential (voltage) differences was discussed in sec. 4.2, p. 89.

On p. 122, we defined electric current as the rate at which electric

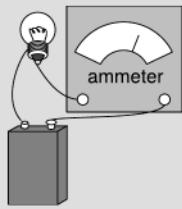


a / 1. Static electricity runs out quickly. 2. A practical circuit. 3. An open circuit.

1



2



b / 1. How a simple ammeter works. 2. Measuring a current with an ammeter.

charge flows across a boundary, $I = dq/dt$. This definition has the theoretical advantage that when moving charges create a magnetic field, the field is exactly proportional to the current, independent of other variables such as the velocity of the charges. Now that we are ready to study circuits in more detail, a second, practical advantage of this definition becomes apparent, which is that current is easy to measure compared to quantities like charge or the velocity of charges. The instrument used to measure current is the ammeter, introduced briefly on p. 123. A simplified ammeter, b/1, simply consists of a coiled-wire magnet whose force twists an iron needle against the resistance of a spring. The greater the current, the greater the force. Although the construction of ammeters may differ, their use is always the same. We break into the path of the electric current and interpose the meter like a tollbooth on a road, b/2. There is still a complete circuit, and as far as the battery and bulb are concerned, the ammeter is just another segment of wire.

Does it matter where in the circuit we place the ammeter? Could we, for instance, have put it in the left side of the circuit instead of the right? Conservation of charge tells us that this can make no difference. Charge is not destroyed or “used up” by the lightbulb, so we will get the same current reading on either side of it. What is “used up” is energy stored in the battery, which is being converted into heat and light energy.

Figure c shows a typical multimeter used as an ammeter. As always with a multimeter, one cable is plugged into the COM plug. The other cable is plugged into the plug marked 10 A, meaning that it’s used for current measurements and can handle large currents, up to 10 A. The rotary dial is turned to the corresponding A setting. The icons indicate that this position of the dial can be used to measure either AC or DC current, and the yellow button has been used to select DC, as verified on the LCD readout. The lines connecting the two plugs that are in use say “FUSED,” which means that if too much current flows, the meter will avoid damaging itself by blowing a fuse. The meter’s reading of 4.572 A verifies that the amount of current being measured is appropriate for this scale. If it had been more than 10 A, the fuse would have blown, and the meter would be reading zero. If we were intending to read a current smaller than 400 mA, then we would be better off changing both the plug and the rotary dial to use the more sensitive setting, which would give us better precision (more sig figs).

The following table summarizes some differences between the use of a voltmeter (sec. 4.2, p. 89) and an ammeter.



c / Practical use of a multimeter as an ammeter.

<i>voltmeter</i>	<i>ammeter</i>
Measures potential difference, $\Delta\phi$, in units of volts.	Measures electric current, I , in units of amperes.
Doesn't require a complete circuit.	Requires a complete circuit.
Is used to probe the circuit without breaking it.	Requires temporarily breaking the circuit in order to insert the meter.
Has no fuse.	Has a fuse, which can be blown.
Used in parallel. If used in series, will cause an open circuit.	Used in series. If used in parallel, will cause a short circuit (sec. 8.3.3, p. 198) and blow the fuse.
Ideally has infinite resistance (sec. 8.3.1, p. 195).	Ideally has zero resistance.

Magnetic levitation

example 1

In figure d, a small, disk-shaped permanent magnet is stuck on the side of a battery, and a wire is clasped loosely around the battery, shorting it. A large current flows. The electrons moving through the wire feel a force from the magnetic field made by the permanent magnet, and this force levitates the wire.

From the photo, it's possible to find the direction of the magnetic field made by the permanent magnet. The electrons in the copper wire are negatively charged, so they flow from the negative (flat) terminal of the battery to the positive terminal (the one with the bump, in front). As the electrons pass by the permanent magnet, we can imagine that they would experience a field either toward the magnet, or away from it, depending on which way the magnet was flipped when it was stuck onto the battery. If q had been positive, then $qv \times \mathbf{B}$ would have been in the direction of $\mathbf{v} \times \mathbf{B}$, determined by the right-hand rule. Since $q < 0$ here, the right-hand rule flips its handedness, as shown in the bottom panel of the figure. The field must be toward the battery.

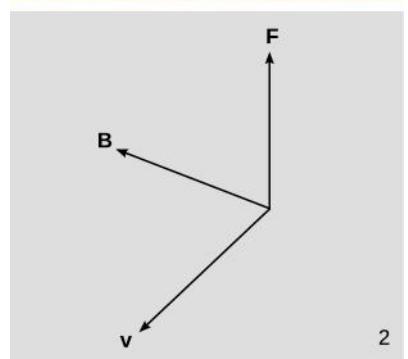
This example is very similar to the design of the simplified ammeter in figure b/1. The main difference is that the roles of the known and unknown quantities is reversed: we know something about the direction of the current, and we use it to find out about the unknown direction of the magnetic field of the permanent magnet.

Discussion question

- A What would have happened if we had analyzed example 1 by assuming that the charge carriers were positively charged? If we already knew the polarity of the magnet, would we be able to use this experiment to prove that the mobile charge carriers in a copper wire are negatively charged? What happens inside the battery, where both positive and negative ions may be flowing?



1



2

d / Example 1.

8.2 Power

Electrical circuits can be used for sending signals, storing information, or doing calculations, but their most common purpose by far is to manipulate energy, as in the battery-and-bulb example of the previous section. We know that lightbulbs are rated in units of watts, i.e., how many joules per second of energy they can convert into heat and light, but how would this relate to the flow of charge as measured in amperes? By way of analogy, suppose your friend, who didn't take physics, can't find any job better than pitching bales of hay. The number of calories he burns per hour will certainly depend on how many bales he pitches per minute, but it will also be proportional to how much mechanical work he has to do on each bale. If his job is to toss them up into a hayloft, he will get tired a lot more quickly than someone who merely tips bales off a loading dock into trucks. In metric units,

$$\frac{\text{joules}}{\text{second}} = \frac{\text{haybales}}{\text{second}} \times \frac{\text{joules}}{\text{haybale}}.$$

Similarly, the rate of energy transformation by a battery will not just depend on how many coulombs per second it pushes through a circuit but also on how much mechanical work it has to do on each coulomb of charge:

$$\frac{\text{joules}}{\text{second}} = \frac{\text{coulombs}}{\text{second}} \times \frac{\text{joules}}{\text{coulomb}}$$

or

$$\text{power} = \text{current} \times \text{voltage}.$$

Units of volt-amps

example 2

- ▷ Doorbells are often rated in volt-amps. What does this combination of units mean?
- ▷ Current times voltage gives units of power, $P = I\Delta V$, so volt-amps are really just a nonstandard way of writing watts. They are telling you how much power the doorbell requires.

Power dissipated by a battery and bulb

example 3

- ▷ If a 9.0-volt battery causes 1.0 A to flow through a lightbulb, how much power is dissipated?
- ▷ The voltage rating of a battery tells us what voltage difference ΔV it is designed to maintain between its terminals.

$$\begin{aligned} P &= I\Delta V \\ &= 9.0 \text{ A} \cdot \text{V} \\ &= 9.0 \frac{\text{C}}{\text{s}} \cdot \frac{\text{J}}{\text{C}} \\ &= 9.0 \text{ J/s} \\ &= 9.0 \text{ W} \end{aligned}$$

The only nontrivial thing in this problem was dealing with the units. One quickly gets used to translating common combinations like $A \cdot V$ into simpler terms.

Discussion questions

A A roller coaster is sort of like an electric circuit, but it uses gravitational forces on the cars instead of electric ones. What would a high-voltage roller coaster be like? What would a high-current roller coaster be like?

B Criticize the following statements:

“He touched the wire, and 10000 volts went through him.”

“That battery has a charge of 9 volts.”

“You used up the charge of the battery.”

C When you touch a 9-volt battery to your tongue, both positive and negative ions move through your saliva. Which ions go which way?

D I once touched a piece of physics apparatus that had been wired incorrectly, and got a several-thousand-volt voltage difference across my hand. I was not injured. For what possible reason would the shock have had insufficient power to hurt me?

8.3 Resistance

8.3.1 Resistance

So far we have simply presented it as an observed fact that a battery-and-bulb circuit quickly settles down to a steady flow, but why should it? Newton’s second law, $a = F/m$, would seem to predict that the steady forces on the charged particles should make them whip around the circuit faster and faster. The answer is that as charged particles move through matter, there are always forces, analogous to frictional forces, that resist the motion. These forces need to be included in Newton’s second law, which is really $a = F_{total}/m$, not $a = F/m$. If, by analogy, you push a crate across the floor at constant speed, i.e., with zero acceleration, the total force on it must be zero. After you get the crate going, the floor’s frictional force is exactly canceling out your force. The chemical energy stored in your body is being transformed into heat in the crate and the floor, and no longer into an increase in the crate’s kinetic energy. Similarly, the battery’s internal chemical energy is converted into heat, not into perpetually increasing the charged particles’ kinetic energy. Changing energy into heat may be a nuisance in some circuits, such as a computer chip, but it is vital in an incandescent lightbulb, which must get hot enough to glow. Whether we like it or not, this kind of heating effect is going to occur any time charged particles move through matter.

What determines the amount of heating? One flashlight bulb designed to work with a 9-volt battery might be labeled 1.0 watts, another 5.0. How does this work? Even without knowing the details



e / Georg Simon Ohm (1787–1854).

of this type of friction at the atomic level, we can relate the heat dissipation to the amount of current that flows via the equation $P = I\Delta V$. If the two flashlight bulbs can have two different values of P when used with a battery that maintains the same ΔV , it must be that the 5.0-watt bulb allows five times more current to flow through it.

For many substances, including the tungsten from which lightbulb filaments are made, experiments show that the amount of current that will flow through it is directly proportional to the voltage difference placed across it. For an object made of such a substance, we define its electrical *resistance* as follows:

definition of resistance

If an object inserted in a circuit displays a current flow proportional to the voltage difference across it, then we define its resistance as the constant ratio

$$R = \Delta V/I.$$

The units of resistance are volts/ampere, usually abbreviated as ohms, symbolized with the capital Greek letter omega, Ω .

Resistance of a lightbulb

example 4

- ▷ A flashlight bulb powered by a 9-volt battery has a resistance of 10Ω . How much current will it draw?
- ▷ Solving the definition of resistance for I , we find

$$\begin{aligned}I &= \Delta V/R \\&= 0.9 \text{ V}/\Omega \\&= 0.9 \text{ V}/(\text{V/A}) \\&= 0.9 \text{ A}\end{aligned}$$

Ohm's law states that many substances, including many solids and some liquids, display this kind of behavior, at least for voltages that are not too large. The fact that Ohm's law is called a "law" should not be taken to mean that all materials obey it, or that it has the same fundamental importance as Newton's laws, for example. Materials are called *ohmic* or *nonohmic*, depending on whether they obey Ohm's law. Although we will concentrate on ohmic materials in this book, it's important to keep in mind that a great many materials are nonohmic, and devices made from them are often very important. For instance, a transistor is a nonohmic device that can be used to amplify a signal (as in a guitar amplifier) or to store and manipulate the ones and zeroes in a computer chip.

In our previous discussion of conductors (sec. 4.3, p. 91), we saw that a perfect conductor has certain characteristics: the potential is constant throughout it, the electric field is perpendicular to its surface, and any net density of charge is present only at the surface.

These statements are exact for a material with exactly zero resistance, and are often good approximations for objects like a copper wire, but in general there is no sharp distinction between insulators and conductors. Some materials, such as silicon, lie midway between the two extremes, and are called semiconductors.

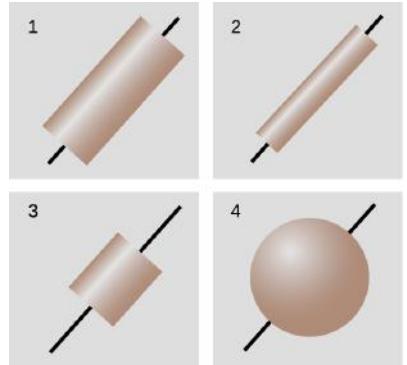
On an intuitive level, we can understand the idea of resistance by making the sounds “hhhhh” and “fffff.” To make air flow out of your mouth, you use your diaphragm to compress the air in your chest. The pressure difference between your chest and the air outside your mouth is analogous to a voltage difference. When you make the “h” sound, you form your mouth and throat in a way that allows air to flow easily. The large flow of air is like a large current. Dividing by a large current in the definition of resistance means that we get a small resistance. We say that the small resistance of your mouth and throat allows a large current to flow. When you make the “f” sound, you increase the resistance and cause a smaller current to flow.

Note that although the resistance of an object depends on the substance it is made of, we cannot speak simply of the “resistance of gold” or the “resistance of wood.” Figure f shows four examples of objects that have had wires attached at the ends as electrical connections. If they were made of the same substance, they would all nevertheless have different resistances because of their different sizes and shapes. A more detailed discussion will be more natural in the context of the following chapter, but it should not be too surprising that the resistance of f/2 will be greater than that of f/1 — the image of water flowing through a pipe, however incorrect, gives us the right intuition. Object f/3 will have a smaller resistance than f/1 because the charged particles have less of it to get through.

8.3.2 Superconductors

All materials display some variation in resistance according to temperature (a fact that is used in thermostats to make a thermometer that can be easily interfaced to an electric circuit). More spectacularly, most metals have been found to exhibit a sudden change to *zero* resistance when cooled to a certain critical temperature. They are then said to be superconductors. Currently, the most important practical application of superconductivity is in medical MRI (magnetic resonance imaging) scanners. When your body is inserted into one of these devices, you are being immersed in an extremely strong magnetic field produced by electric currents flowing through the coiled wires of an electromagnet. If these wires were not superconducting, they would instantly burn up because of the heat generated by their resistance.

There are many other potential applications for superconductors, but most of these, such as power transmission, are not currently economically feasible because of the extremely low temperatures re-



f / Four objects made of the same substance have different resistances.

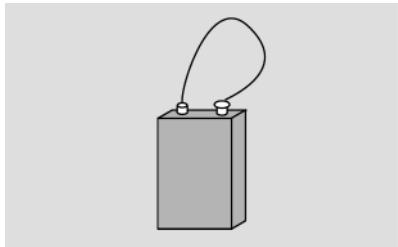


g / A medical MRI scanner, which uses superconductors.

quired for superconductivity to occur.

However, it was discovered in 1986 that certain ceramics are superconductors at less extreme temperatures. The technological barrier is now in finding practical methods for making wire out of these brittle materials. Wall Street is currently investing billions of dollars in developing superconducting devices for cellular phone relay stations based on these materials.

There is currently no satisfactory theory of superconductivity in general, although superconductivity in metals is understood fairly well. Unfortunately I have yet to find a fundamental explanation of superconductivity in metals that works at the introductory level.



h / Short-circuiting a battery.
Warning: you can burn yourself this way or start a fire! If you want to try this, try making the connection only very briefly, use a low-voltage battery, and avoid touching the battery or the wire, both of which will get hot.

8.3.3 Short circuits

So far we have been assuming a perfect conductor. What if it is a good conductor, but not a perfect one? Then we can solve for $\Delta V = IR$. An ordinary-sized current will make a very small result when we multiply it by the resistance of a good conductor such as a metal wire. The voltage throughout the wire will then be nearly constant. If, on the other hand, the current is extremely large, we can have a significant voltage difference. This is what happens in a *short-circuit*: a circuit in which a low-resistance pathway connects the two sides of a voltage source. Note that this is much more specific than the popular use of the term to indicate any electrical malfunction at all. If, for example, you short-circuit a 9-volt battery as shown in figure h, you will produce perhaps a thousand amperes of current, leading to a very large value of $P = I\Delta V$. The wire gets hot!

self-check A

What would happen to the battery in this kind of short circuit? ▷

Answer, p. 432

8.3.4 Resistors

Inside any electronic gadget you will see quite a few little circuit elements like the one shown in the photo. These *resistors* are simply a cylinder of ohmic material with wires attached to the end.

Many electrical devices are based on electrical resistance and Ohm's law, even if they do not have little components in them that look like the usual resistor. The following are some examples.

Lightbulb

There is nothing special about a lightbulb filament — you can easily make a lightbulb by cutting a narrow waist into a metallic gum wrapper and connecting the wrapper across the terminals of a 9-volt battery. The trouble is that it will instantly burn out. Edison solved this technical challenge by encasing the filament in an evacuated bulb, which prevented burning, since burning requires



Resistors.



i / The symbol used in schematics to represent a resistor.

oxygen.

Polygraph

The polygraph, or “lie detector,” is really just a set of meters for recording physical measures of the subject’s psychological stress, such as sweating and quickened heartbeat. The real-time sweat measurement works on the principle that dry skin is a good insulator, but sweaty skin is a conductor. Of course a truthful subject may become nervous simply because of the situation, and a practiced liar may not even break a sweat. The method’s practitioners claim that they can tell the difference, but you should think twice before allowing yourself to be polygraph tested. Most U.S. courts exclude all polygraph evidence, but some employers attempt to screen out dishonest employees by polygraph testing job applicants, an abuse that ranks with such pseudoscience as handwriting analysis.

Fuse

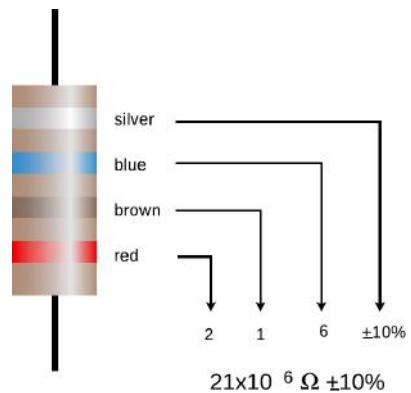
A fuse is a device inserted in a circuit tollbooth-style in the same manner as an ammeter. It is simply a piece of wire made of metals having a relatively low melting point. If too much current passes through the fuse, it melts, opening the circuit. The purpose is to make sure that the building’s wires do not carry so much current that they themselves will get hot enough to start a fire. Most modern houses use circuit breakers instead of fuses, although fuses are still common in cars and small devices. A circuit breaker is a switch operated by a coiled-wire magnet, which opens the circuit when enough current flows. The advantage is that once you turn off some of the appliances that were sucking up too much current, you can immediately flip the switch closed. In the days of fuses, one might get caught without a replacement fuse, or even be tempted to stuff aluminum foil in as a replacement, defeating the safety feature.

Discussion questions

A Explain why it would be incorrect to define resistance as the amount of charge the resistor allows to flow.

8.4 Flow of energy

We use “flipping a switch” as a metaphor for making something happen instantly, but it’s not obvious how this happens physically in an actual circuit. When we turn on a light, we convert an open circuit into a complete circuit. Charge starts to flow through the copper wires, but if we make an order-of-magnitude estimate of the speed at which it needs to flow in order to carry a typical current, the result is on the order of centimeters per second ([202](#)). This slow speed is called the drift velocity, and clearly it is much less than the speed at which energy flows. Figure 1 shows a mechanical analogy. A torque applied to one wheel is transmitted to the other



j / An example of a resistor with a color code.

black	0
brown	1
red	2
orange	3
yellow	4
green	5
blue	6
violet	7
gray	8
white	9
silver	$\pm 10\%$
gold	$\pm 5\%$

k / Color codes used on resistors.

wheel very rapidly, at a speed much greater than the speed at which the links in the roller chain travel. If the chain is tight, the speed at which the energy travels could be similar to the speed of sound in the metal. In an electric circuit, we are concerned with the speed of light rather than the speed of sound. Relativity guarantees that the flow of energy and information cannot be greater than c , but in most cases it is some significant fraction of c .

I / A mechanical analogy for the speed of transmission of energy in an electric circuit.

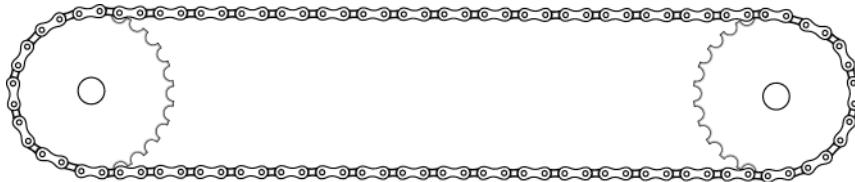


Figure n shows the results of a fairly realistic simulation of how the energy flows in an actual electric circuit shaped like a rectangle, with most of the resistance being supplied by the resistor on the right, but a smaller amount also present in the wires. This is a simulation of the steady state of the circuit, when the current has been flowing for a long time and has had time to settle down. The equipotentials were found by solving Laplace's equation on a computer. The flow of energy, measured by the Poynting vector \mathbf{S} , is almost entirely through the empty space surrounding the wires, not the wires themselves — because the wires are good conductors, $\mathbf{E} \approx 0$ inside them, and therefore the Poynting vector $\mathbf{S} \propto \mathbf{E} \times \mathbf{B}$ inside the wires themselves is not enough to transmit any significant amount of energy. Energy flows out of the battery and into the resistor, causing the resistor to heat up.¹

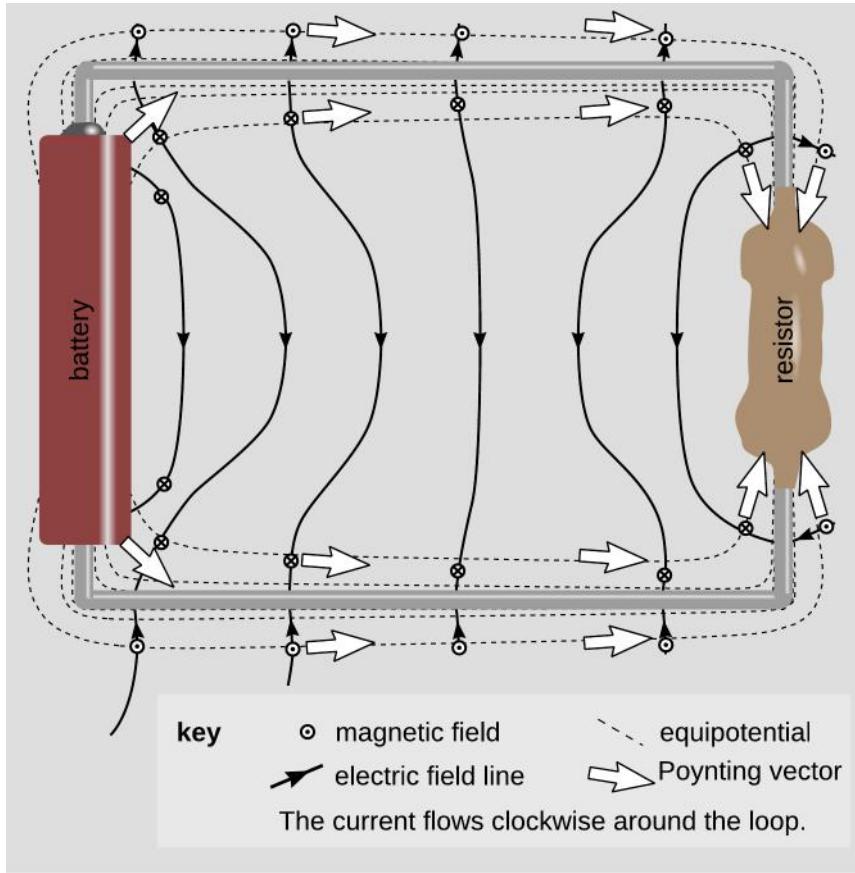
Note that the current is sometimes in the same direction as the energy flow and sometimes in the opposite direction. The energy is *not* transported by the electrons inside the wires.

If we were to suddenly open or close a switch in a circuit like this, the situation would be more complicated, but roughly speaking, we would expect that the changes in the energy flow would be initiated near the switch and then spread out through the circuit as electromagnetic waves. These would not be plane waves (and therefore would not travel at exactly c). They would be guided around the circuit by the wires. That is, the wires act not so much like pipes through which the energy flows, and more like railroad tracks that tell the cars where to go — the term “switch” originated in railroads, where it means the mechanism that allows cars to be directed onto one set of tracks or another when the train comes to a “Y” (figure m).



m / A complex set of switches in a big rail yard. By analogy with an electric circuit, the passengers and freight don't travel inside the rails. The rails merely guide the cars.

¹A little energy also flows into the wires, since the Poynting vectors near the wires are angled slightly inward. This causes a little bit of heating in the wires, which have nonzero resistance.

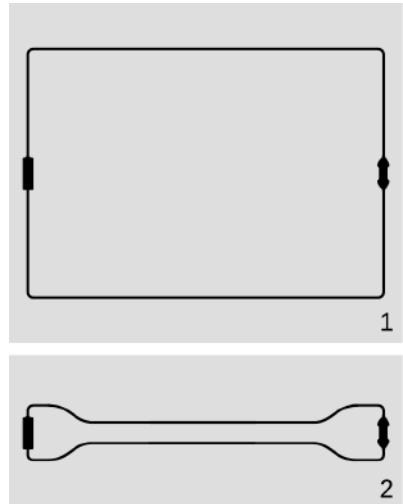


n / The flow of energy in a DC circuit. The flow of energy (white arrows) is from the battery, through empty space, and into the resistive elements.

Discussion questions

A Suppose that we take a pair of wire cutters to the circuit in figure n and make it into an open circuit. Electrons will soon stop flowing, but the energy is not transmitted by the electrons, so why does that matter?

B Figure o shows two variations on figure n. Because the batteries and resistors are identical in cases 1 and 2, the currents, voltage drops, and therefore the power are all identical as well. But this is not so obvious in terms of the flow of energy as determined by the Poynting vector. In the skinny version, we would expect that the magnetic fields of the two opposite currents would cancel rather well, and that would tend to make the Poynting vector small. What compensates for this?



o / Discussion question B.

Notes for chapter 8

≥199 Drift velocity

An order-of-magnitude estimate of the drift velocity.

Suppose that a copper wire contains one electron per atom that is free to move. Then the number density n of charge carriers is the same as the number density of copper atoms, which is just the inverse of the volume of an atom, $n \sim (1 \text{ nm})^{-3} = 10^{27} \text{ m}^{-3}$. When these electrons flow with an average velocity v , the absolute value of the current is $I = nevA$, where A is the wire's cross-sectional area. Putting in a typical current of 1 A and a cross-section area of one square millimeter, we find $v \sim 1 \text{ cm/s}$.

Problems

Key

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

1 In a wire carrying a current of 1.0 pA , how long do you have to wait, on the average, for the next electron to pass a given point? Express your answer in units of microseconds. The charge of an electron is $-e = -1.60 \times 10^{-19} \text{ C}$.

▷ Solution, p. 429

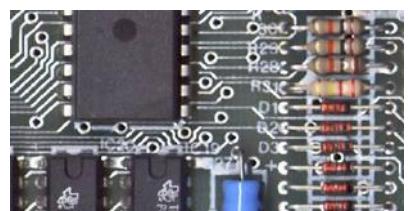
2 A silk thread is uniformly charged by rubbing it with llama fur. The thread is then dangled vertically above a metal plate and released. As each part of the thread makes contact with the conducting plate, its charge is deposited onto the plate. Since the thread is accelerating due to gravity, the rate of charge deposition increases with time, and by time t the cumulative amount of charge is $q = ct^2$, where c is a constant. (a) Find the current flowing onto the plate. ✓
(b) Suppose that the charge is immediately carried away through a resistance R . Find the power dissipated as heat. ✓

3 Lightning discharges a cloud during an electrical storm. Suppose that the current in the lightning bolt varies with time as $I = bt$, where b is a constant. Find the cloud's charge as a function of time. ✓

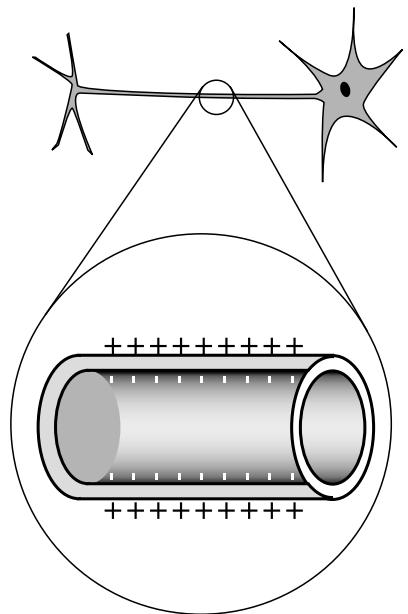
4 In AM (amplitude-modulated) radio, an audio signal $f(t)$ is multiplied by a sine wave $\sin \omega t$ in the megahertz frequency range. For simplicity, let's imagine that the transmitting antenna is a whip, and that charge goes back and forth between the top and bottom. Suppose that, during a certain time interval, the audio signal varies linearly with time, giving a charge $q = (a + bt) \sin \omega t$ at the top of the whip and $-q$ at the bottom. Find the current as a function of time. ✓

5 If a typical light bulb draws about 900 mA from a 110 V household circuit, what is its resistance? (Don't worry about the fact that it's alternating current.) ✓

6 You have to do different things with a circuit to measure current than to measure a voltage difference. Which would be more practical for a printed circuit board, in which the wires are actually strips of metal embedded inside the board? ▷ Solution, p. 429



A printed circuit board, like the kind referred to in problem 6.



Problem 4. Top: A realistic picture of a neuron. Bottom: A simplified diagram of one segment of the tail (axon).

7 A resistor has a voltage difference ΔV across it, causing a current I to flow.

(a) Find an equation for the power it dissipates as heat in terms of the variables I and R only, eliminating ΔV . ✓

(b) If an electrical line coming to your house is to carry a given amount of current, interpret your equation from part a to explain whether the wire's resistance should be small, or large.

8 (a) Express the power dissipated by a resistor in terms of R and ΔV only, eliminating I . ✓

(b) Electrical receptacles in your home are mostly 110 V, but circuits for electric stoves, air conditioners, and washers and dryers are usually 220 V. The two types of circuits have differently shaped receptacles. Suppose you rewire the plug of a drier so that it can be plugged in to a 110 V receptacle. The resistor that forms the heating element of the drier would normally draw 200 W. How much power does it actually draw now? ✓

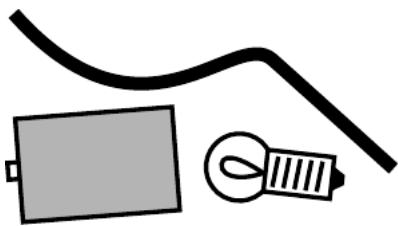
9 You are given a battery, a flashlight bulb, and a single piece of wire. Draw at least two configurations of these items that would result in lighting up the bulb, and at least two that would not light it. (Don't draw schematics.) Note that the bulb has two electrical contacts: one is the threaded metal jacket, and the other is the tip (at the bottom in the figure).

If you're not sure what's going on, there are a couple of ways to check. The best is to try it in real life by either borrowing the materials from your instructor or scrounging the materials from around the house. (If you have a flashlight with this type of bulb, you can remove the bulb.) Another method is to use the simulation at phet.colorado.edu/en/simulation/circuit-construction-kit-dc.

[Problem by Arnold Arons.]

10 (a) You take an LP record out of its sleeve, and it acquires a static charge of 1 nC. You play it at the normal speed of $33\frac{1}{3}$ r.p.m., and the charge moving in a circle creates an electric current. What is the current, in amperes? ✓

(b) Although the planetary model of the atom can be made to work with any value for the radius of the electrons' orbits, more advanced models that we will study later in this course predict definite radii. If the electron is imagined as circling around the proton at a speed of 2.2×10^6 m/s, in an orbit with a radius of 0.05 nm, what electric current is created? The charge of an electron is $-e = -1.60 \times 10^{-19}$ C. ✓



Problem 9.



An LP record, problem 10.

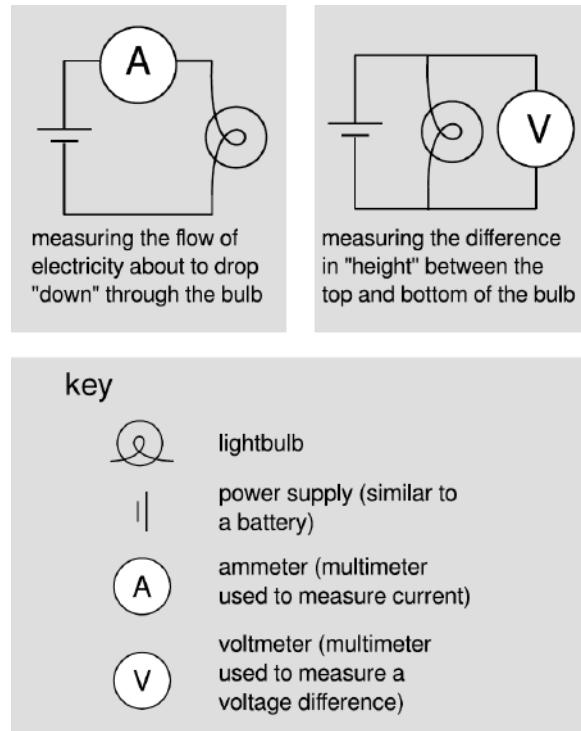
11 We have referred to resistors *dissipating* heat, i.e., we have assumed that $P = I\Delta V$ is always greater than zero. Could $I\Delta V$ come out to be negative for a resistor? If so, could one make a refrigerator by hooking up a resistor in such a way that it absorbed heat instead of dissipating it?

12 Hybrid and electric cars have been gradually gaining market share, but during the same period of time, manufacturers such as Porsche have also begun designing and selling cars with “mild hybrid” systems, in which power-hungry parts like water pumps are powered by a higher-voltage battery rather than running directly on shafts from the motor. Traditionally, car batteries have been 12 volts. Car companies have dithered over what voltage to use as the standard for mild hybrids, building systems based on 36 V, 42 V, and 48 V. For the purposes of this problem, we consider 36 V.

- (a) Suppose the battery in a new car is used to run a device that requires the same amount of power as the corresponding device in the old car. Based on the sample figures above, how would the currents handled by the wires in one of the new cars compare with the currents in the old ones? ✓
- (b) The real purpose of the greater voltage is to handle devices that need *more* power. Can you guess why they decided to change to higher-voltage batteries rather than increasing the power without increasing the voltage?

Exercise 8: Measuring voltage, current, and resistance

As shown in the figure, measuring current and voltage requires hooking the meter into the circuit in two completely different ways.



The arrangement for the ammeter is called a series circuit, because every charged particle that travels the circuit has to go through each component, one after another. The series circuit is arranged like beads on a necklace.

The setup for the voltmeter is an example of a parallel circuit. A charged particle flowing, say, clockwise around the circuit passes through the power supply and then reaches a fork in the road, where it has a choice of which way to go. Some particles will pass through the bulb, others (not as many) through the meter; all of them are reunited when they reach the junction on the right.

Students tend to have a mental block against setting up the ammeter correctly in series, because it involves breaking the circuit apart in order to insert the meter. To drive home this point, we will act out the process using stu-

dents to represent the circuit components. If you hook up the ammeter incorrectly, in parallel rather than in series, the meter provides an easy path for the flow of current, so a large amount of current will flow. To protect the meter from this surge, there is a fuse inside, which will blow, and the meter will stop working. This is not a huge tragedy; just ask your instructor for a replacement fuse and open up the meter to replace it.

Unscrew your lightbulb from its holder and look closely at it. Note that it has two separate electrical contacts: one at its tip and one at the metal screw threads.

Turn the power supply's off-on switch to the off position, and turn its knob to zero. Set up the basic lightbulb circuit without any meter in it. There is a rack of cables in the back of the room with banana-plug connectors on the end, and most of your equipment accepts these plugs. To connect to the two brass screws on the lightbulb's base, you'll need to stick alligator clips on the banana plugs.

Check your basic circuit with your instructor, then turn on the power switch and *slowly* turn up the knob until the bulb lights.²

Once you have your bulb lit, do not mess with the knob on the power supply anymore. You do not even need to switch the power supply off while rearranging the circuit for the two measurements with the meter; the voltage that lights the bulb is only about a volt or a volt and a half (similar to a battery), so it can't hurt you.

We have a single meter that plays both the role of the voltmeter and the role of the ammeter in this lab. Because it can do both these things, it is referred to as a multimeter. Multimeters are highly standardized, and the following instructions are generic ones that will work with

²On the power supplies we use at Fullerton College, the knob is uncalibrated and highly nonlinear; as you turn it up, the voltage it produces goes zerozerozerozerozerozerosix! To light the bulb without burning it out, you will need to find a position for the knob in the narrow range where it rapidly ramps up from 0 to 6 V.

whatever meters you happen to be using in this lab.

Voltage difference

Two wires connect the meter to the circuit. At the places where three wires come together at one point, you can plug a banana plug into the back of another banana plug. At the meter, make one connection at the “common” socket (“COM”) and the other at the socket labeled “V” for volts. The common plug is called that because it is used for every measurement, not just for voltage.

Many multimeters have more than one scale for measuring a given thing. For instance, a meter may have a millivolt scale and a volt scale. One is used for measuring small voltage differences and the other for large ones. You may not be sure in advance what scale is appropriate, but that’s not a big problem — once everything is hooked up, you can try different scales and see what’s appropriate. Use the switch or buttons on the front to select one of the voltage scales. By trial and error, find the most precise scale that doesn’t cause the meter to display an error message about being overloaded.

Write down your measurement, with the units of volts, and stop for a moment to think about what it is that you’ve measured. Imagine holding your breath and trying to make your eyeballs pop out with the pressure. Intuitively, the voltage difference is like the pressure difference between the inside and outside of your body.

What do you think will happen if you unscrew the bulb, leaving an air gap, while the power supply and the voltmeter are still going? Try it. Interpret your observation in terms of the breath-holding metaphor.

Current

The procedure for measuring the current differs only because you have to hook the meter up in series and because you have to use the “A” (amps) plug on the meter and select a current scale.

In the breath-holding metaphor, the number you’re measuring now is like the rate at which air flows through your lips as you let it hiss out. Based on this metaphor, what do you think will happen to the reading when you unscrew the bulb? Try it.

Discuss with your group and check with your instructor:

(1) What *goes through* the wires? Current? Voltage? Both?

(2) Using the breath-holding metaphor, explain why the voltmeter needs *two* connections to the circuit, not just one. What about the ammeter?

While waiting for your instructor to come around and discuss these questions with you, you can go on to the next part of the lab.

Resistance

The ratio of voltage difference to current is called the resistance of the bulb, $R = \Delta V/I$. Its units of volts per amp can be abbreviated as ohms, Ω (capital Greek letter omega).

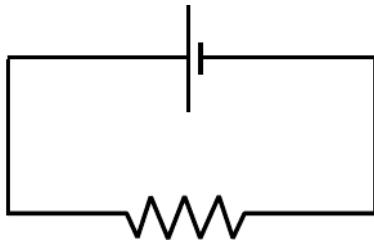
Calculate the resistance of your lightbulb. Resistance is the electrical equivalent of kinetic friction. Just as rubbing your hands together heats them up, objects that have electrical resistance produce heat when a current is passed through them. This is why the bulb’s filament gets hot enough to heat up.

When you unscrew the bulb, leaving an air gap, what is the resistance of the air?

Ohm’s law is a generalization about the electrical properties of a variety of materials. It states that the resistance is constant, i.e., that when you increase the voltage difference, the flow of current increases exactly in proportion. If you have time, test whether Ohm’s law holds for your lightbulb, by cutting the voltage to half of what you had before and checking whether the current drops by the same factor. (In this condition, the bulb’s filament doesn’t get hot enough to create enough visible light for your eye to see, but it does emit infrared light.)

Minilab 8: Electrical measurements

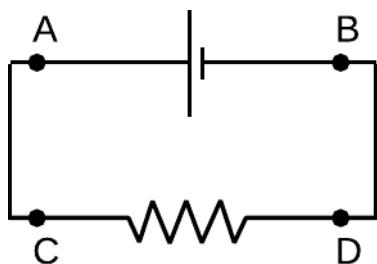
1. How many different currents could you measure in this circuit? Make a prediction, and then try it.



What do you notice? How does this make sense in terms of the roller coaster metaphor introduced in discussion question 8.2A on p. 195?

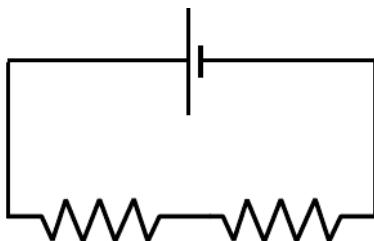
What is being *used up* in the resistor?

2. By connecting probes to these points, how many ways could you measure a voltage? How many of them would be different numbers? Make a prediction, and then do it.



What do you notice? Interpret this using the roller coaster metaphor, and color in parts of the circuit that represent constant voltages.

3. The resistors are unequal. How many *different* voltages and currents can you measure? Make a prediction, and then try it.



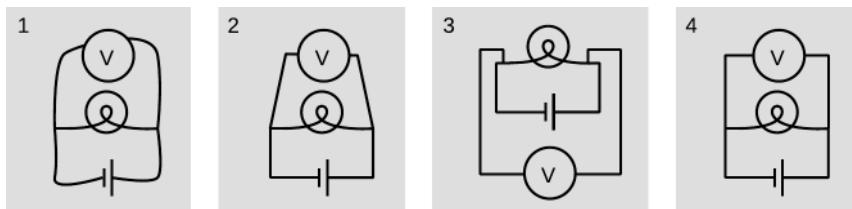
What do you notice? Interpret this using the roller coaster metaphor, and color in parts of the circuit that represent constant voltages.

Chapter 9

DC circuits

9.1 Schematics

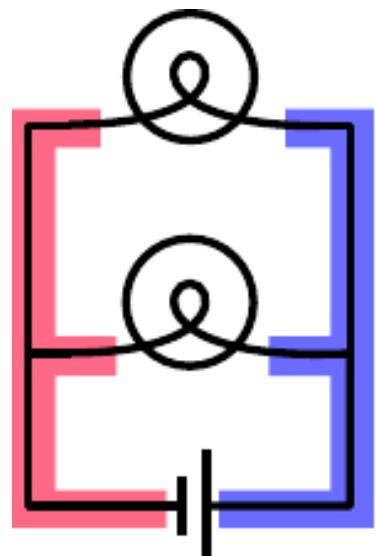
I see a chess position; Kasparov sees an interesting Ruy Lopez variation. To the uninitiated a schematic may look as unintelligible as Mayan hieroglyphs, but even a little bit of eye training can go a long way toward making its meaning leap off the page. A schematic is a stylized and simplified drawing of a circuit. The purpose is to eliminate as many irrelevant features as possible, so that the relevant ones are easier to pick out.



a / 1. Wrong: The shapes of the wires are irrelevant. 2. Wrong: Right angles should be used. 3. Wrong: A simple pattern is made to look unfamiliar and complicated. 4. Right.

An example of an irrelevant feature is the physical shape, length, and diameter of a wire. In nearly all circuits, it is a good approximation to assume that the wires are perfect conductors, so that any piece of wire uninterrupted by other components has constant potential throughout it. Changing the length of the wire, for instance, does not change this fact. (Of course if we used miles and miles of wire, as in a telephone line, the wire's resistance would start to add up, and its length would start to matter.) The shapes of the wires are likewise irrelevant, so we draw them with standardized, stylized shapes made only of vertical and horizontal lines with right-angle bends in them. This has the effect of making similar circuits look more alike and helping us to recognize familiar patterns, just as words in a newspaper are easier to recognize than handwritten ones. Figure a shows some examples of these concepts.

The most important first step in learning to read schematics is to learn to recognize contiguous pieces of wire which must have constant potential throughout. In figure b, for example, the two shaded E-shaped pieces of wire must each have constant potential. This focuses our attention on two of the main unknowns we'd like to be able to predict: the potential of the left-hand E and the potential of the one on the right.



b / The two shaded areas shaped like the letter "E" are both regions of constant potential.

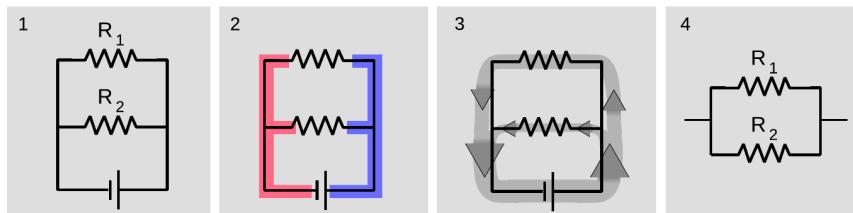
9.2 Parallel resistances and the junction rule

One of the simplest examples to analyze is the parallel resistance circuit, of which figure b was an example. In general we may have unequal resistances R_1 and R_2 , as in c/1. Since there are only two constant-potential areas in the circuit, c/2, all three components have the same potential difference across them. A battery normally succeeds in maintaining the potential differences across itself for which it was designed, so the voltage drops ΔV_1 and ΔV_2 across the resistors must both equal the voltage of the battery:

$$\Delta V_1 = \Delta V_2 = \Delta V_{battery}.$$

Each resistance thus feels the same potential difference as if it was the only one in the circuit, and Ohm's law tells us that the amount of current flowing through each one is also the same as it would have been in a one-resistor circuit. This is why household electrical circuits are wired in parallel. We want every appliance to work the same, regardless of whether other appliances are plugged in or unplugged, turned on or switched off. (The electric company doesn't use batteries of course, but our analysis would be the same for any device that maintains a constant voltage.)

- c / 1. Two resistors in parallel. 2. There are two constant-potential areas. 3. The current that comes out of the battery splits between the two resistors, and later reunites. 4. The two resistors in parallel can be treated as a single resistor with a smaller resistance value.



Of course the electric company can tell when we turn on every light in the house. How do they know? The answer is that we draw more current. Each resistance draws a certain amount of current, and the amount that has to be supplied is the sum of the two individual currents. The current is like a river that splits in half, c/3, and then reunites. The total current is

$$I_{total} = I_1 + I_2.$$

This is an example of a general fact called the junction rule:

the junction rule

In any circuit that is not storing or releasing charge, conservation of charge implies that the total current flowing out of any junction must be the same as the total flowing in.

Coming back to the analysis of our circuit, we apply Ohm's law to each resistance, resulting in

$$\begin{aligned} I_{total} &= \Delta V/R_1 + \Delta V/R_2 \\ &= \Delta V \left(\frac{1}{R_1} + \frac{1}{R_2} \right). \end{aligned}$$

As far as the electric company is concerned, your whole house is just one resistor with some resistance R , called the *equivalent resistance*. They would write Ohm's law as

$$I_{total} = \Delta V/R,$$

from which we can determine the equivalent resistance by comparison with the previous expression:

$$\begin{aligned} 1/R &= \frac{1}{R_1} + \frac{1}{R_2} \\ R &= \left(\frac{1}{R_1} + \frac{1}{R_2} \right)^{-1} \end{aligned}$$

[equivalent resistance of two resistors in parallel]

Two resistors in parallel, $c/4$, are equivalent to a single resistor with a value given by the above equation.

Two lamps on the same household circuit *example 1*

▷ You turn on two lamps that are on the same household circuit. Each one has a resistance of 1 ohm. What is the equivalent resistance, and how does the power dissipation compare with the case of a single lamp?

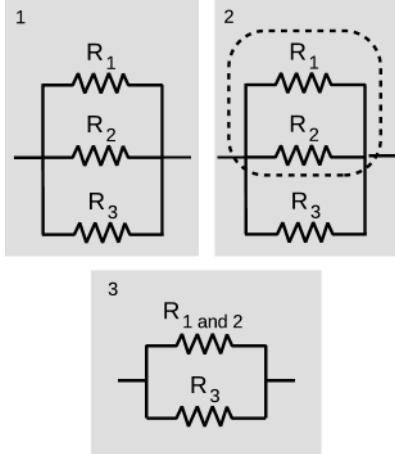
▷ The equivalent resistance of the two lamps in parallel is

$$\begin{aligned} R &= \left(\frac{1}{R_1} + \frac{1}{R_2} \right)^{-1} \\ &= \left(\frac{1}{1\ \Omega} + \frac{1}{1\ \Omega} \right)^{-1} \\ &= \left(1\ \Omega^{-1} + 1\ \Omega^{-1} \right)^{-1} \\ &= \left(2\ \Omega^{-1} \right)^{-1} \\ &= 0.5\ \Omega \end{aligned}$$

The potential difference across the whole circuit is always the 110 V set by the electric company (it's alternating current, but that's irrelevant). The resistance of the whole circuit has been cut in half by turning on the second lamp, so a fixed amount of voltage will produce twice as much current. Twice the current flowing across the same potential difference means twice as much power dissipation, which makes sense.

The cutting in half of the resistance surprises many students, since we are “adding more resistance” to the circuit by putting in the second lamp. Why does the equivalent resistance come out to be less than the resistance of a single lamp? This is a case where purely verbal reasoning can be misleading. A resistive circuit element, such as the filament of a lightbulb, is neither a perfect insulator nor a perfect conductor. Instead of analyzing this type of circuit in terms of “resistors,” i.e., partial insulators, we could have spoken of “conductors.” This example would then seem reasonable, since we “added more conductance,” but one would then have the incorrect expectation about the case of resistors in series, discussed in the following section.

Perhaps a more productive way of thinking about it is to use mechanical intuition. By analogy, your nostrils resist the flow of air through them, but having two nostrils makes it twice as easy to breathe.



Three resistors in parallel

example 2

- ▷ What happens if we have three or more resistors in parallel?
- ▷ This is an important example, because the solution involves an important technique for understanding circuits: breaking them down into smaller parts and then simplifying those parts. In the circuit 9.2/1, with three resistors in parallel, we can think of two of the resistors as forming a single resistor, 9.2/2, with equivalent resistance

$$R_{12} = \left(\frac{1}{R_1} + \frac{1}{R_2} \right)^{-1}.$$

We can then simplify the circuit as shown in 9.2/3, so that it contains only two resistances. The equivalent resistance of the whole circuit is then given by

$$R_{123} = \left(\frac{1}{R_{12}} + \frac{1}{R_3} \right)^{-1}.$$

Substituting for R_{12} and simplifying, we find the result

$$R_{123} = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \right)^{-1},$$

which you probably could have guessed. The interesting point here is the divide-and-conquer concept, not the mathematical result.

Example 2.

An arbitrary number of identical resistors in parallel example 3

- ▷ What is the resistance of N identical resistors in parallel?
- ▷ Generalizing the results for two and three resistors, we have

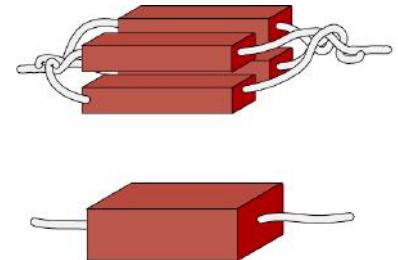
$$R_N = \left(\frac{1}{R_1} + \frac{1}{R_2} + \dots \right)^{-1},$$

where “...” means that the sum includes all the resistors. If all the resistors are identical, this becomes

$$\begin{aligned} R_N &= \left(\frac{N}{R} \right)^{-1} \\ &= \frac{R}{N} \end{aligned}$$

Dependence of resistance on cross-sectional area example 4

We have alluded briefly to the fact that an object's electrical resistance depends on its size and shape, but now we are ready to begin making more mathematical statements about it. As suggested by figure 4, increasing a resistor's cross-sectional area is equivalent to adding more resistors in parallel, which will lead to an overall decrease in resistance. Any real resistor with straight, parallel sides can be sliced up into a large number of pieces, each with cross-sectional area of, say, $1 \mu\text{m}^2$. The number, N , of such slices is proportional to the total cross-sectional area of the resistor, and by application of the result of the previous example we therefore find that the resistance of an object is inversely proportional to its cross-sectional area.



Example 4: Uniting four resistors in parallel is equivalent to making a single resistor with the same length but four times the cross-sectional area. The result is to make a resistor with one quarter the resistance.



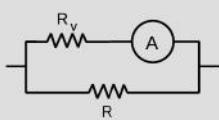
A fat pipe has less resistance than a skinny pipe.

An analogous relationship holds for water pipes, which is why high-flow trunk lines have to have large cross-sectional areas. To make lots of water (current) flow through a skinny pipe, we'd need an impractically large pressure (voltage) difference.

1



2



e / A voltmeter is really an ammeter with an internal resistor. When we measure the voltage difference across a resistor, 1, we are really constructing a parallel resistance circuit, 2.

Incorrect readings from a voltmeter

example 5

A voltmeter is really just an ammeter with an internal resistor, and we use a voltmeter in parallel with the thing that we're trying to measure the potential difference across. This means that any time we measure the voltage drop across a resistor, we're essentially putting two resistors in parallel. The ammeter inside the voltmeter can be ignored for the purpose of analyzing how current flows in the circuit, since it is essentially just some coiled-up wire with a very low resistance.

Now if we are carrying out this measurement on a resistor that is part of a larger circuit, we have changed the behavior of the circuit through our act of measuring. It is as though we had modified the circuit by replacing the resistance R with the smaller equivalent resistance of R and R_v in parallel. It is for this reason that voltmeters are built with the largest possible internal resistance. As a numerical example, if we use a voltmeter with an internal resistance of $1 \text{ M}\Omega$ to measure the voltage drop across a one-ohm resistor, the equivalent resistance is 0.999999Ω , which is not different enough to make any difference. But if we tried to use the same voltmeter to measure the voltage drop across a $2 \text{ M}\Omega$ resistor, we would be reducing the resistance of that part of the circuit by a factor of three, which would produce a drastic change in the behavior of the whole circuit.

This is the reason why you can't use a voltmeter to measure the potential difference between two different points in mid-air, or between the ends of a piece of wood. This is by no means a stupid thing to want to do, since the world around us is not a constant-potential environment, the most extreme example being when an electrical storm is brewing. But it will not work with an ordinary voltmeter because the resistance of the air or the wood is many gigaohms. The effect of waving a pair of voltmeter probes around in the air is that we provide a reuniting path for the positive and negative charges that have been separated — through the voltmeter itself, which is a good conductor compared to the air. This reduces to zero the potential difference we were trying to measure.

In general, a voltmeter that has been set up with an open circuit (or a very large resistance) between its probes is said to be "floating." An old-fashioned analog voltmeter of the type described here will read zero when left floating, the same as when it was sitting on the shelf. A floating digital voltmeter usually shows an error message.

9.3 Series resistances and the loop rule

The two basic circuit layouts are parallel and series, so a pair of resistors in series, f/1, is another of the most basic circuits we can make. By conservation of charge, all the current that flows through one resistor must also flow through the other (as well as through the battery):

$$I_1 = I_2.$$

The only way the information about the two resistance values is going to be useful is if we can apply Ohm's law, which will relate the resistance of each resistor to the current flowing through it and the voltage difference across it. Figure f/2 shows the three constant-potential areas. Voltage differences are more physically significant than voltages, so we define symbols for the voltage differences across the two resistors in figure f/3.

We have three constant-potential areas, with symbols for the difference in voltage between every possible pair of them. These three potential differences must be related to each other. It is as though I tell you that Fred is a foot taller than Ginger, Ginger is a foot taller than Sally, and Fred is two feet taller than Sally. The information is redundant, and you really only needed two of the three pieces of data to infer the third. In the case of our voltage differences, we have

$$|\Delta V_1| + |\Delta V_2| = |\Delta V_{battery}|.$$

The absolute value signs are because of the ambiguity in how we define our voltage differences. If we reversed the two probes of the voltmeter, we would get a result with the opposite sign. Digital voltmeters will actually provide a minus sign on the screen if the wire connected to the "V" plug is lower in potential than the one connected to the "COM" plug. Analog voltmeters pin the needle against a peg if you try to use them to measure negative voltages, so you have to fiddle to get the leads connected the right way, and then supply any necessary minus sign yourself.

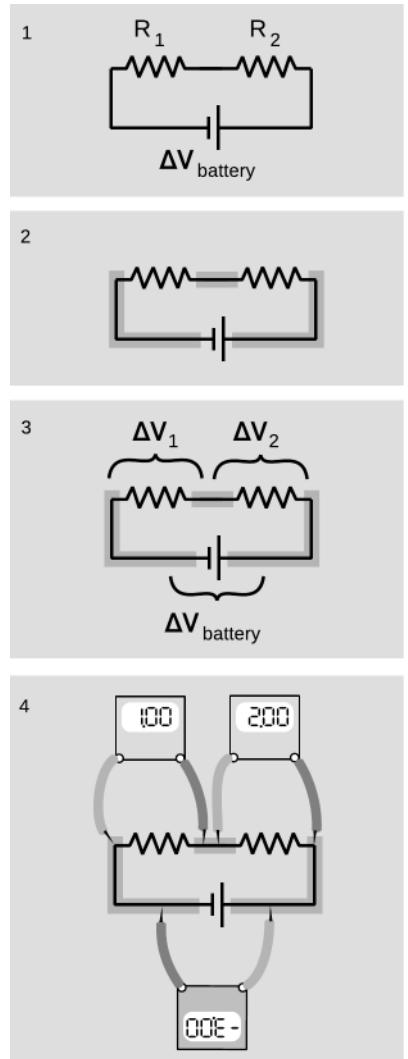
Figure f/4 shows a standard way of taking care of the ambiguity in signs. For each of the three voltage measurements around the loop, we keep the same probe (the darker one) on the clockwise side. It is as though the voltmeter was sidling around the circuit like a crab, without ever "crossing its legs." With this convention, the relationship among the voltage drops becomes

$$\Delta V_1 + \Delta V_2 = -\Delta V_{battery},$$

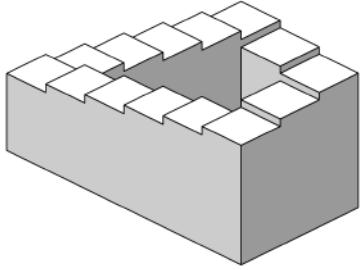
or, in more symmetrical form,

$$\Delta V_1 + \Delta V_2 + \Delta V_{battery} = 0.$$

More generally, this is known as the loop rule for analyzing circuits:



- f / 1. A battery drives current through two resistors in series.
 2. There are three constant-potential regions. 3. The three voltage differences are related.
 4. If the meter crab-walks around the circuit without flipping over or crossing its legs, the resulting voltages have plus and minus signs that make them add up to zero.



g / An impossible staircase.

the loop rule

Assuming the standard convention for plus and minus signs, the sum of the voltage drops around any closed loop in a DC circuit must be zero.

The loop rule and junction rule are credited to Gustav Kirchoff and are therefore often referred to by names such as Kirchoff's rules, Kirchoff's loop rule, Kirchoff's junction rule, and acronyms such as KCL (for current) and KVL (for voltage).

In a DC circuit, we have a well-defined electric potential, which is analogous to height in a gravitational field. Looking for an exception to the loop rule would be like asking for a hike that would be downhill all the way and that would come back to its starting point, or a staircase like the one in figure g. The sum of voltage drops described in the loop rule is the work done per unit charge as we bring a charge around the loop and back to its starting point. If this work were nonzero for a static field, then as argued in sec. 4.1, p. 85, we would have a perpetual motion machine, violating conservation of energy. The assumption of a DC circuit (static field) is necessary, and the loop rule can be violated when the fields are time-varying ([222](#)).

For the circuit we set out to analyze, the equation

$$\Delta V_1 + \Delta V_2 + \Delta V_{battery} = 0$$

can now be rewritten by applying Ohm's law to each resistor:

$$I_1 R_1 + I_2 R_2 + \Delta V_{battery} = 0.$$

The currents are the same, so we can factor them out:

$$I (R_1 + R_2) + \Delta V_{battery} = 0,$$

and this is the same result we would have gotten if we had been analyzing a one-resistor circuit with resistance $R_1 + R_2$. Thus the equivalent resistance of resistors in series equals the sum of their resistances.

Two lightbulbs in series

example 6

- ▷ If two identical lightbulbs are placed in series, how do their brightnesses compare with the brightness of a single bulb?
- ▷ Taken as a whole, the pair of bulbs act like a doubled resistance, so they will draw half as much current from the wall. Each bulb will be dimmer than a single bulb would have been.

The total power dissipated by the circuit is $I\Delta V$. The voltage drop across the whole circuit is the same as before, but the current is halved, so the two-bulb circuit draws half as much total power as

the one-bulb circuit. Each bulb draws one-quarter of the normal power.

Roughly speaking, we might expect this to result in one quarter the light being produced by each bulb, but in reality lightbulbs waste quite a high percentage of their power in the form of heat and wavelengths of light that are not visible (infrared and ultraviolet). Less light will be produced, but it's hard to predict exactly how much less, since the efficiency of the bulbs will be changed by operating them under different conditions.

More than two equal resistances in series *example 7*

By straightforward application of the divide-and-conquer technique discussed in the previous section, we find that the equivalent resistance of N identical resistances R in series will be NR .

Dependence of resistance on length *example 8*

In the previous section, we proved that resistance is inversely proportional to cross-sectional area. By equivalent reason about resistances in series, we find that resistance is proportional to length. Analogously, it is harder to blow through a long straw than through a short one.

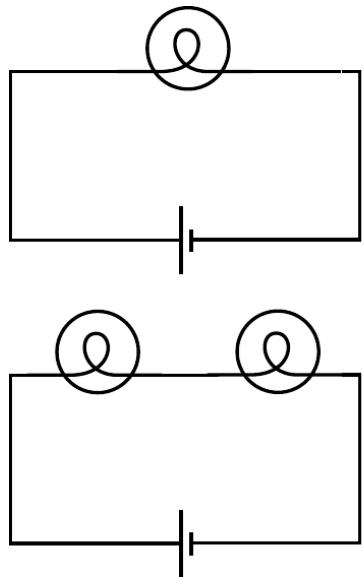
Combining the results of examples 4 and 8, we find that the resistance of an object with straight, parallel sides is given by

$$R = (\text{constant}) \cdot L/A$$

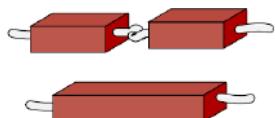
The proportionality constant is called the resistivity, and it depends only on the substance of which the object is made. A resistivity measurement could be used, for instance, to help identify a sample of an unknown substance.

Choice of high voltage for power lines *example 9*

Thomas Edison got involved in a famous technological controversy over the voltage difference that should be used for electrical power lines. At this time, the public was unfamiliar with electricity, and easily scared by it. The president of the United States, for instance, refused to have electrical lighting in the White House when it first became commercially available because he considered it unsafe, preferring the known fire hazard of oil lamps to the mysterious dangers of electricity. Mainly as a way to overcome public fear, Edison believed that power should be transmitted using small voltages, and he publicized his opinion by giving demonstrations at which a dog was lured into position to be killed by a large voltage difference between two sheets of metal on the ground. (Edison's opponents also advocated alternating current rather than direct current, and AC is more dangerous than DC as well. As we will discuss later, AC can be easily stepped up and down to the desired voltage level using a device called a transformer.)



Example 6.



Example 8. Doubling the length of a resistor is like putting two resistors in series. The resistance is doubled.

Now if we want to deliver a certain amount of power P_L to a load such as an electric lightbulb, we are constrained only by the equation $P_L = I\Delta V_L$. We can deliver any amount of power we wish, even with a low voltage, if we are willing to use large currents. Modern electrical distribution networks, however, use dangerously high voltage differences of tens of thousands of volts. Why did Edison lose the debate?

It boils down to money. The electric company must deliver the amount of power P_L desired by the customer through a transmission line whose resistance R_T is fixed by economics and geography. The same current flows through both the load and the transmission line, dissipating power usefully in the former and wastefully in the latter. The efficiency of the system is

$$\begin{aligned}\text{efficiency} &= \frac{\text{power paid for by the customer}}{\text{power paid for by the utility}} \\ &= \frac{P_L}{P_L + P_T} \\ &= \frac{1}{1 + P_T/P_L}\end{aligned}$$

Putting ourselves in the shoes of the electric company, we wish to get rid of the variable P_T , since it is something we control only indirectly by our choice of ΔV_T and I . Substituting $P_T = I\Delta V_T$, we find

$$\text{efficiency} = \frac{1}{1 + \frac{I\Delta V_T}{P_L}}$$

We assume the transmission line (but not necessarily the load) is ohmic, so substituting $\Delta V_T = IR_T$ gives

$$\text{efficiency} = \frac{1}{1 + \frac{I^2 R_T}{P_L}}$$

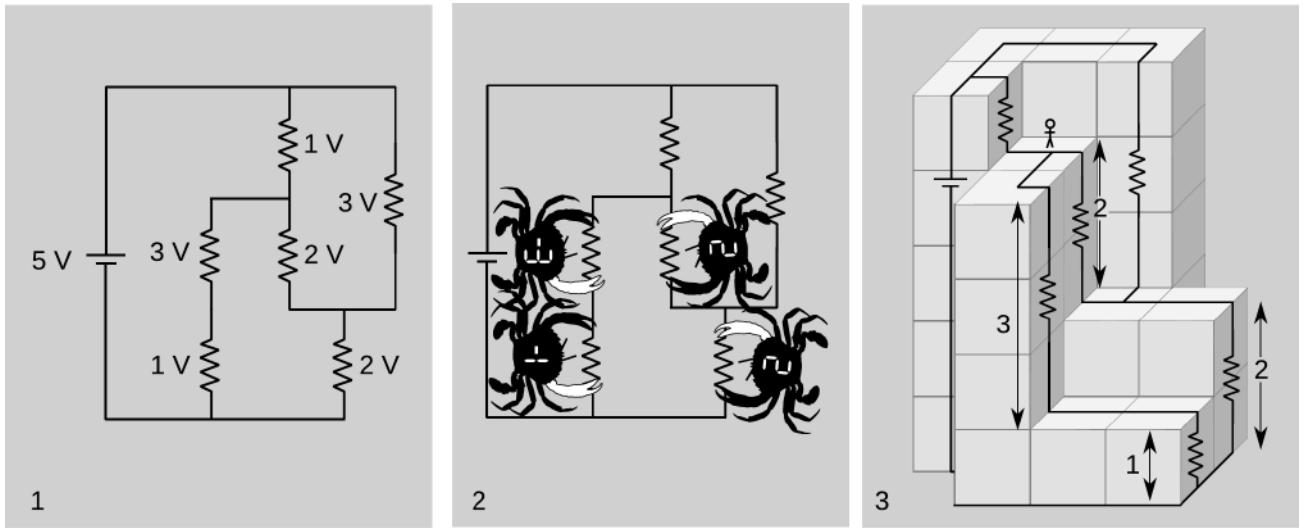
This quantity can clearly be maximized by making I as small as possible, since we will then be dividing by the smallest possible quantity on the bottom of the fraction. A low-current circuit can only deliver significant amounts of power if it uses high voltages, which is why electrical transmission systems use dangerous high voltages.

Two ways of handling signs

example 10

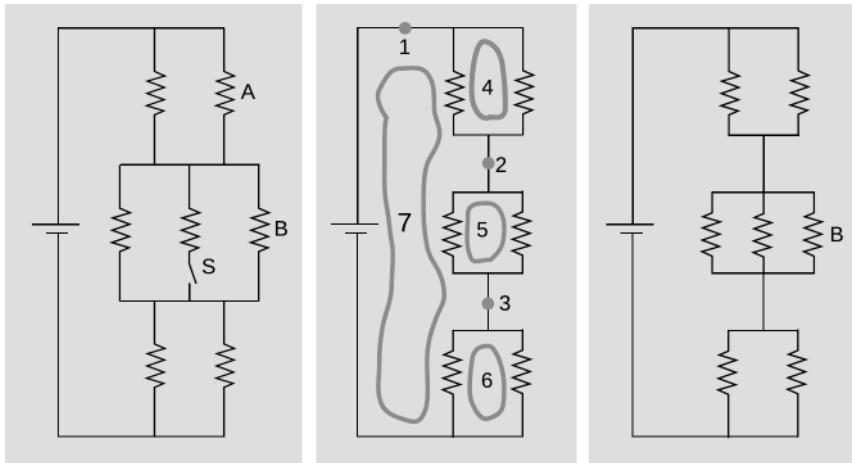
The figure above shows two ways of visualizing the loop rule and handling the signs involved. In panel 1, each circuit element is labeled with the voltage drop across it.

In 2, the crab is a voltmeter whose reading is the potential on the white claw minus the potential on the black claw. The crab can't flip over. It can only scuttle sideways as it moves around the loop



that we've chosen, consisting of four resistors. The sum of the four readings is zero.

Panel 3 shows a visualization of the same circuit in which potential is like height. The stick figure on the ledge wants to get down to the ground by doing a series of hops. He has two ways: do the 3 V drop and then the 1 V drop, or do the 2 V and the other 2 V. Here we treat all the voltage differences as positive numbers. This method works nicely if you're pretty sure for each resistor in the circuit which end is the higher voltage.



Example 11.

A complicated circuit

example 11

- ▷ All seven resistors in the left-hand panel of figure h are identical. Initially, the switch S is open as shown in the figure, and the current through resistor A is I_0 . The switch is then closed. Find the current through resistor B, after the switch is closed, in terms

of I_0 .

► The second panel shows the circuit redrawn for simplicity, in the initial condition with the switch open. When the switch is open, no current can flow through the central resistor, so we may as well ignore it. I've also redrawn the junctions, without changing what's connected to what. This is the kind of mental rearranging that you'll eventually learn to do automatically from experience with analyzing circuits. The redrawn version makes it easier to see what's happening with the current. Charge is conserved, so any charge that flows past point 1 in the circuit must also flow past points 2 and 3. This would have been harder to reason about by applying the junction rule to the original version, which appears to have nine separate junctions.

In the new version, it's also clear that the circuit has a great deal of symmetry. We could flip over each parallel pair of identical resistors without changing what's connected to what, so that makes it clear that the voltage drops and currents must be equal for the members of each pair. We can also prove this by using the loop rule. The loop rule says that the two voltage drops in loop 4 must be equal, and similarly for loops 5 and 6. Since the resistors obey Ohm's law, equal voltage drops across them also imply equal currents. That means that when the current at point 1 comes to the top junction, exactly half of it goes through each resistor. Then the current reunites at 2, splits between the next pair, and so on. We conclude that each of the six resistors in the circuit experiences the same voltage drop and the same current. Applying the loop rule to loop 7, we find that the sum of the three voltage drops across the three left-hand resistors equals the battery's voltage, V , so each resistor in the circuit experiences a voltage drop $V/3$. Letting R stand for the resistance of one of the resistors, we find that the current through resistor B, which is the same as the currents through all the others, is given by $I_0 = V/3R$.

We now pass to the case where the switch is closed, as shown in the third panel. The battery's voltage is the same as before, and each resistor's resistance is the same, so we can still use the same symbols V and R for them. It is no longer true, however, that each resistor feels a voltage drop $V/3$. The equivalent resistance of the whole circuit is $R/2 + R/3 + R/2 = 4R/3$, so the total current drawn from the battery is $3V/4R$. In the middle group of resistors, this current is split three ways, so the new current through B is $(1/3)(3V/4R) = V/4R = 3I_0/4$.

Interpreting this result, we see that it comes from two effects that partially cancel. Closing the switch reduces the equivalent resistance of the circuit by giving charge another way to flow, and increases the amount of current drawn from the battery. Resistor B, however, only gets a 1/3 share of this greater current, not 1/2.

The second effect turns out to be bigger than first, and therefore the current through resistor B is lessened over all.

Getting killed by your ammeter

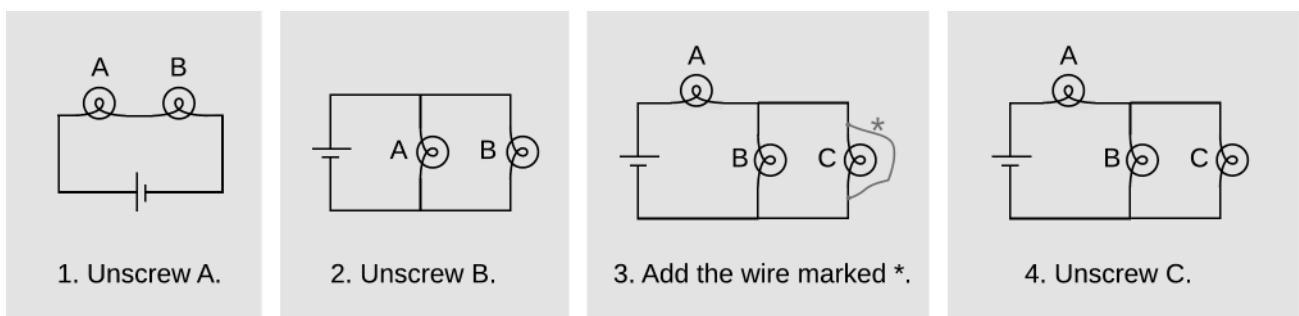
example 12

As with a voltmeter, an ammeter can give erroneous readings if it is used in such a way that it changes the behavior of the circuit. An ammeter is used in series, so if it is used to measure the current through a resistor, the resistor's value will effectively be changed to $R + R_a$, where R_a is the resistance of the ammeter. Ammeters are designed with very low resistances in order to make it unlikely that $R + R_a$ will be significantly different from R .

In fact, the real hazard is death, not a wrong reading! Virtually the only circuits whose resistances are significantly less than that of an ammeter are those designed to carry huge currents. An ammeter inserted in such a circuit can easily melt. When I was working at a laboratory funded by the Department of Energy, we got periodic bulletins from the DOE safety office about serious accidents at other sites, and they held a certain ghoulish fascination. One of these was about a DOE worker who was completely incinerated by the explosion created when he inserted an ordinary Radio Shack ammeter into a high-current circuit. Later estimates showed that the heat was probably so intense that the explosion was a ball of plasma — a gas so hot that its atoms have been ionized.

Discussion questions

A We have stated the loop rule in a symmetric form where a series of voltage drops adds up to zero. To do this, we had to define a standard way of connecting the voltmeter to the circuit so that the plus and minus signs would come out right. Suppose we wish to restate the junction rule in a similar symmetric way, so that instead of equating the current coming in to the current going out, it simply states that a certain sum of currents at a junction adds up to zero. What standard way of inserting the ammeter would we have to use to make this work?



B The lightbulbs are all identical. In each case, a change is proposed to make to the circuit. Predict the change in brightness of each bulb.

Notes for chapter 9

[216](#) Loop rule can be violated by AC circuits

The loop rule is valid only for static fields (DC circuits), and can be violated otherwise.

To see why the assumption of a DC circuit is necessary, consider what happens in an antenna, as in figure o on p. 164. In such an AC circuit, a charge can oscillate back and forth repeatedly, having negative work done on it in each cycle. This is not a violation of conservation of energy, because the energy is being pumped out into an electromagnetic wave. The electric field in the antenna's radiation pattern is curly (which is allowed by Maxwell's equations because $\text{curl } \mathbf{E} = -\partial \mathbf{B} / \partial t$), so the work done by the field is path-dependent, and the construction of the electric potential in 4.2, p. 89, breaks down. In the gravitational metaphor, there is no way to even define a notion of height.

However, it still often happens that the loop rule is approximately valid, even for AC circuits ([2324](#)).

Problems

Key

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

1 (a) Many battery-operated devices take more than one battery. If you look closely in the battery compartment, you will see that the batteries are wired in series. Consider a flashlight circuit. What does the loop rule tell you about the effect of putting several batteries in series in this way?

(b) The cells of an electric eel's nervous system are not that different from ours — each cell can develop a voltage difference across it of somewhere on the order of one volt. How, then, do you think an electric eel can create voltages of thousands of volts between different parts of its body?

2 You have a circuit consisting of two unknown resistors in series, and a second circuit consisting of two unknown resistors in parallel.

- (a) What, if anything, would you learn about the resistors in the series circuit by finding that the currents through them were equal?
- (b) What if you found out the voltage differences across the resistors in the series circuit were equal?

(c) What would you learn about the resistors in the parallel circuit from knowing that the currents were equal?

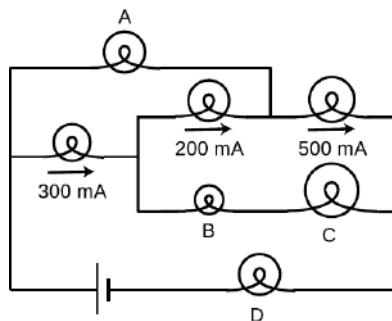
(d) What if the voltages in the parallel circuit were equal?

3 The bulbs all have unequal resistances. Given the three currents shown in the figure, find the currents through bulbs A, B, C, and D. ✓

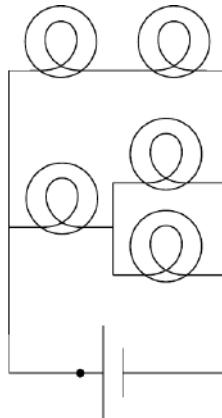
4 The figure shows a circuit containing five lightbulbs connected to a battery. Suppose you're going to connect one probe of a voltmeter to the circuit at the point marked with a dot. How many unique, nonzero voltage differences could you measure by connecting the other probe to other wires in the circuit?

5 The lightbulbs in the figure are all identical. If you were inserting an ammeter at various places in the circuit, how many unique currents could you measure? If you know that the current measurement will give the same number in more than one place, only count that as one unique current.

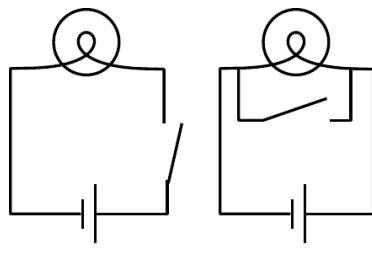
6 The figure shows two possible ways of wiring a flashlight with a switch. Both will serve to turn the bulb on and off, although the switch functions in the opposite sense. Why is method 1 preferable?



Problem 3.



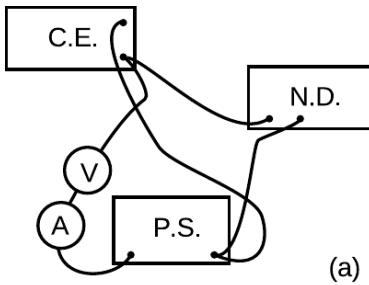
Problems 4 and 5.



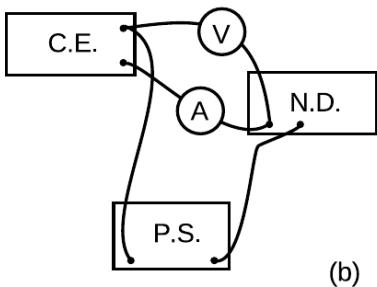
1

2

Problem 6.

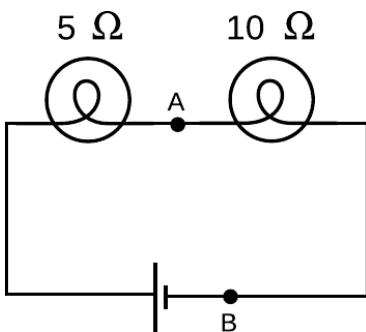


(a)



(b)

Problem 7.



Problem 9.

- 7 A student in a biology lab is given the following instructions: “Connect the cerebral eraser (C.E.) and the neural depolarizer (N.D.) in parallel with the power supply (P.S.). (Under no circumstances should you ever allow the cerebral eraser to come within 20 cm of your head.) Connect a voltmeter to measure the voltage across the cerebral eraser, and also insert an ammeter in the circuit so that you can make sure you don’t put more than 100 mA through the neural depolarizer.” The diagrams show two lab groups’ attempts to follow the instructions. (a) Translate diagram a into a standard-style schematic. What is correct and incorrect about this group’s setup? (b) Do the same for diagram b.

- 8 A $1.0\ \Omega$ toaster and a $2.0\ \Omega$ lamp are connected in parallel with the 110-V supply of your house. (Ignore the fact that the voltage is AC rather than DC.)

- (a) Draw a schematic of the circuit.
 (b) For each of the three components in the circuit, find the current passing through it and the voltage drop across it. ✓
 (c) Suppose they were instead hooked up in series. Draw a schematic and calculate the same things. ✓

- 9 In the figure, the battery is 9 V.

- (a) What are the voltage differences across each light bulb? ✓
 (b) What current flows through each of the three components of the circuit? ✓
 (c) If a new wire is added to connect points A and B, how will the appearances of the bulbs change? What will be the new voltages and currents?
 (d) Suppose no wire is connected from A to B, but the two bulbs are switched. How will the results compare with the results from the original setup as drawn?

- 10 What resistance values can be created by combining a $1\ k\Omega$ resistor and a $10\ k\Omega$ resistor? ▷ Solution, p. 429

- 11 How many different resistance values can be created by combining three unequal resistors? (Don’t count possibilities where not all the resistors are used.)

- 12 Wire is sold in a series of standard diameters, called “gauges.” The difference in diameter between one gauge and the next in the series is about 20%. How would the resistance of a given length of wire compare with the resistance of the same length of wire in the next gauge in the series? ✓

13 It's fairly common in electrical circuits for additional, undesirable resistances to occur because of factors such as dirty, corroded, or loose connections. Suppose that a device with resistance R normally dissipates power P , but due to an additional series resistance r the *total* power is reduced to P' . We might, for example, detect this change because the battery powering our device ran down more slowly than normal.

- (a) Find the unknown resistance r . ✓
- (b) Check that the units of your result make sense.
- (c) Check that your result makes sense in the special cases $P' = P$ and $P' = 0$.
- (d) Suppose we redefine P' as the useful power dissipated in R . For example, this would be the change we would notice because a flashlight was dimmer. Find r . ✓

14 A person in a rural area who has no electricity runs an extremely long extension cord to a friend's house down the road so she can run an electric light. The cord is so long that its resistance, x , is not negligible. Show that the lamp's brightness is greatest if its resistance, y , is equal to x . Explain physically why the lamp is dim for values of y that are too small or too large.

15 Suppose six identical resistors, each with resistance R , are connected so that they form the edges of a tetrahedron (a pyramid with three sides in addition to the base, i.e., one less side than an Egyptian pyramid). What resistance value or values can be obtained by making connections onto any two points on this arrangement?

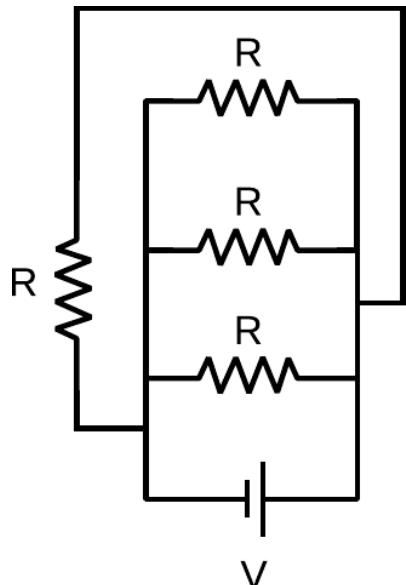
▷ Solution, p. 429

16 Find the current drawn from the battery. ✓

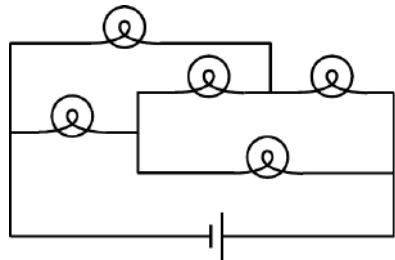
17 The bulbs are all identical. Which one doesn't light up?

18 The heating element of an electric stove is connected in series with a switch that opens and closes many times per second. When you turn the knob up for more power, the fraction of the time that the switch is closed increases. Suppose someone suggests a simpler alternative for controlling the power by putting the heating element in series with a variable resistor controlled by the knob. (With the knob turned all the way clockwise, the variable resistor's resistance is nearly zero, and when it's all the way counterclockwise, its resistance is essentially infinite.) (a) Draw schematics. (b) Why would the simpler design be undesirable? ★

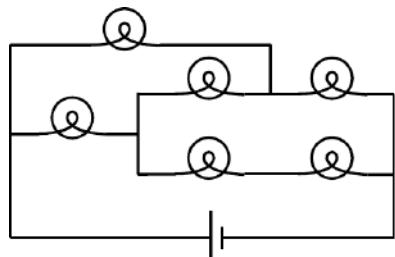
19 Each bulb has a resistance of one ohm. How much power is drawn from the one-volt battery? ✓ ★



Problem 16.



Problem 17.



Problem 19.

Lab 9: Voltage and current

This exercise is based on one created by Virginia Roundy.

Apparatus:

DC power supply

1.5 volt batteries

lightbulbs and holders

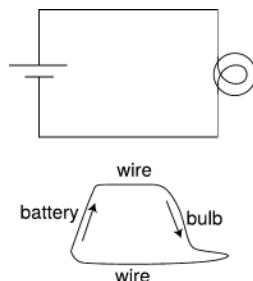
wire

highlighting pens, 3 colors

When you first glance at this exercise, it may look scary and intimidating — all those circuits! However, all those wild-looking circuits can be analyzed using the following four guides to thinking:

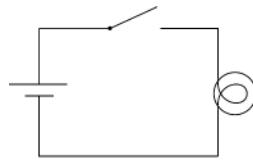
1. *A circuit has to be complete*, i.e., it must be possible for charge to get recycled as it goes around the circuit. If it's not complete, then charge will build up at a dead end. This built-up charge will repel any other charge that tries to get in, and everything will rapidly grind to a stop.
2. *There is constant voltage everywhere along a piece of wire*. To apply this rule during this lab, I suggest you use the colored highlighting pens to mark the circuit. For instance, if there's one whole piece of the circuit that's all at the same voltage, you could highlight it in yellow. A second piece of the circuit, at some other voltage, could be highlighted in blue.
3. *Charge is conserved*, so charge can't "get used up."

4. You can draw a *rollercoaster diagram*, like the one shown below. On this kind of diagram, height corresponds to voltage — that's why the wires are drawn as horizontal tracks.



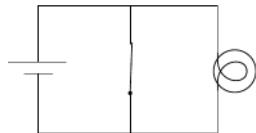
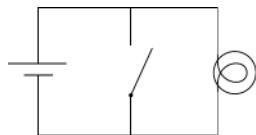
A Bulb and a Switch

Look at circuit 1, and try to predict what will happen when the switch is open, and what will happen when it's closed. Write both your predictions in the table on the following page before you build the circuit. When you build the circuit, you don't need an actual switch like a light switch; just connect and disconnect the banana plugs. Use one of the 1.5 volt batteries as your voltage source.



Circuit 1

	<i>switch open</i>
prediction	
explanation	
observation	



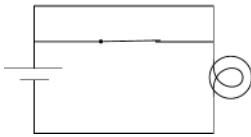
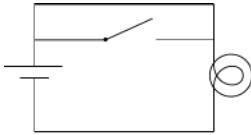
Circuit 2 (Don't leave the switch closed for a long time!)

	<i>switch closed</i>
prediction	
explanation	
observation	
explanation (if different)	

	<i>switch open</i>
prediction	
explanation	
observation	
explanation (if different)	

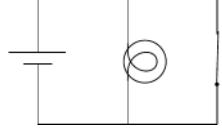
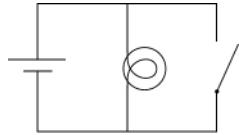
	<i>switch closed</i>
prediction	
explanation	
observation	
explanation (if different)	

Did it work the way you expected? If not, try to figure it out with the benefit of hindsight, and write your explanation in the table above.



Circuit 3

	<i>switch open</i>
<i>prediction</i>	
<i>explanation</i>	
<i>observation</i>	
<i>explanation (if different)</i>	



Circuit 4

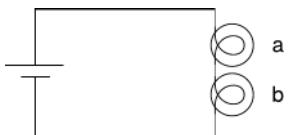
	<i>switch open</i>
<i>prediction</i>	
<i>explanation</i>	
<i>observation</i>	
<i>explanation (if different)</i>	

	<i>switch closed</i>
<i>prediction</i>	
<i>explanation</i>	
<i>observation</i>	
<i>explanation (if different)</i>	

	<i>switch closed</i>
<i>prediction</i>	
<i>explanation</i>	
<i>observation</i>	
<i>explanation (if different)</i>	

Two Bulbs

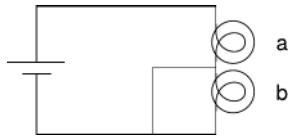
Instead of a battery, use the DC power supply, set to 2.4 volts, for circuits 5 and 6. Analyze this one both by highlighting and by drawing a rollercoaster diagram.



Circuit 5

	<i>bulb a</i>
prediction	
explanation	
observation	
explanation (if different)	

	<i>bulb b</i>
prediction	
explanation	
observation	
explanation (if different)	



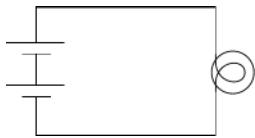
Circuit 6

	<i>bulb a</i>
prediction	
explanation	
observation	
explanation (if different)	

	<i>bulb b</i>
prediction	
explanation	
observation	
explanation (if different)	

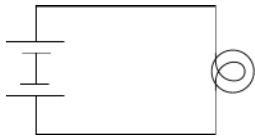
Two Batteries

Use batteries for circuits 7-9. Circuits 7 and 8 are both good candidates for rollercoaster diagrams.



Circuit 7

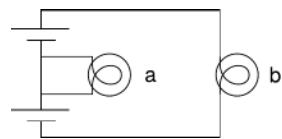
prediction	
explanation	
observation	
explanation (if different)	



Circuit 8

prediction	
explanation	
observation	
explanation (if different)	

A Final Challenge



Circuit 9

	<i>bulb a</i>
prediction	
explanation	
observation	
explanation (if different)	

	<i>bulb b</i>
prediction	
explanation	
observation	
explanation (if different)	

Iterated integrals

Chapter 10

Iterated integrals (optional)

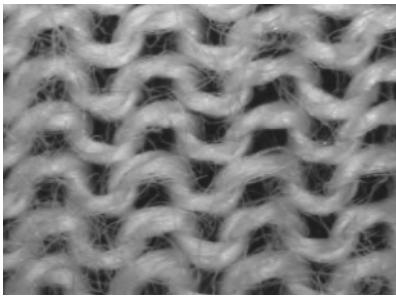
In sec. 2.2, p. 46, we introduced the concept of flux, which we defined using a surface integral. We have avoided up until now any development of the toolbox of techniques that are needed in order to evaluate nontrivial surface integrals, restricting ourselves to cases like example 2, p. 48, in which the integral could be broken down into a few pieces, and the integrand was a constant for each piece. In this optional chapter, we describe the main technique for explicitly evaluating a surface integral, which is to rewrite it as two plain old integrals, one nested inside the other like Russian dolls. These are called iterated integrals, and they have other applications besides the calculation of fluxes. We will also look at some of these other applications as they apply to electricity and magnetism. Knowledge of this material is not assumed later in the book.

10.1 A warm-up: iterated sums

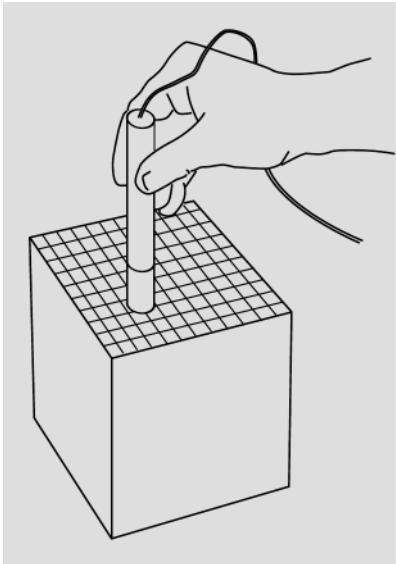
A kindergarten teacher in Chicago is letting her kids out onto the snowy playground for recess. One little boy is wearing a nice red wool cardigan, but he doesn't know how to button it up. Mrs. Kidlove uses the following algorithm to button it up for him:

- 1 For each button,
- 2 push the button through the hole.

This method of describing an algorithm is called pseudocode, because unlike actual programming code, it's written in a format meant to be read by humans. Line 1 introduces a repetitive action, known in programming as a “loop,” because after performing the action in line 2, we come back and do it again. Line 2 is the body of the loop — the thing that we're going to do over and over — and it's indented to show this. The effect is like modern poetry, but not as pretentious.



a / A knitted fabric.



b / Measuring the flux through the box in minilab 2, p. 71.



c / A puzzle involving a discrete sum over a surface.

But wait, now Mrs. Kidlove realizes it's not just this one kid. Red wool cardigans are in style this year, and every little kid needs to be buttoned up before going out. Now her algorithm looks like this:

- 1 For each kid,
- 2 for each button,
- 3 push the button through the hole.

To see how all this relates to surface integrals, let's consider the steps that would have been required in order to knit one of those sweaters by hand. For those unfamiliar with knitting, figure a shows the idea:

- 1 For each row,
- 2 for each stitch,
- 3 do certain things with the knitting needles.

We're now going systematically through all the locations in a surface, as in minilab 2 (fig. b), where we measured the flux through the top of the box:

- 1 For each row,
- 2 for each column,
- 3 add the flux through the square to the total flux.

This is still a discrete sum rather than an integral, but it clearly can be used to approximate an integral. As an example involving a discrete sum that we can write in mathematical notation, consider the following puzzle (figure c) posed by Sam Loyd, who was known around 1900 as the “prince of puzzlers.”

According to encyclopediacal lore, the royal game, or what is now known as chess, was invented by ... Sessa, and the king of [the Gupta Empire], Shevan the Great, asked Sessa what reward he demanded for his wonderful game. Sessa astonished the king by the apparent moderation of his demand, viz., one grain of wheat for the first square of the chess-board, two for the second, four for the third, eight for the fourth, and so on, always doubling for each square up to the sixty-fourth square of the chess board.

This sum can be written as

$$\sum_{i=0}^7 \sum_{j=0}^7 2^{8i+j}.$$

Actually the two-dimensional nature of this sum is not really crucial to the statement or solution of the puzzle, but it's picturesque and illustrates the ideas and notation. The idea here is that we number the rows from 0 to 7, and also the columns. Going across row $i = 0$, the terms in the sum go $2^0 + \dots + 2^7$ as the column runs from $j = 0$ to $j = 7$. Then in row $i = 1$ we have $2^8 + \dots + 2^{15}$, and so on for i running up to 7. Loyd doesn't explicitly state what the reader is supposed to do in order to solve the puzzle, and figuring out the right trick isn't relevant to your study of electricity and magnetism, but if you enjoy this sort of thing:

self-check A

Simplify the sum and then approximate it in your head using scientific notation. (You will find it useful that $10^{10} = 1024 \approx 10^3$, "one k.") ▷ Answer, p. 432

Although this puzzle can be fun to solve, the point here is really to introduce some concepts and notation, using discrete sums as a warm-up for integrals. The sums \sum are like integrals \int , and the symbols i and j are like variables of integration such as dx and dy . In both cases, the variables are "bound." That is, in an expression like $\sum_i \dots$, the symbol i only has meaning *inside* the sum, and similarly in an integral $\int \dots dx$, x doesn't mean anything outside the integral.



d / Iterated sums involve nesting one sum inside another like Russian dolls, and similarly for iterated integrals.

10.2 Iterated integrals

The continuous version of the sum in the Loyd puzzle would be

$$\int_{y=0}^8 \left(\int_{x=0}^8 2^{8x+y} dx \right) dy.$$

We have an "integral sandwich." Usually no ambiguity results if we simplify the notation a little:

$$\int_0^7 \int_0^7 2^{8x+y} dx dy.$$

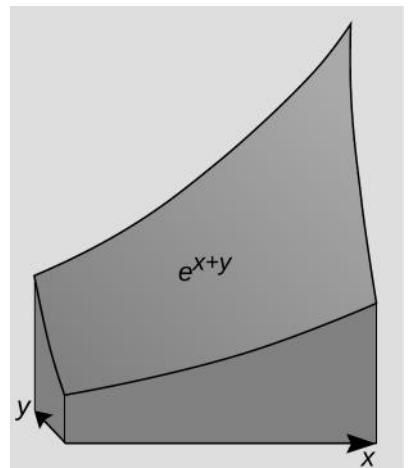
This turns out to be of about the same order of magnitude as the discrete sum.

To make this a little easier to visualize and work with, let's change it to the following:

$$\int_0^1 \left(\int_0^1 e^{x+y} dx \right) dy.$$

Just as an ordinary integral is the area under a curve, this iterated integral can be interpreted as the volume under the surface in figure e. We can handle this computation just by applying some algebra and the familiar methods of integration. First let's break up the exponential into two factors:

$$\int_0^1 \left(\int_0^1 e^x e^y dx \right) dy.$$



e / The function e^{x+y} .

The advantage of doing this is that as far as the inside integral is concerned, x is the variable of integration, and y is just a constant. Then we can do as we are always allowed in integration and take the constant factor outside:

$$\int_0^1 \left(e^y \int_0^1 e^x dx \right) dy.$$

Now the innermost integral is just an ordinary one-dimensional integral, which we can evaluate as $e^x|_0^1 = e - 1$. We then have

$$\begin{aligned} & \int_0^1 (e^y(e - 1)) dy \\ &= (e - 1) \int_0^1 e^y dy \\ &= (e - 1)^2 \\ &\approx 2.95. \end{aligned}$$

As a check to see that this is reasonable, the solid in the figure has a base that is 1×1 in area, so its volume should be equal to 1×1 multiplied by the average value of the function within this square. The function varies from $e^{0+0} = 1$ at the lower corner to $e^{1+1} \approx 7.4$ at the top of the ski jump, so it's pretty plausible that its average value is 2.95.

10.3 Varying limits of integration

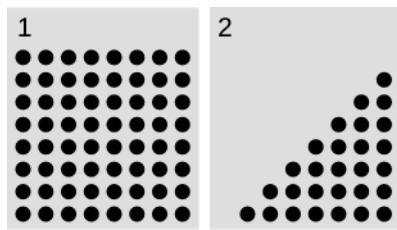
One way to express the number of squares on a chess board is

$$\sum_{i=0}^7 \sum_{j=0}^7 1 = 64,$$

represented visually by figure f/1. This is the world's silliest way to notate the type of standard grade-school arithmetic problem in which Sally has 8 skirts and 8 tops, and we want to know how many outfits she can put together.

Suppose, on the other hand, that Sally has 8 dogs, and she wants to pick two to take on a walk. Then the sum becomes

$$\sum_{i=0}^7 \sum_{j=0}^{i-1} 1 = 28,$$



f / Summing dots.

figure f/2. The restriction on the inner sum prevents double counting, since, e.g., taking Daisy and Lucy on a walk is the same as taking Lucy and Daisy. We've come across sums of this form previously, in example 1, p. 74, where we found the electrical energy of a molecule containing 6 atoms. With a 1 inside the sum, as in the present examples, we got the number of combinations of atoms

that needed to be considered, while with an energy inside the sum we got the total energy.

In the second example, note how the upper limit of the inner sum depends on the summation variable i used in the outer sum. This makes sense, because i is a bound variable that does mean something inside the outer sum. On the other hand, it would not make sense to write a sum like $\sum_{i=0}^{j-1} \sum_{j=0}^7$, because the upper limit $j - 1$ contains the symbol. In the example of the teacher buttoning the kids' sweaters, a similar nonsensical example would be

- 1 For each button on that kid's sweater,
- 2 for each kid,
- 3 push the button through the hole.

If you show this sentence to someone, they'll get confused because there is no "that kid" at the outermost level.

All of these ideas apply to integrals as well. If the limits of integration are constant and finite, then an integral of the form $\int_a^b \int_c^d \dots$ can only be an integral over a rectangle. But if we relax this restriction, then we can, for example, find the area of a triangle:

$$\begin{aligned} \int_0^1 \int_0^y dx dy &= \int_0^1 y dy \\ &= \frac{1}{2}. \end{aligned}$$

Often we want to do a change of variable (" u substitution), and this works out pretty much the same for iterated integrals as for ordinary integrals, including the need to change the differential and the limits of integration.

Area of a right triangle

example 1

Let a right triangle have legs a and b . We want to find its area. As a preliminary, we set up a coordinate system with side a along the x axis and the hypotenuse described as the line $y = (b/a)x$. The setup looks like this.

$$\int_0^a \int_0^{bx/a} dy dx$$

Notice how we had to make the y integral the inside part of the sandwich and x the outside part. If we hadn't done this, then we would have had to fiddle with things to keep from ungrammatically referring to the inner integral's variable in the outer integral's limits of integration. In examples like these, it's almost always advantageous to do a change of variables, if possible, such that the new variables are unitless. Here, the natural way to accomplish that is to define $u = x/a$ and $v = y/b$. We then have

$dy dx = ab dv du$ and for the upper limit of integration on the inside integral, $v = y/b = (bx/a)/b = x/a = u$. Taking the constant factor ab outside, we then have

$$ab \int_0^1 \int_0^u dv du.$$

This is a big win, because the double integral is now a true definite integral, with no adjustable parameters like a or b . In fact, we've already evaluated this integral (with different names for the bound variables, which is irrelevant) and found it to be $1/2$, so the result is $ab/2$, as expected.

10.4 How to set up applications

Up until now we've been doing a lot of integrals where the thing inside the integral is just 1. But in real-world applications we usually have something we're integrating. Here are some typical applications from mechanics, in which we integrate over space:

<i>thing being integrated</i>	<i>integral</i>
density	mass
density $\times r^2$	moment of inertia

In electricity and magnetism:

<i>thing being integrated</i>	<i>integral</i>
charge density	charge
energy density	energy
momentum density	momentum
charge density $\times z$	dipole moment along the z axis

Any of these can exist in one dimension (\int), two ($\int \int$), or three ($\int \int \int$). The following two examples from mechanics illustrate the typical setup process in a one-dimensional case.

Mass of an air column

example 2

- ▷ Above some small area on the surface of the earth, we have an air column with mass per unit height (i.e., linear density) $dm/dz = Ae^{-az}$, where a is a positive constant A tells us how thick the atmosphere is at sea level. Find the total mass in this air column.
- ▷ The general idea is simply to start by saying that the integral is the sum of a whole bunch of little tiny parts:

$$\int dm.$$

If you're not experienced with this sort of thing, it might never occur to you to write this down, because m isn't a variable, and nothing is a function of m . That's OK. Here dm just means "a little bit of mass," and \int means "add up all the..." But in order

to evaluate the integral, we now have to substitute:

$$\begin{aligned}\int dm &= \int \frac{dm}{dz} dz \\ &= \int_0^\infty Ae^{-az} dz \\ &= \frac{A}{a}.\end{aligned}$$

This makes sense, because the units come out right ($(\text{kg}/\text{m})/\text{m}^{-1}$), and the dependence on A and a is right (e.g., smaller a means the density falls off more slowly with height, which should give a bigger total mass).

Moment of inertia of a rod

example 3

- ▷ Find the moment of inertia of a uniform rod for rotation about one end.
- ▷ Let the rod's total mass be M and its length ℓ . The moment of inertia I is

$$\begin{aligned}I &= \int dl \\ &= \int r^2 dm.\end{aligned}$$

Because the rod is uniform, $dm/dr = M/\ell$, and $dm = (M/\ell) dr$.

$$\begin{aligned}I &= \int_0^\ell r^2 \frac{M}{\ell} dr \\ &= \frac{M}{\ell} \int_0^\ell r^2 dr\end{aligned}$$

It's nicer to change to the unitless variable $u = r/\ell$, which gives

$$\begin{aligned}I &= \frac{M}{\ell} \ell^3 \int_0^1 u^2 du \\ &= \frac{1}{3} M \ell^2.\end{aligned}$$

Here's an example involving electromagnetism and three dimensions.

Energy of an electromagnetic wave

example 4

- ▷ Find the average energy density of a sinusoidal electromagnetic plane wave, within a rectangular box oriented so that its axes coincide with **E**, **B**, and **S**. Let the box encompass an integer number of wavelengths.
- ▷ Using the expressions for the energy densities of the fields, and

the fact that $Bc = E$, we have

$$\begin{aligned} U &= \int dU \\ &= \int \frac{1}{4\pi k} E^2 dv \\ &= \frac{1}{4\pi k} \int E^2 dv. \end{aligned}$$

Let the z axis be the direction of propagation, and let x and y be aligned with the other edges of the box. Let $k = 2\pi/\lambda$. Then

$$\begin{aligned} U &= \frac{1}{4\pi k} \int E^2 dx dy dz \\ &= \frac{1}{4\pi k} \int \int \int \tilde{E}^2 \sin^2 kz dx dy dz \\ &= \frac{\tilde{E}^2}{4\pi k} \int \int \int \sin^2 kz dx dy dz. \end{aligned}$$

The integrand doesn't depend on x or y , so

$$\begin{aligned} U &= \frac{\tilde{E}^2}{4\pi k} \int \sin^2 kz \left(\int \int dx dy \right) dz \\ &= \frac{\tilde{E}^2 A}{4\pi k} \int \sin^2 kz dz, \end{aligned}$$

where A is the cross-sectional area of the box in the plane perpendicular to propagation. Because the remaining z integral covers an integer number of wavelengths, its average value is $1/2$, and we can replace it with $1/2$. This fact becomes obvious if you graph the function \sin^2 , or if you consider that $\sin^2 + \cos^2 = 1$. The z integral therefore equals $1/2$ times the length of the box. Grouping these factors together with A , we get $V/2$, half the volume. The result for the total energy is $U = \tilde{E}^2 V / 8\pi k$, which gives the average energy density

$$\frac{U}{V} = \frac{\tilde{E}^2}{8\pi k}.$$

In other words, the electric and magnetic fields have equal energy densities, their average energy densities are equal to half their maximum values, and therefore the average total energy density is equal to the maximum energy density of the electric field.

Energy of a line of charge

example 5

▷ Find the electrical potential energy of a uniform line of charge of finite length.

▷ The total electrical potential energy is

$$U = \int dU = \frac{1}{2} \int \int \frac{k dq_1 dq_2}{r},$$

where the factor of $1/2$ is to undo the effect of double-counting. Let the density of charge be $\lambda = dq/dx$, and let the line go from $x = 0$ to ℓ . Then

$$\begin{aligned} U &= \frac{1}{2}k\lambda^2 \int_{x_1=0}^{\ell} \int_{x_2=0}^{\ell} \frac{dx_1 dx_2}{|x_1 - x_2|} \\ &= k\lambda^2 \int_{x_1=0}^{\ell} \int_{x_2=0}^{x_1} \frac{dx_1 dx_2}{x_1 - x_2}. \end{aligned}$$

In the second line, we've changed the limits of integration so that we only integrate over combinations of points with $x_2 < x_1$. This means we get rid of the $1/2$, and we don't need the absolute value in the denominator anymore. Evaluating the inside integral gives

$$U = -k\lambda^2 \int_{x_1=0}^{\ell} \ln(x_1 - x_2) \Big|_0^{x_1} dx_2.$$

Plugging in the limits of integration gives a divergence at $x_2 = x_1$, so the result is that the energy is $+\infty$. In other words, if we wanted to bring charges from far away and assemble them into this configuration, it would require an infinite amount of mechanical work.

10.5 Electric field of a continuous charge distribution

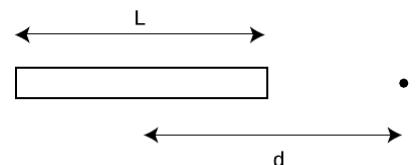
The electric field made by a continuous charge distribution is the sum of the fields created by every part of it. If we let the “parts” become infinitesimally small, we have a sum of an infinitely many infinitesimal numbers: an integral.

Field of a uniformly charged rod

example 6

- ▷ A rod of length L has charge Q spread uniformly along it. Find the electric field at a point a distance d from the center of the rod, along the rod's axis.

- ▷ This is a one-dimensional situation, so we really only need to do a single integral representing the total field along the axis. We imagine breaking the rod down into short pieces of length dz , each with charge dq . Since charge is uniformly spread along the rod, we have $dq = \lambda dz$, where $\lambda = Q/L$ (Greek lambda) is the charge per unit length, in units of coulombs per meter. Since the pieces are infinitesimally short, we can treat them as point charges and use the expression $k dq/r^2$ for their contributions to the field, where $r = d - z$ is the distance from the charge at z to



g / Example 6.

the point in which we are interested.

$$\begin{aligned} E_z &= \int \frac{k dq}{r^2} \\ &= \int_{-L/2}^{+L/2} \frac{k\lambda dz}{r^2} \\ &= k\lambda \int_{-L/2}^{+L/2} \frac{dz}{(d-z)^2} \end{aligned}$$

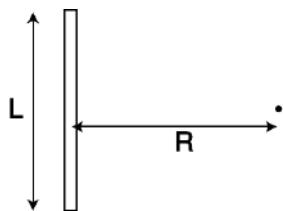
The integral can be looked up in a table, or reduced to an elementary form by substituting a new variable for $d - z$. The result is

$$\begin{aligned} E_z &= k\lambda \left(\frac{1}{d-z} \right) \Big|_{-L/2}^{+L/2} \\ &= \frac{kQ}{L} \left(\frac{1}{d-L/2} - \frac{1}{d+L/2} \right). \end{aligned}$$

For large values of d , this expression gets smaller for two reasons: (1) the denominators of the fractions become large, and (2) the two fractions become nearly the same, and tend to cancel out. This makes sense, since the field should get weaker as we get farther away from the charge. In fact, the field at large distances must approach kQ/d^2 .

It's also interesting to note that the field becomes infinite at the ends of the rod, but is not infinite on the interior of the rod. Can you explain physically why this happens?

Example 6 was one-dimensional. In the general three-dimensional case, we might have to integrate all three components of the field. However, there is a trick that lets us avoid this much complication. The potential is a scalar, so we can find the potential by doing just a single integral, then use the potential to find the field.



h / Example 7.

Potential, then field

example 7

▷ A rod of length L is uniformly charged with charge Q . Find the field at a point lying in the midplane of the rod at a distance R .

▷ By symmetry, the field has only a radial component, E_R , pointing directly away from the rod (or toward it for $Q < 0$). The brute-force approach, then, would be to evaluate the integral $E = \int |d\mathbf{E}| \cos \theta$, where $d\mathbf{E}$ is the contribution to the field from a charge dq at some point along the rod, and θ is the angle $d\mathbf{E}$ makes with the radial line.

It's easier, however, to find the potential first, and then find the field from the potential. Since the potential is a scalar, we simply integrate the contribution dV from each charge dq , without even worrying about angles and directions. Let z be the coordinate that measures distance up and down along the rod, with $z = 0$ at

the center of the rod. Then the distance between a point z on the rod and the point of interest is $r = \sqrt{z^2 + R^2}$, and we have

$$\begin{aligned} V &= \int \frac{k dq}{r} \\ &= k\lambda \int_{-L/2}^{+L/2} \frac{dz}{r} \\ &= k\lambda \int_{-L/2}^{+L/2} \frac{dz}{\sqrt{z^2 + R^2}} \end{aligned}$$

The integral can be looked up in a table, or evaluated using computer software:

$$\begin{aligned} V &= k\lambda \ln \left(z + \sqrt{z^2 + R^2} \right) \Big|_{-L/2}^{+L/2} \\ &= k\lambda \ln \left(\frac{L/2 + \sqrt{L^2/4 + R^2}}{-L/2 + \sqrt{L^2/4 + R^2}} \right) \end{aligned}$$

The expression inside the parentheses can be simplified a little. Leaving out some tedious algebra, the result is

$$V = 2k\lambda \ln \left(\frac{L}{2R} + \sqrt{1 + \frac{L^2}{4R^2}} \right)$$

This can readily be differentiated to find the field:

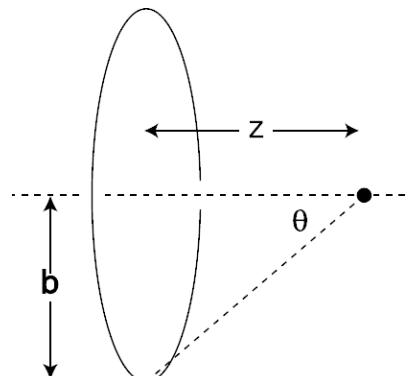
$$\begin{aligned} E_R &= -\frac{dV}{dR} \\ &= (-2k\lambda) \frac{-L/2R^2 + (1/2)(1 + L^2/4R^2)^{-1/2}(-L^2/2R^3)}{L/2R + (1 + L^2/4R^2)^{1/2}}, \end{aligned}$$

or, after some simplification,

$$E_R = \frac{k\lambda L}{R^2 \sqrt{1 + L^2/4R^2}}$$

For large values of R , the square root approaches one, and we have simply $E_R \approx k\lambda L/R^2 = kQ/R^2$. In other words, the field very far away is the same regardless of whether the charge is a point charge or some other shape like a rod. This is intuitively appealing, and doing this kind of check also helps to reassure one that the final result is correct.

The preceding example, although it involved some messy algebra, required only straightforward calculus, and no vector operations at all, because we only had to integrate a scalar function to find the potential. The next example is one in which we can integrate either the field or the potential without too much complication.



i / Example 8.

On-axis field of a ring of charge

example 8

- ▷ Find the potential and field along the axis of a uniformly charged ring.

▷ Integrating the potential is straightforward.

$$\begin{aligned}V &= \int \frac{k dq}{r} \\&= k \int \frac{dq}{\sqrt{b^2 + z^2}} \\&= \frac{k}{\sqrt{b^2 + z^2}} \int dq \\&= \frac{kQ}{\sqrt{b^2 + z^2}},\end{aligned}$$

where Q is the total charge of the ring. This result could have been derived without calculus, since the distance r is the same for every point around the ring, i.e., the integrand is a constant. It would also be straightforward to find the field by differentiating this expression with respect to z .

Instead, let's see how to find the field by direct integration. By symmetry, the field at the point of interest can have only a component along the axis of symmetry, the z axis:

$$\begin{aligned}E_x &= 0 \\E_y &= 0\end{aligned}$$

To find the field in the z direction, we integrate the z components contributed to the field by each infinitesimal part of the ring.

$$\begin{aligned}E_z &= \int dE_z \\&= \int |\mathbf{dE}| \cos \theta,\end{aligned}$$

where θ is the angle shown in the figure.

$$\begin{aligned}E_z &= \int \frac{k dq}{r^2} \cos \theta \\&= k \int \frac{dq}{b^2 + z^2} \cos \theta\end{aligned}$$

Everything inside the integral is a constant, so we have

$$\begin{aligned}E_z &= \frac{k}{b^2 + z^2} \cos \theta \int dq \\&= \frac{kQ}{b^2 + z^2} \cos \theta \\&= \frac{kQ}{b^2 + z^2} \frac{z}{r} \\&= \frac{kQz}{(b^2 + z^2)^{3/2}}\end{aligned}$$

In all the examples presented so far, the charge has been confined to a one-dimensional line or curve. Although it is possible, for example, to put charge on a piece of wire, it is more common to encounter practical devices in which the charge is distributed over a two-dimensional surface, as in the flat metal plates used in Thomson's experiments. Mathematically, we can approach this type of calculation with the divide-and-conquer technique: slice the surface into lines or curves whose fields we know how to calculate, and then add up the contributions to the field from all these slices. In the limit where the slices are imagined to be infinitesimally thin, we have an integral.

Field of a uniformly charged disk

example 9

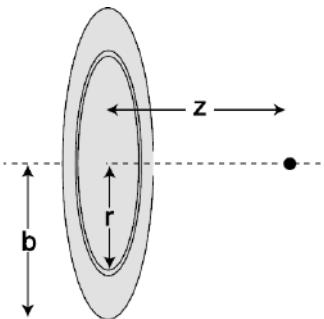
- ▷ A circular disk is uniformly charged. (The disk must be an insulator; if it was a conductor, then the repulsion of all the charge would cause it to collect more densely near the edge.) Find the field at a point on the axis, at a distance z from the plane of the disk.
- ▷ We're given that every part of the disk has the same charge per unit area, so rather than working with Q , the total charge, it will be easier to use the charge per unit area, conventionally notated σ (Greek sigma), $\sigma = Q/\pi b^2$.

Since we already know the field due to a ring of charge, we can solve the problem by slicing the disk into rings, with each ring extending from r to $r + dr$. The area of such a ring equals its circumference multiplied by its width, i.e., $2\pi r dr$, so its charge is $dq = 2\pi\sigma r dr$, and from the result of example 8, its contribution to the field is

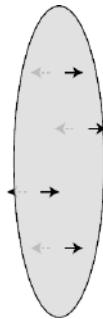
$$\begin{aligned} dE_z &= \frac{kz dq}{(r^2 + z^2)^{3/2}} \\ &= \frac{2\pi\sigma k z r dr}{(r^2 + z^2)^{3/2}} \end{aligned}$$

The total field is

$$\begin{aligned} E_z &= \int dE_z \\ &= 2\pi\sigma k z \int_0^b \frac{r dr}{(r^2 + z^2)^{3/2}} \\ &= 2\pi\sigma k z \left[\frac{-1}{\sqrt{r^2 + z^2}} \right]_{r=0}^{r=b} \\ &= 2\pi\sigma k \left(1 - \frac{z}{\sqrt{b^2 + z^2}} \right) \end{aligned}$$



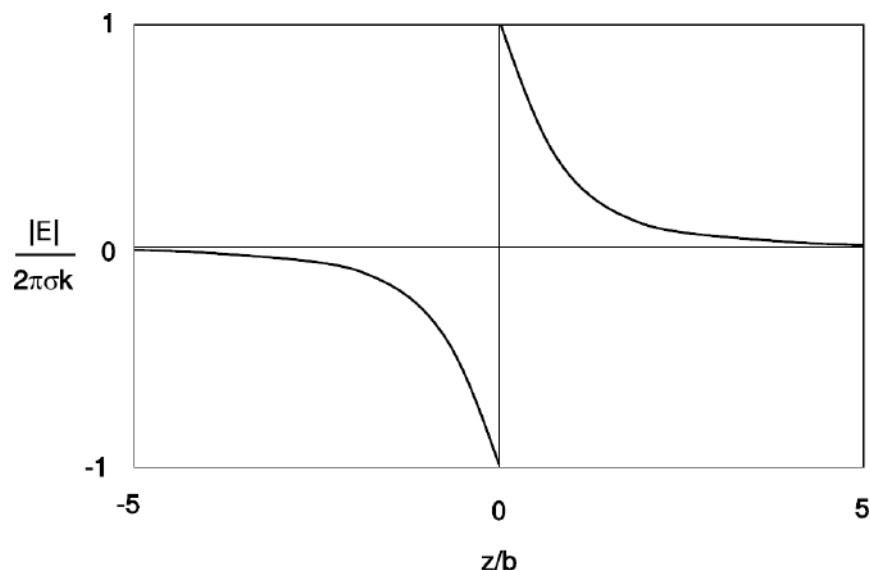
j / Example 9: geometry.



k / Example 9: the field on both sides (for $\sigma > 0$).

The result of example 9 has some interesting properties. First, we note that it was derived on the unspoken assumption of $z > 0$. By symmetry, the field on the other side of the disk must be equally strong, but in the opposite direction, as shown in figures k and l. Thus there is a discontinuity in the field at $z = 0$. In reality, the disk will have some finite thickness, and the switching over of the field will be rapid, but not discontinuous.

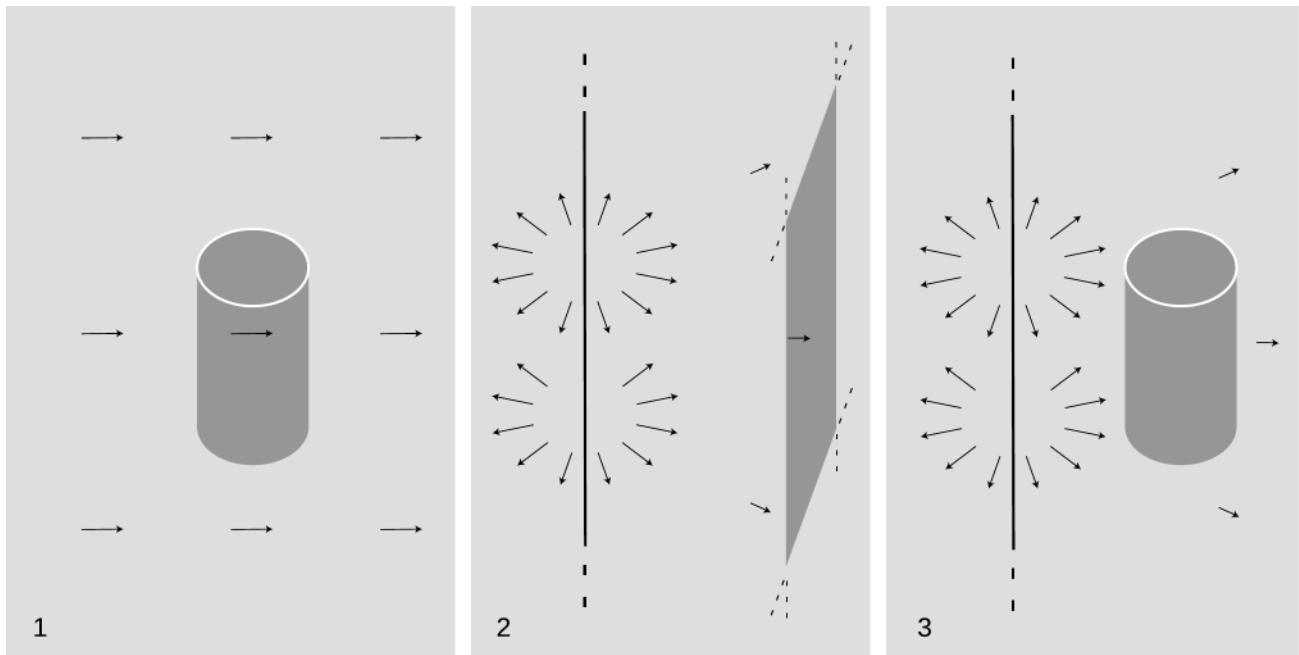
At large values of z , i.e., $z \gg b$, the field rapidly approaches the $1/r^2$ variation that we expect when we are so far from the disk that the disk's size and shape cannot matter.



I / Example 9: variation of the field ($\sigma > 0$).

10.6 Surface integrals

The integral form of Maxwell's equations involves the concept of flux, so that surface integrals involving flux are of fundamental importance. Up until now, we've avoided many of the mathematical complications involved in actually carrying out these integrals. The typical procedure was the one used in example 8, p. 58, in which we found the electric field of a line of charge. We used a tricky choice of the Gaussian surface that conformed to the cylindrical symmetry of the problem. This brought the integral into the form $\int(\text{constant}) dA$, and we could then bring the constant outside the integral.



m / Three examples of flux integrals, with different levels of difficulty and complication.

Figure m shows some examples of what can happen when this trick doesn't apply. One relatively simple case is the one in which we have a uniform field and a curved Gaussian surface, m/1. We can also have a nonuniform field and a flat surface, m/2. And finally we have the most general case, m/3, in which the field is nonuniform and the surface curved. In this section we will carry out the computations in these three examples in order to demonstrate the relevant techniques. All three are examples in which we already know the answer from Gauss's law, so we can easily check that our results are correct.

10.6.1 Curved surface, uniform field

Consider the example of m/1, in which a cylindrical Gaussian surface is immersed in a uniform electric field, with the field perpendicular to the cylinder's axis. Let the field be in the x direction, and let the axis of the cylindrical Gaussian surface be in the z direction. Let the radius of the cylinder be a . We define coordinates (θ, z) on the surface, so that any point on the surface can be described by these two numbers. The angle θ is defined counterclockwise from the x axis when we project the given point into the x - y plane at $z = 0$. For infinitesimal changes dz and $d\theta$ in the coordinates, the area is $dA = (\text{arc length})dz = a d\theta dz$. Let the height of the cylinder be H .

The area vector $d\mathbf{A}$ has magnitude dA and points outward, perpendicular to the cylinder. The angle between $d\mathbf{A}$ and \mathbf{E} is θ , so we have $\mathbf{E} \cdot d\mathbf{A} = E \cos \theta dA$. The flux integral is

$$\begin{aligned}\Phi &= \int E \cos \theta dA \\ &= E \int \cos \theta dA \\ &= E \int_0^H \int_0^{2\pi} \cos \theta d\theta dz \\ &= EH \int_0^{2\pi} \cos \theta d\theta.\end{aligned}$$

But here we are integrating a sine wave over one full cycle, so it averages to zero by symmetry, and the result is that $\Phi = 0$.

To confirm this by Gauss's law, we need to know how much charge is inside the cylinder. The divergence of a uniform field is zero, so there is zero charge density everywhere. Thus there is zero charge enclosed, and we confirm by Gauss's law that we should have $\Phi = 0$.

10.6.2 Flat surface, varying field

In figure m/2, we consider the flux through an infinite plane due to an infinite, uniform line of charge lying parallel to the plane. Let the line of charge have density λ , and let its distance from the plane be b . We choose coordinates such that the line of charge is in the z direction, and the x axis is in the direction from the line of charge to the closest part of the plane. We coordinatize our Gaussian surface as (y, z) , which gives the area element $dA = dy dz$. Let $r = \sqrt{b^2 + y^2}$ be the distance from a point in the plane to the nearest point on the line. At a particular point (y, z) , the cosine of the angle between the the field and the area vector is $\cos \theta = b/r$. The electric field is $E = 2k\lambda/r$ (example 8, p. 58). The flux integral

is therefore

$$\begin{aligned} d\Phi &= \mathbf{E} \cdot d\mathbf{A} \\ &= (E)(dA)(\cos \theta) \\ &= \left(\frac{2k\lambda}{r}\right)(dy dz) \left(\frac{b}{r}\right) \\ &= 2k\lambda b \frac{dy dz}{r^2}. \end{aligned}$$

Because both the line of charge and the surface are infinite, integrating this will give infinity. This is a common issue, with the common solution that we switch from calculating a given quantity such as flux to calculating the amount of that quantity per unit length. In the present example, the simple way to do this is simply to refrain from integrating in the z direction. Then the flux per unit length is

$$\begin{aligned} \frac{d\Phi}{dz} &= 2k\lambda b \int_{-\infty}^{\infty} \frac{dy}{r^2} \\ &= 2k\lambda b \int_{-\infty}^{\infty} \frac{dy}{b^2 + y^2}. \end{aligned}$$

This is actually an indefinite integral because it involves the parameter b , but we can turn it into a true definite integral by doing, as is always a good idea when possible, a change of variable to a unitless variable. Letting $u = y/b$, we then have

$$\frac{d\Phi}{dz} = 2k\lambda \int_{-\infty}^{\infty} \frac{du}{1+u^2}.$$

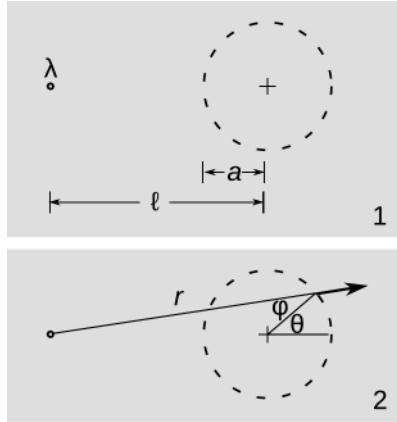
The definite integral can be done either using software or by figuring out that the correct trig substitution is $v = \tan^{-1} u$. The result is π , so we have

$$\frac{d\Phi}{dz} = 2\pi k\lambda.$$

We now compare with the result to be expected from Gauss's law. Gauss's law refers to a closed surface, which we don't have here. However, if we put a similar surface on the other side, we can make a geometry sort of like a pencil sandwich. Each piece of bread receives half the flux from the pencil. It doesn't matter that there is a gap between the two pieces of bread, because the bread slices are infinite, so zero flux gets out without passing through the bread. The result is that we expect the flux per unit length, through one surface, to be half the value given by Gauss's law for a closed surface:

$$\frac{d\Phi}{dz} = \frac{4\pi k dq/dz}{2} = 2\pi k\lambda.$$

This agrees with the result found by direct computation.



n / The flux through a cylindrical circuit due to a line of charge.

10.6.3 The general case

In the most general case, the surface is not flat, and the field is nonuniform, as in figure m/3, shown in cross-section in n/1. As our Gaussian surface, we take a cylinder of radius a and finite height H , whose axis is parallel to the line of charge and lies at a distance ℓ from it. We expect the flux to be zero, since the charge is outside the surface.

Figure n/2 shows the setup for directly calculating this flux. A certain point on the surface lies at a distance r from the line of charge, and the electric field at this point has magnitude $E = 2k\lambda/r$, and the direction shown by the arrow. To describe the point, we use the same (θ, z) coordinates as in section 10.6.1, and $dA = a d\theta dz$. We write φ for the angle between the area vector and the field. Then the flux is given by

$$\begin{aligned}\Phi &= \int \mathbf{E} \cdot d\mathbf{A} \\ &= \int E dA \cos \varphi \\ &= 2ka\lambda \int_0^H \int_0^{2\pi} \frac{\cos \varphi}{r} d\theta dz \\ &= 2kaH\lambda \int_0^{2\pi} \frac{\cos \varphi}{r} d\theta.\end{aligned}$$

Since we expect to prove Φ is zero, let's drop the constant factors in front. The law of cosines can be applied in two different ways here, as $\ell^2 = a^2 + r^2 - 2ar \cos \varphi$ and $r^2 = \ell^2 + a^2 + 2a\ell \cos \theta$. Using the first of these to eliminate $\cos \varphi$, we have for our integral

$$\Phi = \int_0^{2\pi} \left(1 - \frac{\ell^2 - a^2}{r^2} \right) d\theta.$$

Now using the second one to eliminate r^2 , we obtain

$$\Phi \propto \int_0^{2\pi} \left(1 - \frac{\ell^2 - a^2}{\ell^2 + a^2 + 2a\ell \cos \theta} \right) d\theta.$$

The writing can be simplified by defining the dimensionless constant $c = a/\ell$, with $c < 1$ because the cylinder doesn't contain the charge, and it also turns out to be helpful to let $\beta = 2c/(1+c^2)$, which gives $\beta < 1$. Then

$$\Phi \propto 2\pi - \frac{1-c^2}{1+c^2} \int_0^{2\pi} \frac{d\theta}{1+\beta \cos \theta}.$$

I didn't know how to do this integral, so I resorted to computer software, which verified that the result was indeed just right in order to cancel out the 2π term. Later I learned that integrals of this form can be approached systematically by using a set of rules called Bioche's rules. There is a Wikipedia article on the topic, including a sketch of how to apply the rules to this example.

Problems

Key

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

1 When we talk about rigid-body rotation, the concept of a perfectly rigid body can only be an idealization. In reality, any object will compress, expand, or deform to some extent when subjected to the strain of rotation. However, if we let it settle down for a while, perhaps it will reach a new equilibrium. As an example, suppose we fill a centrifuge tube with some compressible substance like shaving cream or Wonder Bread. We can model the contents of the tube as a one-dimensional line of mass, extending from $r = 0$ to $r = \ell$. Once the rotation starts, we expect that the contents will be most compressed near the “floor” of the tube at $r = \ell$; this is both because the inward force required for circular motion increases with r for a fixed ω , and because the part at the floor has the greatest amount of material pressing “down” (actually outward) on it. The linear density dm/dr , in units of kg/m, should therefore increase as a function of r . Suppose that we have $dm/dr = \mu e^{r/\ell}$, where μ is a constant. Find the moment of inertia. ✓

2 Show that when a thin, uniform ring rotates about a diameter, the moment of inertia is half as big as for rotation about the axis of symmetry. ➤ Solution, p. 429

3 An arc of a circle of radius b subtends an angle α and possesses a uniform linear charge density λ . Find the field at the center. ✓

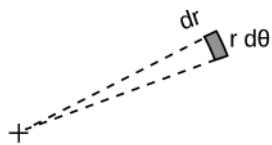
4 Find the moment of inertia of a solid rectangular box of mass M and uniform density, whose sides are of length a , b , and c , for rotation about an axis through its center parallel to the edges of length a . ✓

5 (a) As suggested in the figure, find the area of the infinitesimal region expressed in polar coordinates as lying between r and $r + dr$ and between θ and $\theta + d\theta$. ✓

(b) Generalize this to find the infinitesimal element of volume in cylindrical coordinates (r, θ, z) , where the Cartesian z axis is perpendicular to the directions measured by r and θ . ✓

(c) Find the moment of inertia for rotation about its axis of a cone whose mass is M , whose height is h , and whose base has a radius b . ✓

6 Astronomers believe that the mass distribution (mass per unit volume) of some galaxies may be approximated, in spherical coordinates, by $\rho = ae^{-br}$, for $0 \leq r \leq \infty$, where ρ is the density. Find the total mass. ✓



Problem 5.

7 Let two sides of a triangle be given by the vectors \mathbf{A} and \mathbf{B} , with their tails at the origin, and let mass m be uniformly distributed on the interior of the triangle. (a) Show that the distance of the triangle's center of mass from the intersection of sides \mathbf{A} and \mathbf{B} is given by $\frac{1}{3}|\mathbf{A} + \mathbf{B}|$.

(b) Consider the quadrilateral with mass $2m$, and vertices at the origin, \mathbf{A} , \mathbf{B} , and $\mathbf{A} + \mathbf{B}$. Show that its moment of inertia, for rotation about an axis perpendicular to it and passing through its center of mass, is $\frac{m}{6}(A^2 + B^2)$.

(c) Show that the moment of inertia for rotation about an axis perpendicular to the plane of the original triangle, and passing through its center of mass, is $\frac{m}{18}(A^2 + B^2 - \mathbf{A} \cdot \mathbf{B})$.

★



Problem 8.

8 Consider the electric field created by a uniformly charged cylindrical surface that extends to infinity in one direction.

(a) Show that the field at the center of the cylinder's mouth is $2\pi k\sigma$, which happens to be the same as the field of an infinite flat sheet of charge!

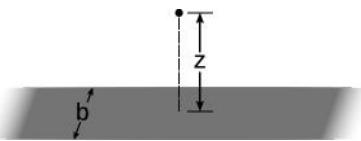
(b) This expression is independent of the radius of the cylinder. Explain why this should be so. For example, what would happen if you doubled the cylinder's radius?

9 (a) Show that the energy in the electric field of a point charge is infinite! Does the integral diverge at small distances, at large distances, or both? ▷ Hint, p. 425

(b) Now calculate the energy in the electric field of a uniformly charged sphere with radius b . Based on the shell theorem, it can be shown that the field for $r > b$ is the same as for a point charge, while the field for $r < b$ is kqr/b^3 .

Remark: The implications of this blow-up were sketched in section 5.5.3, p. 139.

✓



Problem 10.

10 (a) An infinite strip of width b has a surface charge density σ . Find the field at a point at a distance z from the strip, lying in the plane perpendicularly bisecting the strip. ✓

(b) Show that this expression has the correct behavior in the limit where z approaches zero, and also in the limit of $z \gg b$.

11 A solid cylinder of radius b and length ℓ is uniformly charged with a total charge Q . Find the electric field at a point at the center of one of the flat ends.

12 Find the potential at the edge of a uniformly charged disk. (Define $V = 0$ to be infinitely far from the disk.) ✓ ▷ Hint, p. 425

13 A rectangular box is uniformly charged with a charge density ρ . The box is extremely long and skinny, and its cross-section is a square with sides of length b . The length is so great in comparison to b that we can consider it as being infinite. Find the electric field

at a point lying on the box's surface, at the midpoint between the two edges. Your answer will involve an integral that is most easily done using computer software.

14 A hollow cylindrical pipe has length ℓ and radius b . Its ends are open, but on the curved surface it has a charge density σ . A charge q with mass m is released at the center of the pipe, in unstable equilibrium. Because the equilibrium is unstable, the particle accelerates off in one direction or the other, along the axis of the pipe, and comes shooting out like a bullet from the barrel of a gun. Find the velocity of the particle when it's infinitely far from the "gun." Your answer will involve an integral that is difficult to do by hand; you may want to look it up in a table of integrals, do it online at integrals.com, or download and install the free Maxima symbolic math software from maxima.sourceforge.net.

Sources of magnetism



This electric doorbell ringer uses two electromagnets made out of coils of copper wire with a layer of enamel or polymer insulation. The main practical goal of this chapter is to be able to calculate the magnetic field made by currents.

Chapter 11

Sources of magnetism

11.1 The current density

11.1.1 Definition

We've interpreted the junction rule (sec. 9.2, p. 210) as a statement of conservation of charge, but this interpretation only works if we assume that no charge ever gets stashed away temporarily at some location in the circuit and then brought back into action later. By analogy, the junction rule can be disobeyed by cars, if there are lots of cars coming into a parking garage in the morning and staying there all day until they leave after working hours. Circuits can actually accumulate charge, e.g., on the plates of a capacitor, so we would like to have a more general way of describing conservation of charge (which is always true) than the junction rule (which is only sometimes true).

By the way, this kind of thing arises in many physical situations, not just in electricity and magnetism. In figure a, the stream of water is fatter near the mouth of the faucet, and skinnier lower down. This is because the water speeds up as it falls. If the cross-sectional area of the stream was equal all along its length, then the rate of flow (kilograms per second) through a lower cross-section would be greater than the rate of flow through a cross-section higher up. Since the flow is *steady*, the amount of water between the two cross-sections stays constant. Conservation of mass therefore requires that the cross-sectional area of the stream shrink in inverse proportion to the increasing speed of the falling water. Notice how we had to assume a *steady* flow, so that no region of space had any net influx



a / The mass of water is conserved.

or outflow of water.

The first step in formulating this sort of thing mathematically for electric charge and currents is to recognize that we're describing a law of physics — conservation of charge — and the laws of physics are really always *local*. This means that the electric current, I , in units of amperes, is fundamentally ill suited to our purposes, since the definition of current involves the net charge flowing across some surface, which can be large. We want some way of talking about the flow of current at a particular *point* in space, which would be a kind of current *density* and should be a vector. The mathematical setup is the same as as the one occurring in the definition of flux (sec. 2.2, p. 47), but with different variables. The current through an infinitesimal area, with area vector $d\mathbf{A}$, is

$$dI = \mathbf{j} \cdot d\mathbf{A},$$

which implicitly defines the current density \mathbf{j} . Integrating this gives the current through a finite surface, $I = \int \mathbf{j} \cdot d\mathbf{A}$, which looks just like our definition of flux, but with different letters of the alphabet and a different physical interpretation. The current density has SI units of A/m^2 . It is a vector pointing in the net direction of flow of electric charge, with the flow of negative charge being represented by a vector in the opposite direction.

Current density in a copper wire

example 1

- ▷ Electrical codes in the U.S. require that a copper wire carrying 20 A should be at least of a certain size (called 12 gauge), which has a cross-sectional area of 3.31 mm^2 . What is the corresponding current density, assuming that the current is uniformly distributed across the wire's entire cross-section?
- ▷ Since the current density is stated to be constant, we can take it outside the integral, $I = \int \mathbf{j} \cdot d\mathbf{A} = \mathbf{j} \cdot \int \mathbf{A} = \mathbf{j} \cdot \mathbf{A}$. The cross-sectional area is stated for a cross-section perpendicular to the wire, so that the area vector points along the wire's axis, and therefore \mathbf{j} and \mathbf{A} are parallel, meaning that the dot product $\mathbf{j} \cdot \mathbf{A}$ is simply the product of the magnitudes jA . We then have

$$\begin{aligned} j &= \frac{I}{A} \\ &= 6.0 \times 10^6 \text{ A/m}^2. \end{aligned}$$

Comparing with the requirements in the code for other amounts of current, we find that the area is required to scale up somewhat faster than the current, so that the current density has to be somewhat smaller for thicker wires. This is presumably because a thicker wire has a smaller surface-to-volume ratio, and therefore cannot get rid of its heat as quickly.

Discussion questions

- A** Compare the currents in the bars. Compare the current densities.
- B** Each of the four figures shows a short section of a long current-carrying copper bar, whose cross-section is a square with sides of length b . In this side view, the height of the bar is b . The current density \mathbf{j} is constant and is in the direction shown by the white arrows. In example 1, the black line shows a surface cutting perpendicularly through the wire. We define the orientation of this surface to be to the right, as shown by the black arrow. Integrating both sides of $dI = \mathbf{j} \cdot d\mathbf{A}$, we obtain $I = jb^2$.
1. If we wanted to compare this calculation to reality by a measurement with a real ammeter, what would we actually have to do?
 2. Suppose we flip the orientation of the \mathbf{A} vector. What would this mean in terms of actual measurements?
 3. We now tilt the surface by 45 degrees. Recalculate I .
 4. Suppose the surface is now a sphere of diameter b . Find the error in the following calculation: $I = 4\pi(b/2)^2j = \pi b^2 j$.

11.1.2 Continuity equation

We can now imagine a “div-meter” (figure y, p. 61) that measures the divergence of the current density rather than the divergence of the electric field. If $\text{div } \mathbf{j}$ is nonzero — let’s say positive — at a certain point, then either conservation of charge is being violated at that point, or charge that was stored in that location is being taken out, like the cars in the parking garage leaving at the end of the day. Conservation of charge can now be stated succinctly as

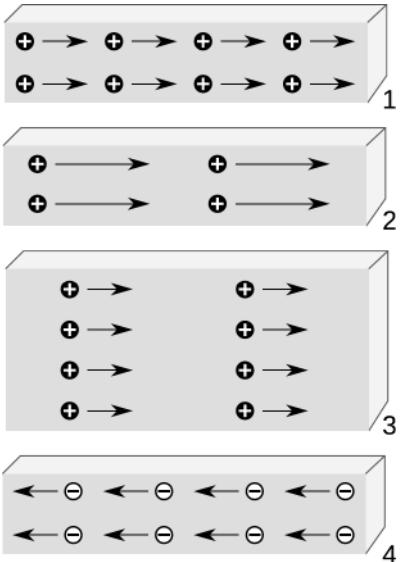
$$\text{div } \mathbf{j} = -\frac{\partial \rho}{\partial t},$$

where ρ is the charge density. An equation of this form also holds in other physical cases such as the flow of water, and is known more generally as an equation of continuity.

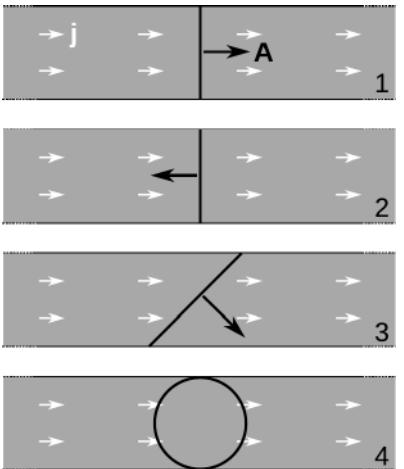
Skin depth

example 2

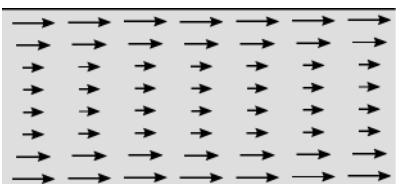
There is a general phenomenon in AC circuits that the flow of current in a wire is nonuniform: the current density is higher as we get closer to the surface of the wire. This is referred to as the “skin effect,” and it occurs because of induced electric fields. Figure b shows a typical profile of current density throughout a wire. If we visualize a div-meter in this field, we can see that although it would tend to spin and be swept downstream, it would not change its volume, which is how a div-meter registers a divergence. Therefore $\text{div } \mathbf{j} = 0$ in this example, and by conservation of charge we find that $\partial \rho / \partial t = 0$, i.e., charge is not being stored or taken out of storage anywhere in the wire. Skin depth is discussed further in example 9, p. 267.



Discussion question A.
The arrows are velocity vectors.
The charges are $\pm e$.

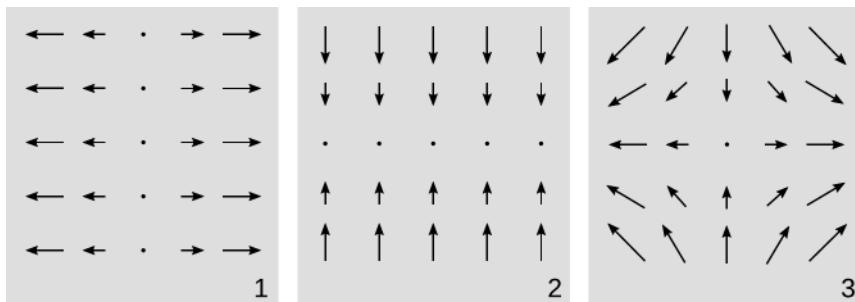


Discussion question B.



b / Example 2.

c / Example 3. Each sea-of-arrows diagram represents a current density \mathbf{j} . Which are physically possible?



Possible and impossible patterns of steady flow example 3

For a steady flow of electric charge, nothing is changing over time, so $\partial\rho/\partial t = 0$ and therefore the continuity equation requires $\operatorname{div} \mathbf{j} = 0$ everywhere: the flow must be divergence-free. Figure c shows some possible and impossible examples.

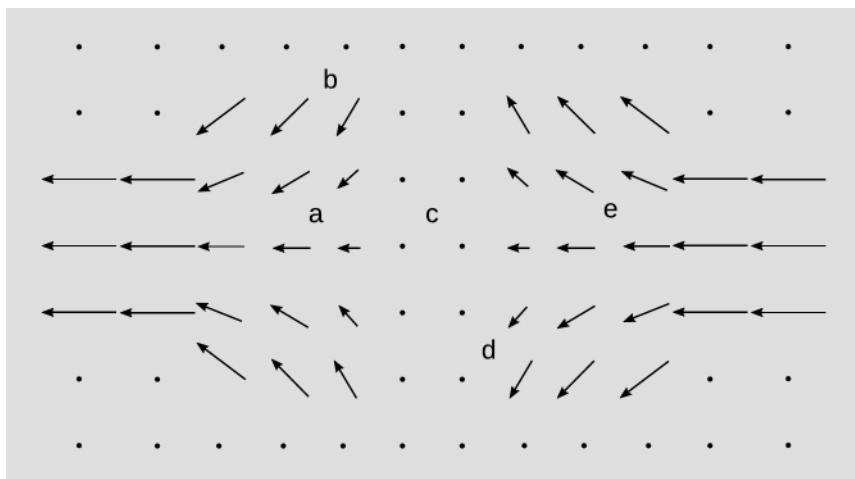
Example c/1 is impossible. The current density is of the form $\mathbf{j} = bx\hat{\mathbf{x}}$, with $b > 0$, and we showed in example 10, p. 62, that such a field has a divergence equal to b , which is nonzero.

The flow in c/2 is also impossible. This is like $\mathbf{j} = -by\hat{\mathbf{y}}$, which can be obtained from c/1 by rotating 90 degrees and then reversing all the vectors. Because the divergence is a scalar, the 90-degree rotation doesn't change the divergence. The divergence is also a *linear* operator (like any kind of derivative operator), so flipping the arrows works like $\operatorname{div}(-\mathbf{j}) = -\operatorname{div} \mathbf{j}$, i.e., it flips the sign of the result. We therefore find that the divergence of this flow is $-b$, and this is also impossible for a steady flow.

The flow in figure c/3 is the point-by-point vector sum of c/1 and c/2. Since the divergence is again a linear operator, we have $\operatorname{div}(\mathbf{j}_1 + \mathbf{j}_2) = \operatorname{div} \mathbf{j}_1 + \operatorname{div} \mathbf{j}_2$. But this is $b - b = 0$, so this *is* a possible steady flow of charges.

Discussion questions

d / Discussion question A.

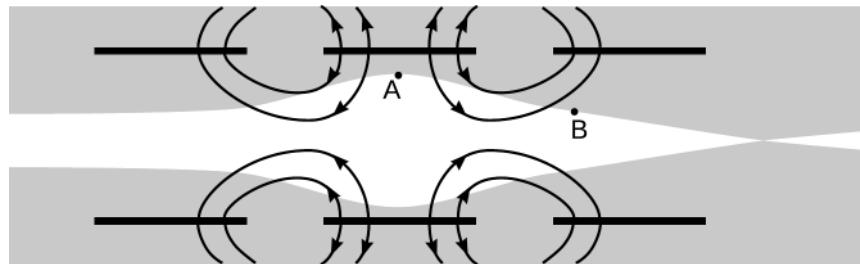


A The sea-of-arrows diagram in figure d represents a current density \mathbf{j} . The portions with the slanting arrows are identical to pieces of figure c/3. Imagine putting a div-meter at each of the marked points, and determine whether the divergence is positive, negative, or zero. Describe what is happening to the charge. If a positive charge enters from the right, where does it end up?

B (1) The figure shows a freeway on which we'll say that the cars are all initially uniformly spaced and moving at the same velocity. Suppose that every car begins accelerating at the same time and with the same acceleration. Describe how the continuity equation works out. (2) Now do the case where cars are flowing on a freeway at constant speed, but then begin to accelerate starting at some point in space. The flow is steady.



Discussion question B.



The einzel lens

example 4

Figure e shows a common electrostatic focusing device called an einzel lens. Einzel lenses are used to focus beams of charged particles in, for example, scanning electron microscopes and old-fashioned CRT video tubes. This one consists of three cylindrical pieces of metal. The axes coincide, and run from left to right in the figure, which shows a cross-sectional view. The two dark lines on the left are the top and bottom of the left-hand cylinder, and similarly in the middle and right. The left and right cylinders are at ground, the middle at $\phi > 0$.

As the electrons enter from the left side, they first encounter strong electric fields when they reach the gap between the left and middle cylinders. If we consider an electron at the top edge of the beam, the field initially accelerates it and moves it away from the axis, to point A. Continuing from A to B, the electron is slowed back down and deflected toward a point on the right at which the beam reaches a focus on the axis.

In the approximation that the electrons have constant velocity as they went from A to B and then to the focus, the beam forms a perfect cone, and its density varies as $1/r^2$, where r is the distance from the focus. This is a current density \mathbf{j} identical in form to the electric field of a point charge, and since we know that that field has zero divergence, so does this \mathbf{j} . The physics is different from that of the faucet in figure a, since the water is incompressible.

11.1.3 ★ Transformation properties (optional)

Clearly the values of ρ and \mathbf{j} depend on our frame of reference. By analogy with the flow of water, if Huck and Jim are floating down the Mississippi River on a raft, the current density is zero in their frame of reference. If this analogy were to hold in detail, then we would expect, however, that ρ would stay the same. This turns out not to be true. The purpose of this optional section is to work out the transformation properties of ρ and \mathbf{j} correctly. Some of the discussion will only be understandable to the reader who has studied optional chapter 7.

Suppose that in a certain frame of reference, a long, straight rod holds charge density ρ and carries a uniform current density j_z in the longitudinal direction. It will be surrounded by an electric field that is proportional to ρ , and a magnetic field proportional to j_z . Even if you haven't read ch. 7, you know that when we change frames of reference, we will have a new mix of electric and magnetic fields. Therefore if we know the transformation of the electric and magnetic fields, we can immediately infer the same transformation properties for ρ and j_z . We have worked out the transformation of the fields on p. 176, so we obtain the transformation of ρ and j_z simply by swapping in the new variables and putting in factors of c to make the units work:

$$c\rho' = \gamma c\rho - \frac{v}{c} \gamma j_z$$
$$j'_z = -\frac{v}{c} \gamma c\rho + \gamma j_z.$$

Because ρ and \mathbf{j} are things we can measure at a point in space, they have their own independent physical existence, and therefore these relationships hold regardless of whether the physical context is the one involving the straight rod.

Length contraction

example 5

Suppose that in a certain frame of reference, a wire carries zero current but has a charge. Setting $j_z = 0$ in the equation above for ρ' gives $\rho' = \gamma \rho$. This is the result of relativistic length contraction (sec. 7.2.2, p. 180).

11.2 Maxwell's equations

11.2.1 Adding a current term

In sec. 6.7 we studied Maxwell's equations in a vacuum. Now we wish to generalize them to their full form, including the possibility that there are both charges and currents. We've already seen in sec. 2.8 how to incorporate a charge density ρ , which acts as a source of electric fields. The only finishing touch left is to add a term describing how a current density \mathbf{j} produces a curly magnetic field. We first present Maxwell's equations (also summarized for convenience on p. 444) and then give some justification. They are:

$$\begin{aligned}\operatorname{div} \mathbf{E} &= 4\pi k\rho \\ \operatorname{div} \mathbf{B} &= 0 \\ \operatorname{curl} \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \operatorname{curl} \mathbf{B} &= \frac{1}{c^2} \frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi k}{c^2} \mathbf{j}.\end{aligned}$$

The only new feature is the final term involving \mathbf{j} . Something like this is required physically because we know that currents create magnetic fields. The k/c^2 has to be there because of units. The positive sign of this term expresses the right-hand relationship between the direction of the current and the curliness of the magnetic field it creates, which was proved in [2185](#). (If aliens on another planet define their magnetic field to be the opposite direction compared to ours, then they can flip all the signs of terms involving \mathbf{B} , but they have to do so in a consistent way.) The only thing left to justify is the factor of 4π , which we'll come back to after example 6.

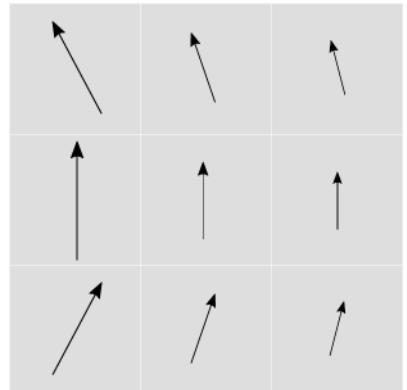
The field surrounding a wire

example 6

The figure shows the magnetic field in a region of space near a long, straight, current-carrying wire. A checkerboard grid is superimposed. The wire is three squares to the left of the figure, and its current is coming out of the page. We found the form of this field in example 2, p. 121: its magnitude falls off like $1/r$, and the field lines are circles.

Many people looking casually at this diagram would say that it has a nonzero curl that is out of the page (meaning, by the right-hand rule, counterclockwise as seen from the front). But this has to be wrong, because according to Maxwell's equations, $\operatorname{curl} \mathbf{B}$ is proportional to the current density \mathbf{j} that exists *here*, in this region of space. The wire isn't here, it's over to the left, so here $\mathbf{j} = 0$, and according to Maxwell, the curl of the field should be zero here.

A closer inspection shows that it is indeed plausible that the curl is zero. Suppose we put a curl-meter (p. 86) on top of the diagram. For simplicity, let's approximate the effect on the curl-meter by concentrating on four of the squares: top-middle, bottom-middle,



Example 6.

right-center, and left-center. The fields in the top-middle and bottom-middle squares create counterclockwise torques, i.e., a torque vector in the direction out of the page. However, the upward field in the center-left square is stronger than the one in the center-right square, and this imbalance creates a countervailing torque that is clockwise, i.e., into the page. It's not visually obvious that these torques exactly cancel, but it's plausible.

Returning to the factor of 4π , a nice way to verify this, which is fundamentally of more physical interest than the 4π itself, is to show that Maxwell's equations imply conservation of charge. The 4π then arises as the correct factor to put in so that charge conservation comes out correctly. This can be done in the following vector calculus calculation, which requires almost no real knowledge of vector calculus other than understanding generic ideas about the linearity of derivatives. You can skip ahead to sec. 11.2.2 if this kind of thing doesn't seem exciting.

We take the divergence of both sides of the fourth Maxwell's equation:

$$\operatorname{div}(\operatorname{curl} \mathbf{B}) = \operatorname{div} \left[\frac{1}{c^2} \frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi k}{c^2} \mathbf{j} \right].$$

The left side is zero by symmetry ([2277](#)). On the right-hand side, we can use the fact that the divergence is a kind of derivative, so it's linear, just like with a plain old derivative, which has $(c_1 f + c_2 g)' = c_1 f' + c_2 g'$. This gives

$$0 = \operatorname{div} \frac{\partial \mathbf{E}}{\partial t} + 4\pi k \operatorname{div} \mathbf{j}.$$

In general it's legitimate to swap the order of derivative operators ([2277](#)), so we can make this into

$$0 = \frac{\partial}{\partial t} \operatorname{div} \mathbf{E} + 4\pi k \operatorname{div} \mathbf{j}.$$

But now the first Maxwell's equation $\operatorname{div} \mathbf{E} = 4\pi k\rho$ allows us to make this into

$$0 = \frac{\partial \rho}{\partial t} + \operatorname{div} \mathbf{j},$$

which is the equation of continuity, stating that charge is conserved.

11.2.2 The view from the top of the mountain

In a similar way, it is also possible to show from Maxwell's equations that energy is conserved, a result known as Poynting's theorem. The basic concept is the same as in sec. 6.6.2, where we showed that the Poynting vector could be interpreted as a rate of energy flow, but extending it to apply more generally than in the case of a plane wave in vacuum. Way back at the beginning of ch. 6, p. 154, I claimed that we would eventually lay out the full set of tightly interlocking logical relationships behind Maxwell's equations. Here's

how that works. The logical development up until now has been the following:

- Assumptions:
1. Time is relative (sec. 1.1, p. 15).
 2. All frames of reference are equally valid, regardless of their motion how we orient them.
 3. Charge is conserved.
 4. Energy is conserved.
 5. Electric and magnetic fields have some basic properties (observability, vectors, and superposition — sec. 1.2, p. 18).
 6. Maxwell's equations
 7. $E = mc^2$ (sec. 5.5, p. 136)
 8. Time dilation, length contraction, and other facts about the structure of space and time (sec. 7.2, p. 178).
- Results:

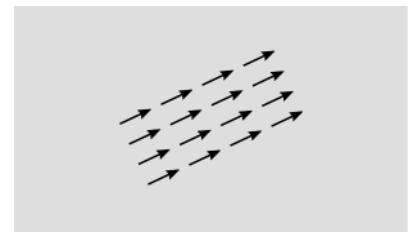
On the other hand, we can see that it's possible to run the logic in the opposite direction as well. If we consider Maxwell's equations to be a starting assumption (originally deduced by Maxwell from experimentalists' observations), then we can deduce facts such as conservation of charge. Historically, the idea of Einstein's two ground-breaking 1905 papers on relativity was to show that, starting from facts 2-6 as assumptions, he could prove facts 1, 7, and 8.

Out of nowhere?

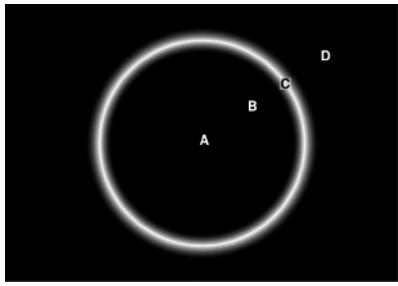
example 7

Adding the final current term to Maxwell's equations means that we know enough laws of physics that it should be possible, in principle, to predict the magnetic field made by a set of currents, as in the doorbell ringer on p. 257. A seemingly reasonable approach would be to break down the wire into short segments, like a dot-to-dot puzzle, find the field of one such segment, and then add up the fields of all the segments. This divide-and-conquer technique would then reduce the hard problem to the easier problem of finding the magnetic field created by a current distribution like the one in figure f.

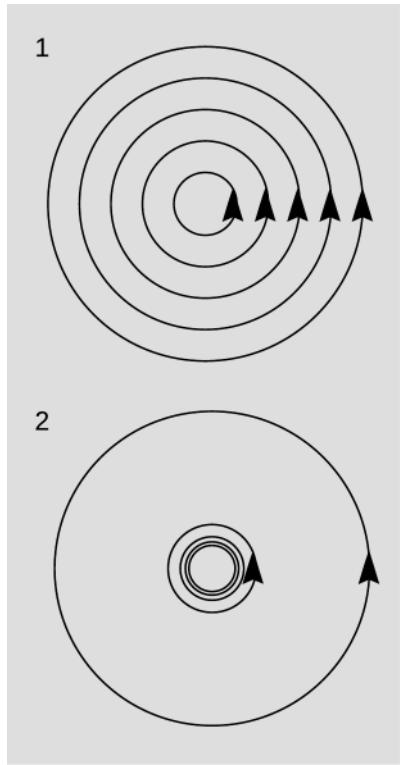
The trouble here is that the laws of physics can't predict the magnetic field in this situation, because the laws of physics forbid this situation from happening. Although the total amount of charge in the figure is staying constant, charge is springing into existence on the left and disappearing into nothingness on the right. If Maxwell's equations are true, then the continuity equation is true as well, and the continuity equation is a *local* law of physics: it forbids us from creating or destroying charge in one place, even if we try to make up for it somewhere else.



f / Example 7. A current distribution describing a segment of wire, isolated and hanging in space.



Discussion question A. Charge exists where there is white in the drawing, which is a snapshot of one moment in time.



Discussion question B.

Discussion questions

A The figure shows a cross-sectional view of a spherical shell of positive charge that is exploding outward, accelerating under the influence of its own repulsion. Describe \mathbf{E} , \mathbf{B} , and \mathbf{j} at points A, B, C, and D, and qualitatively verify the Maxwell's equation $\text{curl } \mathbf{B} = (1/c^2)\partial\mathbf{E}/\partial t + (4\pi k/c^2)\mathbf{j}$.

B The figure shows two magnetic field patterns. Pattern 1 is constant in magnitude, while 2's field falls off much faster than $1/r$. Compare their curls with the that of the external field of a current-carrying wire (example 6, p. 263.) Because of this, there must be a nonzero current density away from the axis. Infer the directions of these current densities.

11.3 Ohm's law in local form

If we could look inside a resistor with a DC current flowing through it, and zoom in to a very small scale, then the only things we would be able to probe locally through measurements — the only “observables” — would be the electric field \mathbf{E} and the current density \mathbf{j} . Whereas at a global scale we would say that the voltage drop ΔV causes a current I , locally it must be that \mathbf{E} causes \mathbf{j} . To get the strict proportionality in Ohm's law, with no nonlinearity for an idealized ohmic material, we must have $\mathbf{j} = \sigma\mathbf{E}$, for some constant of proportionality σ , called the conductivity. This is the local form of Ohm's law.

The conductivity depends on the material, and generally tells us how many free charge carriers are available, as well as the frequency of collisions that stop the charge carriers. For a perfect conductor, the conductivity is infinite, but electric fields are excluded, so the product $\sigma\mathbf{E}$ becomes the indeterminate form $\infty \cdot 0$. The units of conductivity can be expressed as $1/(\Omega \cdot \text{m})$. An example of a good conductor is copper, which has $\sigma \approx 6 \times 10^7 \Omega^{-1} \cdot \text{m}^{-1}$.

Conductivity of flesh

example 8

▷ The DC resistance of a human arm, measured from end to end, is about 300Ω . Estimate the conductivity of human flesh.

▷ If the arm has length L and cross-sectional area A , then $E = \Delta V/L$ and $j = I/A$. We then have $\sigma = j/E = L/AR$. Taking a human arm to have a diameter of 7 cm and a length of 60 cm, σ comes out to be about $0.5 \Omega^{-1} \cdot \text{m}^{-1}$, or eight orders of magnitude less than the conductivity of copper. Cf. problem 6, p. 342.

Skin depth's dependence on frequency example 9

Figure g is a copy of the one from example 2, p. 259, in which we said that the arrows represented the current density, and used the equation of continuity to find out that $\partial\rho/\partial t = 0$, i.e., charge is not being stored or taken out of storage anywhere in the wire. We can now use Maxwell's equations to see why this phenomenon occurs only for AC circuits.

By $\mathbf{j} = \sigma\mathbf{E}$, we see that the figure can be taken as a drawing of either the current or the electric field. The only difference would be the constant scalar factor σ , which is irrelevant since we haven't defined a numerical scale for the length of the arrows. The figure shows a pretty strong skin depth effect, but we could imagine making it either stronger or weaker than this. If it were much stronger, then appreciable fields and currents would exist only near the surface, like a skin. If it were much weaker, or nonexistent, then we would see a nearly uniform condition throughout the cross-section of the wire. We will argue that the effect must depend on the *frequency* of the current. In some common examples from everyday life, this frequency is zero for a DC circuit such as a flashlight, 60 Hz for most household appliances in the US, and ~ 1 GHz for a cell phone. In the case of 60 Hz, the current is switching directions back and forth 60 times per second, with a sinusoidal variation.

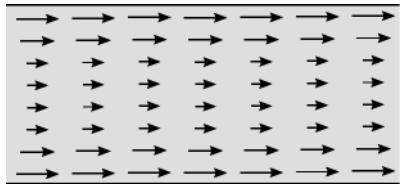
We apply the following two Maxwell's equations:

$$\begin{aligned}\text{curl } \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \text{curl } \mathbf{B} &= \frac{1}{c^2} \frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi k}{c^2} \mathbf{j}.\end{aligned}$$

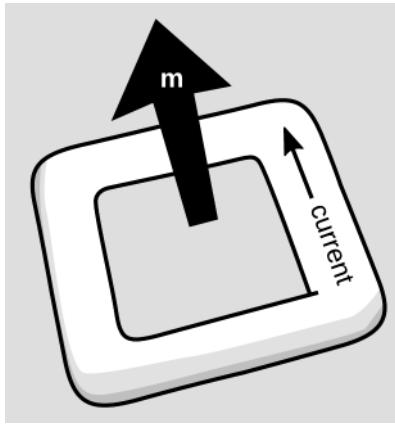
In the DC limit, all the time derivatives must vanish, and therefore $\text{curl } \mathbf{E} = 0$. If we imagine inserting a curl-meter for the electric field into figure g, we can see that a skin depth effect requires a nonzero $\text{curl } \mathbf{E}$. Therefore the skin effect cannot occur at DC, i.e., the skin depth δ goes to infinity as the frequency f approaches zero. A more detailed analysis shows that $\delta \propto f^{-1/2}$. Since the only unitful quantities available are σ , f , and the fundamental constants k and c , units then require $\delta \propto (\sigma fk)^{-1/2}c$.

11.4 The magnetic dipole

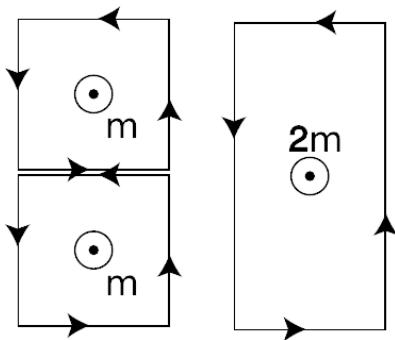
Example 7, p. 265, shows that if we want to find the field made by a current distribution like the one in the doorbell ringer, we can't necessarily chop up the current distribution into building blocks that look like little line segments. Let's instead investigate magnetic dipoles as building blocks.



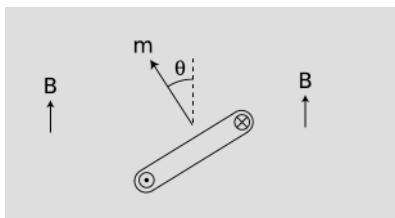
g / Example 9. The cross-section of a wire with an AC current flowing through it.



h / A magnetic dipole made out of a square current loop. The dipole vector is perpendicular to the loop and is related to the direction of the current by a right-hand rule.



i / Adding two dipole vectors. The dipole moment increases in proportion to the area.



j / Side view of a square current dipole, with the direction of the current as indicated, coming out of the page on one side and going back in on the other. The dipole is immersed in an externally imposed uniform field shown by the vertical arrows.

11.4.1 Modeling the dipole using a current loop

Our discussion of the dipole in sec. 5.4, p. 132, focused mainly on the electric dipole. Here we discuss the magnetic dipole in more detail. We've defined two types of dipoles in terms of the energy they have when they interact with an external field,

$$U = -\mathbf{D} \cdot \mathbf{E} \quad [\text{definition of the electric dipole moment } \mathbf{D}]$$

$$U = -\mathbf{m} \cdot \mathbf{B} \quad [\text{definition of the magnetic dipole moment } \mathbf{m}],$$

and the perfect mathematical analogy between the two definitions automatically implies that electric and magnetic dipoles have many of the same properties. Both dipole moments are measured by vectors, and this implies that they behave as vectors when we rotate them, and also that they add like vectors.

But this is somewhat abstract, and it's nice to have a more concrete physical picture in mind. Because our universe doesn't seem to come equipped with magnetic charges, we can't make a magnetic dipole by gluing such charges to the ends of a stick. Instead, the simplest embodiment of a magnetic dipole would be something like the current loop shown in figure h. Figure i shows an example of why it makes sense that dipole moments add as vectors, and that the dipole moment is proportional to the area of the loop. In exercise , p. 286, we will verify using the right-hand rule that the torque in figure j is in the direction that tends to align the dipole vector with the magnetic field. A calculation ([Z277](#)) shows that the dipole moment is

$$\mathbf{m} = IA,$$

where I is the current and \mathbf{A} is the area vector.

The magnetic dipole moment of an atom example 10

Let's make an order-of-magnitude estimate of the magnetic dipole moment of an atom. A hydrogen atom is about 10^{-10} m in diameter, and the electron moves at speeds of about $10^{-2} c$. We don't know the shape of the orbit, and indeed it turns out that according to the principles of quantum mechanics, the electron doesn't even have a well-defined orbit, but if we're brave, we can still estimate the dipole moment using the cross-sectional area of the atom, which will be on the order of $(10^{-10} \text{ m})^2 = 10^{-20} \text{ m}^2$. The electron is a single particle, not a steady current, but again we throw caution to the winds, and estimate the current it creates as $e/\Delta t$, where Δt , the time for one orbit, can be estimated by dividing the size of the atom by the electron's velocity. (This is only a rough estimate, and we don't know the shape of the orbit, so it would be silly, for instance, to bother with multiplying the diameter by π based on our intuitive visualization of the electron as moving around the circumference of a circle.) The result for the dipole moment is $m \sim 10^{-23} \text{ A} \cdot \text{m}^2$.

Should we be impressed with how small this dipole moment is, or with how big it is, considering that it's being made by a single atom? Very large or very small numbers are never very interesting by themselves. To get a feeling for what they mean, we need to compare them to something else. An interesting comparison here is to think in terms of the total number of atoms in a typical object, which might be on the order of 10^{26} (Avogadro's number). Suppose we had this many atoms, with their moments all aligned. The total dipole moment would be on the order of $10^3 \text{ A}\cdot\text{m}^2$, which is a pretty big number. To get a dipole moment this strong using human-scale devices, we'd have to send a thousand amps of current through a one-square meter loop of wire! The insight to be gained here is that, even in a permanent magnet, we must not have all the atoms perfectly aligned, because that would cause more spectacular magnetic effects than we really observe. Apparently, nearly all the atoms in such a magnet are oriented randomly, and do not contribute to the magnet's dipole moment.

11.4.2 Dipole moment related to angular momentum

In example 10 we made a crude estimate of the typical magnetic dipole moment of an atom. There is another way of going about this, which is potentially much more accurate and of interest as a way of probing the structure of the atom.

Suppose that a particle of charge q and mass m is whizzing around and around some closed path. We don't even care whether the trajectory is a square or a circle, an orbit or a random wiggle. But let's say for convenience that it's a planar shape. The magnetic dipole moment (averaged over time) is $\mathbf{m} = I\mathbf{A}$. But the angular momentum of a unit mass can also be interpreted as twice the area it sweeps out per unit time. Aside from the factor of two, which is just a historical glitch in the definitions, this mathematical analogy is exact: mass is to charge as angular momentum \mathbf{L} is to magnetic dipole moment \mathbf{m} . Therefore we have the identity

$$\frac{q}{m} \cdot \frac{|\mathbf{L}|}{|\mathbf{m}|} = 2$$

(where \mathbf{m} is the dipole moment, while m is the mass). The left-hand side is called the g factor. We expect $g = 2$ for a single orbiting particle.

Now suppose that we have a collection of particles with identical values of q/m . Then vector addition of the \mathbf{L} and \mathbf{m} values gives the same $g = 2$ for the system as a whole. On the other hand, if the different members of the system do *not* all have the same q/m , then the g of the system as a whole need not be 2. For example, a collection of positive and negative charges could easily have zero net charge but $\mathbf{m} \neq 0$, giving $g = 0$.

Particles such as the electron, the neutron, and the proton may

be pointlike, or they may be composites of other particles. The electron and proton, which are charged, have the expected g factors of exactly 2 when we measure the \mathbf{L} and \mathbf{m} that they have due to their motion through space. But we also find that electrons, neutrons, and protons all come equipped with a built-in angular momentum, present even when they are at rest. This intrinsic angular momentum, called spin, is fixed in magnitude but can vary in direction, like that of a gyroscope. Thus if we measure the \mathbf{L} and \mathbf{m} of these particles *at rest*, they have fixed g factors, figure k.

electron	2.002319304361
neutron	0
proton	5.58569471

k / g factors of several particles.

The electron's intrinsic g factor is extremely close to 2, and if we ignore the small discrepancy for now, we are led to imagine that the electron is either a pointlike particle or a composite of smaller particles, each of which has the same charge-to-mass ratio. The neutron does have a nonvanishing dipole moment, so its zero g factor suggests that it is a composite of other particles whose charges cancel. The proton's g factor is quite different from 2, so we infer that it, too, is composite. The current theory is that protons and neutrons are clusters of particles called quarks. Quarks come in different types, and the different types have different values of q/m .

The magnetic dipole moment of the proton is of considerable importance in our lives because of its use in the MRI (magnetic resonance imaging) scans used in medicine. As described on p. 197, a large DC magnetic field, generated by superconducting magnets, is used to cause the protons in the body's hydrogen atoms to align partially (about 10^{-5} of full alignment). These magnetic moments are then manipulated and observed using AC fields.

From a physicist's point of view, it is also remarkable that we can infer these facts about the internal structures of neutrons and protons without having to do any experiments that directly probe their interior structure. We don't need a super-powerful microscope, nor do we need a particle accelerator that can supply enough energy to shake up their internal structure, like shaking a gift-wrapped box to tell what's inside. Merely by measuring the external, aggregate properties of the "box," we can get clues about the structure inside.¹

11.4.3 Field of a dipole

An electric dipole, unlike a magnetic one, can be built out of two opposite monopoles, i.e., charges, separated by a certain distance, and it is then straightforward to show by vector addition that the field of an electric dipole, far away, is

$$E_z = kD(3\cos^2\theta - 1)r^{-3}$$

$$E_R = kD(3\sin\theta \cos\theta)r^{-3},$$

¹This is closely analogous to the Tolman-Stewart experiment (example 4, p. 94), in which the subatomic structure of metals was probed by measuring inertial effects in an electric circuit.

where r is the distance from the dipole to the point of interest, θ is the angle between the dipole vector and the line connecting the dipole to this point, and E_z and E_R are, respectively, the components of the field parallel to and perpendicular to the dipole vector. We have already found this field in the special cases of $\theta = \pi/2$ (example 3, p. 51) and $\theta = 0$ (problem 9, p. 69).

This is the field pattern that exists far away from the dipole, in empty space. Because the vacuum form of Maxwell's equations treats the electric and magnetic fields totally symmetrically, the magnetic field of a magnetic dipole has to have the same form. With the correct constant of proportionality, it turns out to be

$$B_z = \frac{km}{c^2} (3 \cos^2 \theta - 1) r^{-3}$$

$$B_R = \frac{km}{c^2} (3 \sin \theta \cos \theta) r^{-3}.$$

Discussion question

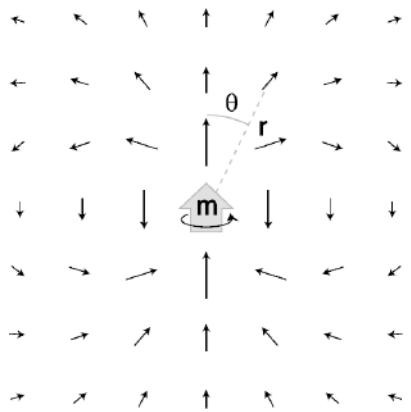
- A** Find the regions of three-dimensional space in which the magnetic field of a dipole is (1) in the same direction as the dipole vector (parallel), (2) in the opposite direction (antiparallel), and (3) perpendicular to it.

11.5 Magnetic fields found by summing dipoles

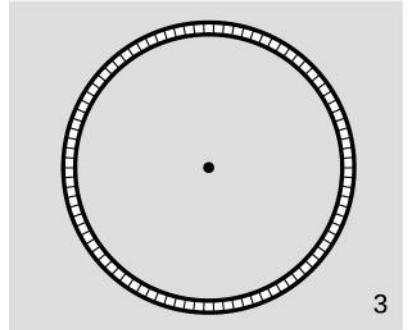
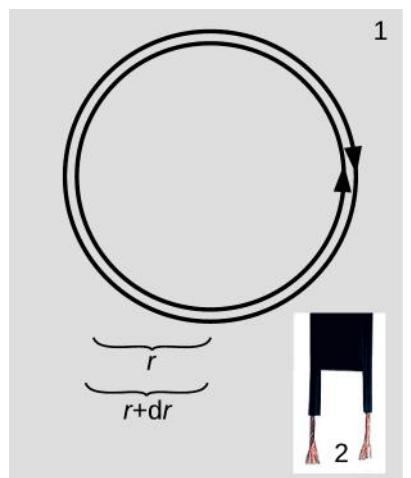
Many real-world electromagnets are built out of multiple circular loops of wire. In this section we use a trick to calculate the useful result for the field at the center of a single circular loop. Usually I'm not a big fan of tricks, but we'll see later that this trick can be generalized in a useful way.

We start by considering a problem, figure m/1, that looks harder but is actually easier. We have not one but two loops of wire, with slightly different radii r and $r + dr$. Here the "d" just means "a little bit of," i.e., writing the radii this way is just a way of saying that they're almost, but not quite, the same. The currents in the two rings are both I , but they flow in opposite directions. This is actually a reasonably realistic setup; the inset m/2 shows a type of cable, called "twin lead," that is often used in this way, with current flowing one way through one conductor and then coming back through the other conductor in the opposite direction. Of course in the real circuit we would have to have a battery and connections to the loops from the outside, and we would probably hook up the two loops in series so that their currents were guaranteed to be exactly equal in absolute value. These complications are not shown in the diagram.

The trick, as shown in m/3, is to take the circular strip between the rings and break it up into imaginary squares, then place a square



I / The field of a dipole.



m / Two counterrotating rings of current. The inset photo 2 shows a twin lead cable with the insulation stripped off of its ends.

current loop with current I on each one. These are all dipoles, and each one contributes a field at the center which we can calculate by plugging $\theta = \pi/2$ in to our expressions from section 11.4.3. The result is that each tiny dipole dm contributes $-(k/c^2r^3) dm$, where the minus sign means that the field points out of the page (in the opposite direction compared to the vector dm). The total field is found by adding up all these small contributions to the field,

$$B = - \int \frac{k}{c^2 r^3} dm.$$

This integral, like any integral, represents a sum of infinitely many infinitesimal things. We're not integrating with respect to some variable m , though; dm here just means an infinitesimal dipole moment of one of the squares. But we don't actually need any calculus to do this integral. Moving all the constants outside and substituting $dm = I dA$, we have

$$B = -\frac{kI}{c^2 r^3} \int dA,$$

where $\int dA$ is the area of the strip, $A = (\text{width})(\text{circumference}) = (dr)(2\pi r)$. Our final result is

$$B = -\frac{2\pi kI}{c^2 r^2} dr \quad [\text{field at the center of figure m/1}].$$

We weren't actually that excited about finding the field in the somewhat artificial example of figure m. What would be more useful would be to find the field of a *single* circular loop. Now we just play a similar trick again. We imagine an infinite set of concentric rings, extending from some radius a all the way out to infinity. The current on the outer edge of each ring is canceled out by the current in the overlapping inner edge of the next ring, so that the only loop of current that *doesn't* cancel out is the very innermost one, at $r = a$. The result is then that the field at the center of a circular loop is $\int_a^\infty [-(2\pi kI)/(c^2 r^2)] dr$, or

$$B = \frac{2\pi kI}{c^2 a} \quad [\text{field at the center of a ring of current}].$$

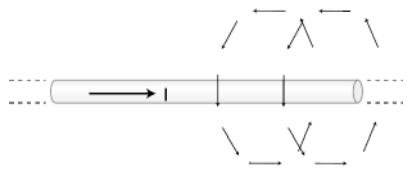
The positive sign means that the direction of the field is right-handed, e.g., out of the page if the current is counterclockwise.

11.6 Magnetic fields for some practical examples

Figure n shows the equations for some of the more commonly encountered configurations in which wires produce a magnetic field, with illustrations of their field patterns. Of these three results, we've only previously derived the first and a special case of the second. The remaining derivations are given later in the book, but the results are presented together at this point for reference.

Field created by a long, straight wire carrying current I:

$$B = \frac{k}{c^2} \cdot \frac{2I}{r}$$



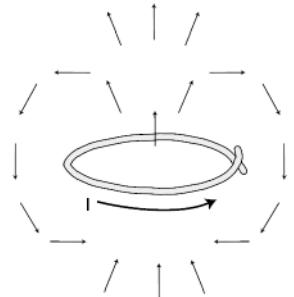
Here r is the distance from the center of the wire. The field vectors trace circles in planes perpendicular to the wire, in a direction given by a right-hand rule where the thumb is the current and the fingers are the field. The form of this result was derived in example 2, p. 121, and the unitless factor of 2 in example 6, p. 181.

Field created by a single circular loop of current:

The field vectors form a dipole-like pattern, coming through the loop and back around on the outside. The orientation of the loops is such that in the middle region there is a right-hand relationship in which the thumb is the field; or, alternatively, one can use the same right-hand rule as for a straight wire, applying it to the area close to the wire. There is no simple equation for a field at an arbitrary point in space, but for a point lying *along the central axis* perpendicular to the loop, the field is

$$B = \frac{k}{c^2} \cdot 2\pi Ib^2 (b^2 + z^2)^{-3/2},$$

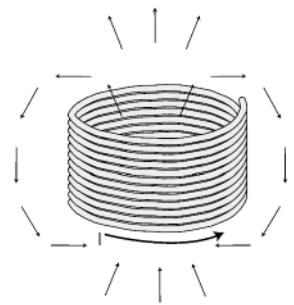
where b is the radius of the loop and z is the distance of the point from the plane of the loop.



Field created by a solenoid (cylindrical coil):

The field pattern is similar to that of a single loop, but for a long solenoid the field lines become very straight on the inside of the coil and on the outside immediately next to the coil. For a sufficiently long solenoid, the interior field also becomes very nearly uniform, with a magnitude of

$$B = \frac{k}{c^2} \cdot 4\pi IN/\ell,$$



n / Some magnetic fields.

where N is the number of turns of wire and ℓ is the length of the solenoid. This result is derived in example 4, p. 350. The field near the mouths or outside the coil is not constant, and is more difficult to calculate (problem 13, p. 283). For a long solenoid, the exterior field is much smaller than the interior field.

Some other cases of interest can be solved by superposing the fields above. An example is the Helmholtz coil (problem 6, p. 280).

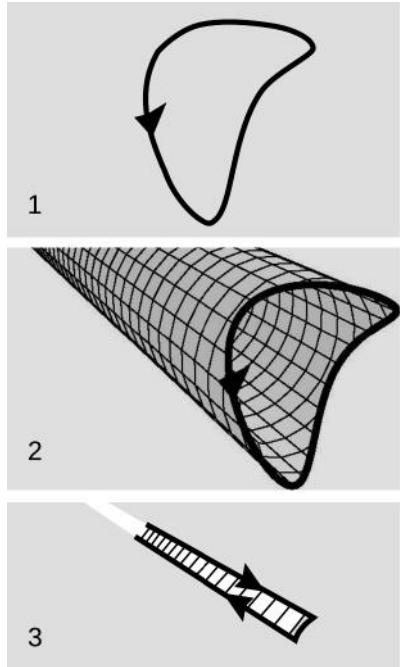
self-check A

1. Let a current-carrying wire lie along the x axis, carrying current in the positive x direction. At points on the y axis, which component of the field is nonzero? Sketch this component as a function of y .
2. Sketch the function $B_z(z)$ for a circular loop as described above. ▷

Answer, p. 432

11.7 ★ The Biot-Savart law (optional)

In sec. 11.5, p. 271, we were able to find the magnetic field at the center of a current loop by the trick of turning the current distribution into a superposition of small, square dipoles. It's usually a waste of time to learn a trick that only works in one case, but this trick can be extended to apply more generally. Consider the wire loop shown in figure o/1, which has a randomly chosen, asymmetric shape like a potato chip. If it carries a steady current, it will create a static magnetic field pattern. (The assumption of a steady current is necessary, since otherwise, e.g., it could act like an antenna and radiate electromagnetic waves.)



o / Extending the dipole trick to handle a randomly chosen, asymmetric current loop.

Figure o/2 shows an idea for extending our dipole trick to handle this case. We form an imaginary tube that starts on the loop and extends off to infinity, then split it up into strips like railroad tracks, where each railroad track contains an infinite number of square dipoles, o/3. The key here is that although each track has a pair of long, straight rails that bring current in from infinity and then send it back out, the currents along these rails cancel when we overlay each track with its neighbor. Therefore all we really need to do is find the magnetic field of *one* such semi-infinite strip, and by adding it up we can find the field of an arbitrary object like the potato-chip loop.

The field of such a strip is probably pretty complicated, and likely to be mainly the field due to the two long rails. However, the currents in the rails are destined to cancel out when we add everything up, sort of like a politically opposed married couple who vote in every election and cancel each other out — they might as well have stayed home. We therefore conjecture that the correct final result can be found by adding up fields that depend only on the properties of the little end-caps. This is not guaranteed to be correct, for the reasons described in example 7 on p. 265, but let's go ahead based on this questionable assumption and see where it gets us.

Given some point in space, we want to find the contribution to the field at that point coming from a particular end-cap. Let the vector from the end-cap to our point be \mathbf{r} . The contribution to the field has to be of the form $\mathbf{f}(\mathbf{j}, \mathbf{r}) dv$, where the dv is the volume of the wire constituting the end-cap. The mystery function has to be proportional to its two vector inputs, and as its output it has to give a vector with units of tesla. This severely constrains the form of the function. The only rotationally invariant way to combine two vectors like this is the cross product, and the only way of getting the right units is by throwing in a factor of $k c^{-2} r^{-3}$. The only wiggle room is a possible unitless factor in front. This unitless factor turns out to be 1, which we will prove later, in example 11, by comparing with a configuration whose field we already know. So

the contribution to the field from this end-cap is

$$\frac{k}{c^2} \frac{\mathbf{j} \times \mathbf{r}}{r^3} dv.$$

Within the small volume of the end-cap, the integrand is constant, and the end-cap is approximately straight, so that it forms a cylinder with volume $dv = dA dl$, where dA is the cross-sectional area. We also assume that, within this short segment of wire, the current flows in the direction of the wire and is uniform across the wire's cross-sectional area. This allows us to rewrite the end-cap's field as

$$\frac{Ik}{c^2} \frac{dl \times \mathbf{r}}{r^3}.$$

Integrating over the contributions of all the end-caps gives the formula known as the Biot-Savart law (rhymes with "Leo bazaar"),

$$\mathbf{B} = \frac{Ik}{c^2} \int \frac{dl \times \mathbf{r}}{r^3}.$$

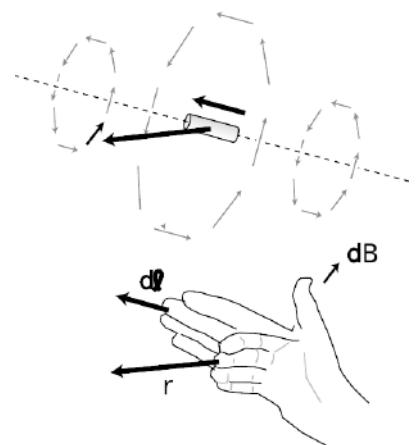
The field at the center of a circular loop example 11

In section 11.5 we had to use a trick to find the field at the center of a circular loop of current of radius a . The Biot-Savart law routinizes the trick for us and eliminates the need for that kind of creativity. Dividing the loop into many short segments, each dl is perpendicular to the \mathbf{r} vector that goes from it to the center of the circle, and every \mathbf{r} vector has magnitude a . Therefore every cross product $dl \times \mathbf{r}$ has the same magnitude, adl , as well as the same direction along the axis perpendicular to the loop. The field is

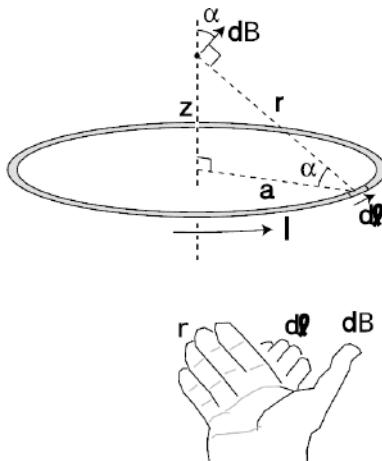
$$\begin{aligned} B &= \int \frac{kIadl}{c^2 a^3} \\ &= \frac{kI}{c^2 a^2} \int dl \\ &= \frac{kI}{c^2 a^2} (2\pi a) \\ &= \frac{2\pi kI}{c^2 a}. \end{aligned}$$

The fact that the field in example 11 comes out the same as in the previous calculation verifies that we have the right unitless factor in the Biot-Savart law.

Although the Biot-Savart law seems to have given us a correct result, we highlighted an assumption in its derivation that was not guaranteed to be correct. To give a full proof of the law, we would really need to prove that when we plug the result back in to Maxwell's equations, it gives a solution. That unfortunately requires a level of vector calculus beyond the scope of this book.



p / The geometry of the Biot-Savart law. The small arrows show the result of the Biot-Savart law at various positions relative to the current segment dl . The Biot-Savart law involves a cross product, and the right-hand rule for this cross product is demonstrated for one case.



q / Example 12.

Out-of-the-plane field of a circular loop example 12

▷ What is the magnetic field of a circular loop of current at a point on the axis perpendicular to the loop, lying a distance z from the loop's center?

▷ Again, let's write a for the loop's radius. The \mathbf{r} vector now has magnitude $\sqrt{a^2 + z^2}$, but it is still perpendicular to the $d\ell$ vector. By symmetry, the only nonvanishing component of the field is along the z axis,

$$\begin{aligned} B_z &= \int |d\mathbf{B}| \cos \alpha \\ &= \int \frac{kI r d\ell}{c^2 r^3} \frac{a}{r} \\ &= \frac{kI a}{c^2 r^3} \int d\ell \\ &= \frac{2\pi k I a^2}{c^2 (a^2 + z^2)^{3/2}}. \end{aligned}$$

For $z = 0$ we recover the result of example 11. For $z \gg a$, the field is that of a dipole with the correct dipole moment (ex. 11A, p. 286).

Notes for chapter 11

Z264 Divergence of a curl

The divergence of a curl is zero

The kind of field that looks like it would have a nonvanishing value of this operation is something like $\mathbf{F} = zx\hat{\mathbf{y}}$, the idea being that we take a field like $x\hat{\mathbf{y}}$ that has a curl in the z direction, and then we give it some z dependence so that there could be a divergence. In fact, any field that is differentiable near the origin can have its components approximated in that neighborhood by a function of this general form (a second-order mixed polynomial), so if we can prove that $\text{div}(\text{curl } \mathbf{F}) = 0$ for this particular \mathbf{F} , it follows that the div of a curl is zero for any \mathbf{F} .

Now we show based on symmetry that $\text{div}(\text{curl } \mathbf{F}) = 0$ for this \mathbf{F} . Suppose we rotate our coordinate system by 180 degrees about the y axis. This doesn't change the field or its components, but the curl lies in the x - z plane, so the output of the curl operation has its components sign-flipped because of the new coordinate system used to describe it. This is an over-all reversal of the curl's direction, and since the div is linear, the effect is to sign-flip $\text{div}(\text{curl } \mathbf{F})$.

But the divergence is a scalar, so the final result of taking $\text{div}(\text{curl } \mathbf{F})$ cannot change just because we change our coordinates.

We have proved that $\text{div}(\text{curl } \mathbf{F})$ flips its sign, but also that it doesn't change its sign. This is only possible if it is zero, as claimed.

Z264 Order of derivatives

Typically it's OK to freely interchange the order of derivative operations

Of course an example doesn't prove a general rule, but let's consider a simple example in order to get the idea of what is being discussed. We have

$$\frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} (xy) \right) = \frac{\partial}{\partial x} x = 1$$

but also

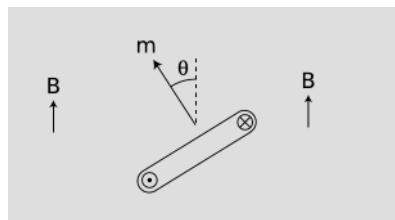
$$\frac{\partial}{\partial y} \left(\frac{\partial}{\partial x} (xy) \right) = \frac{\partial}{\partial y} y = 1.$$

In this example, it didn't matter which derivative we applied first. A theorem called Clairaut's theorem says that if a function $f(x, y)$ is well-behaved, in the sense that all its second derivatives ($\partial^2 f / \partial x^2$, $\partial(\partial f / \partial x) / \partial y$, and $\partial(\partial f / \partial y) / \partial x$) are continuous at a certain point, then the first derivatives can be interchanged at that point, so that $\partial(\partial f / \partial x) / \partial y = \partial(\partial f / \partial y) / \partial x$.

Because derivative operators like div, grad, and curl can be expressed in terms of partial derivatives, it follows that they can also be interchanged under the same conditions.

Z268 Magnetic dipole moment of a current loop

The magnetic dipole moment of a square current loop is given by $\mathbf{m} = IA$.



Side view of a square current dipole, with the direction of the current as indicated, coming out of the page on one side and going back in on the other. The dipole is immersed in an externally imposed uniform field shown by the vertical arrows.

Consider the geometry shown in figure j. Let the mobile charge carriers in the wire have linear density λ , and let the sides of the loop have length h , so that we have $I = \lambda v$. We want to show that $m = IA = h^2 \lambda v$ is consistent with the definition of the dipole moment $U = -\mathbf{m} \cdot \mathbf{B}$, where \mathbf{B} is an externally applied field. We do this by computing the torque and then finding the work done when the dipole is reoriented.

The only nonvanishing torque comes from the

forces on the left and right sides. The currents in these sides are perpendicular to the field, so the magnitude of the cross product $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$ is simply $|\mathbf{F}| = qvB$. The torque supplied by each of these forces is $\mathbf{r} \times \mathbf{F}$, where the lever arm \mathbf{r} has length $h/2$, and makes an angle θ with respect to the force vector. The magnitude of the total torque acting on the loop is therefore

$$\begin{aligned} |\boldsymbol{\tau}| &= 2 \frac{h}{2} |\mathbf{F}| \sin \theta \\ &= h q v B \sin \theta, \end{aligned}$$

and substituting $q = \lambda h$ and $v = m/h^2\lambda$, we have

$$\begin{aligned} |\boldsymbol{\tau}| &= h \lambda h \frac{m}{h^2 \lambda} B \sin \theta \\ &= m B \sin \theta. \end{aligned}$$

The work done to reorient the dipole is $W = \int \boldsymbol{\tau} d\theta = -mB \cos \theta$ (ignoring the irrelevant constant of integration), and this is the same as $U = -\mathbf{m} \cdot \mathbf{B}$.

Problems

Key

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

1 The photo shows the cross-section of an electrical transmission line designed to be hung in the air between towers. To prevent the cable from sagging too much, there is a central core of radius a made of a carbon-glass composite, which is stronger and lighter than steel, but nonconductive. Surrounding this is a conducting aluminum sheath with outer radius b . Because the frequency is low, the current density is nearly uniform.

- (a) If the cable carries current I , find the magnitude of the current density \mathbf{j} . ✓
- (b) Evaluate your answer numerically for $a = 4.76$ mm and $b = 14.07$ mm, at this cable's nominal current-carrying capacity of 1.00 kA. ✓



Problem 1.

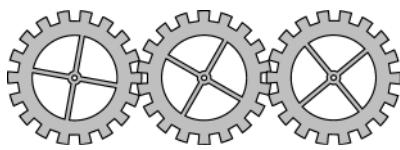
2 Magnetic dipole 1 has its dipole moment oriented along the z axis, so that its z component m_1 is either positive or negative, and its x and y components are zero. It interacts with a second dipole m_2 , also purely along the z axis. They lie on the x axis at distance r from one another. Find the energy of their interaction with each other. What is their stable orientation? You should find that the sign of this result agrees with experience from playing with a pair of bar magnets, as well as with the result of ch. 5, problem 2a, p. 143. ✓

3 Deuterium is an isotope of hydrogen in which the nucleus has one proton and one neutron. The nucleus is referred to as a deuteron. The deuteron has angular momentum and magnetic moment

$$L = 1.05457 \times 10^{-34} \text{ J}\cdot\text{s} \quad \text{and}$$
$$m = 4.33 \times 10^{-27} \text{ A}\cdot\text{m}^2.$$

Find the g factor of the deuteron. You should find that it roughly makes sense if we consider the nucleus as two pointlike particles with identical masses but with one particle's charge being zero.

✓

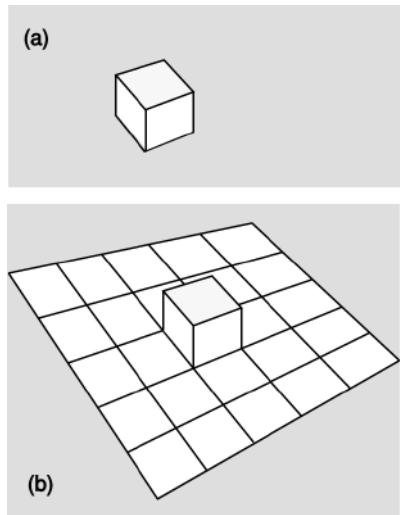


Problem 4.

4 N identical gears of radius r are arranged with their axes parallel and coplanar. The figure shows the $N = 3$ case as an example. Each gear is an insulator, and has charge q distributed uniformly about its circumference. If the system spins at frequency f , find the total dipole moment. How is this different from an example like the one in figure o/3, p. 274?

Remark: This is not an unreasonable model of the magnetic properties of a linear molecule, if the magnetic interactions are like the ones described in problem 2.

▷ Hint, p. 425



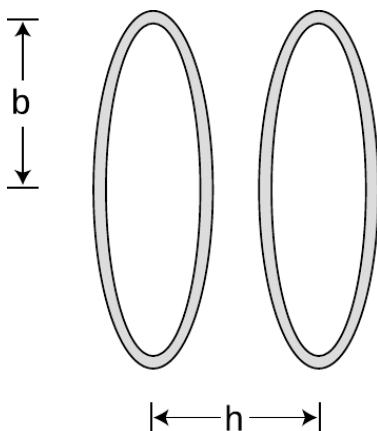
Problem 5.

5 (a) The first figure shows a cube with six sides, including a floor underneath. Each square has sides of length b and is a current loop carrying current I in the same orientation, i.e., an ant exploring the outside surface and inspecting all the current loops will see each current rotating in the same direction as it stands on that panel of the box. Why is the total dipole moment not $6Ib^2$?

(b) The second figure shows a landscape consisting of a 5×5 grid of squares, interrupted by a “little house on the prairie” in the middle: a cube with four walls and a roof. The cube does not have a floor, so the total area is $29b^2$. Find the magnitude of the total magnetic dipole moment.

Remark: This is not quite as silly and artificial as it might seem. In condensed matter physics, it's common to have things like surface layers of dipoles, and it's also common to have defects in such a surface such as bumps and scratches.

✓



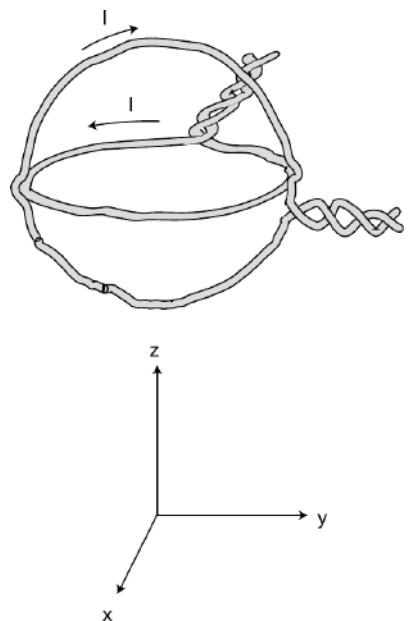
Problem 6.

6 A Helmholtz coil is defined as a pair of identical circular coils lying in parallel planes and separated by a distance, h , equal to their radius, b . Each coil has N turns of wire. Current circulates in the same direction in each coil, so the fields tend to reinforce each other in the interior region. This configuration has the advantage of being fairly open, so that other apparatus can be easily placed inside and subjected to the field while remaining visible from the outside. The choice of $h = b$ results in the most uniform possible field near the center. Find the field at the center.

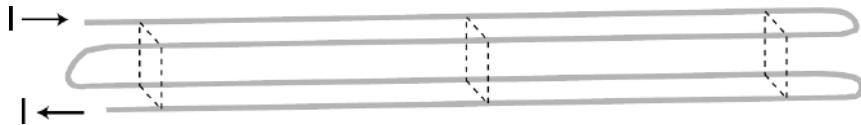
✓

7 The figure shows a nested pair of circular wire loops used to create magnetic fields. (The twisting of the leads is a practical trick for reducing the magnetic fields they contribute, so the fields are very nearly what we would expect for an ideal circular current loop.) The coordinate system below is to make it easier to discuss directions in space. One loop is in the $y - z$ plane, the other in the $x - y$ plane. Each of the loops has a radius of 1.0 cm, and carries 1.0 A in the direction indicated by the arrow.

- (a) Calculate the magnetic field that would be produced by *one* such loop, at its center. ✓
- (b) Describe the direction of the magnetic field that would be produced, at its center, by the loop in the $x - y$ plane alone.
- (c) Do the same for the other loop.
- (d) Calculate the magnitude of the magnetic field produced by the two loops in combination, at their common center. Describe its direction. ✓



Problem 7.



Problem 8.

8 Four long wires are arranged, as shown, so that their cross-section forms a square, with connections at the ends so that current flows through all four before exiting. Note that the current is to the right in the two back wires, but to the left in the front wires. If the dimensions of the cross-sectional square (height and front-to-back) are b , find the magnetic field (magnitude and direction) along the long central axis. ✓

9 This problem will lead you through the steps of applying the Biot-Savart law to prove that the magnetic field of a long, straight wire has magnitude

$$B = \frac{2kI}{c^2 R}.$$

Almost everything in this equation has to be the way it is because of units, the only exception being the unitless factor of 2, so this problem amounts to proving that it really does come out to be 2, a fact that we previously proved in example 6 on page 181, using relativity.

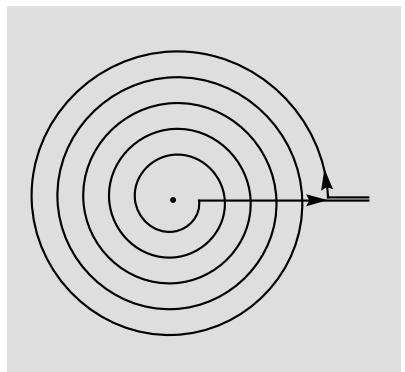
- (a) Set up the integral prescribed by the Biot-Savart law, and simplify it so that it involves only scalar variables rather than a vector cross product, but do not evaluate it yet.
- (b) Your integral will contain several different variables, each of which is changing as we integrate along the wire. These will probably include a position on the wire, a distance from the point on the wire to the point at which the field is to be found, and an angle between the wire and this point-to-point line. In order to evaluate the integral, it is necessary to express the integral in terms of only one of these variables. It's not obvious, but the integral turns out to be easiest to evaluate if you express it in terms of the angle and eliminate the other variables. Do so. Note that the $d\dots$ part of the integral has to be reexpressed in the same way we would do any time we attacked an integral by substitution ("u-substitution").
- (c) Pulling out all constant factors now gives a definite integral. Evaluate this integral, which you should find is a trivial one, and show that it equals 2.

10 A square current loop with sides of length $2h$ carries current I , creating a magnetic field B_{square} at its own center. We wish to compare this with the field B_{circle} of a circular current loop of radius h . Find $B_{\text{square}}/B_{\text{circle}}$. ✓

11 A regular polygon with n sides can be inscribed within a circle of radius R and can have a circle inscribed inside it with radius h . Let $\rho = \sqrt{hR}$ be the geometric mean of these two radii. A current loop is constructed in the shape of a regular n -gon. Show that the magnetic field at the center can be calculated in a simple way from the perimeter and ρ , and make sense of the result in the extreme cases $n = 2$ (a degenerate polygon enclosing no area) and $n \rightarrow \infty$. *

12 Magnet coils are often wrapped in multiple layers. The figure shows the special case where the layers are all confined to a single plane, forming a spiral. Since the thickness of the wires (plus their insulation) is fixed, the spiral that results is a mathematical type known as an Archimedean spiral, in which the turns are evenly spaced. The equation of the spiral is $r = w\theta$, where w is a constant. For a spiral that starts from $r = a$ and ends at $r = b$, show that the field at the center is given by $(kI/c^2w) \ln b/a$.

▷ Solution, p. 430 *



Problem 12.

13 Let the interior field of a certain infinite solenoid be B_0 . Now suppose that we build a finite solenoid with all the same design parameters except that it has a finite length, and consider the field B at a point on the axis. Show by integrating the field of a loop of current that

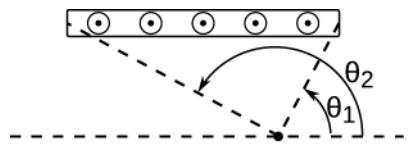
$$\frac{B}{B_0} = \frac{\cos \theta_1 - \cos \theta_2}{2},$$

where the angles θ_1 and θ_2 are defined in the figure.

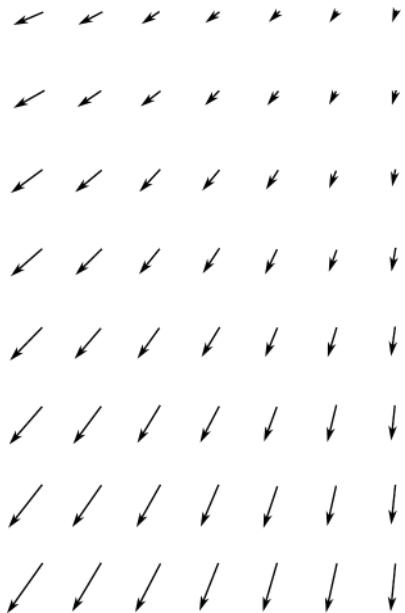
*

14 In problem 26, p. 111, you characterized the curl and divergence of the field shown in the figure.

- (a) Suppose someone tells you that the figure shows an electric field. What further interpretation can you give of the physical situation?
- (b) What if they say instead that it shows a current density?
- (c) What if they say that it shows a magnetic field?



Problem 13.



Problem 14.

Lab 11: Charge-to-mass ratio of the electron

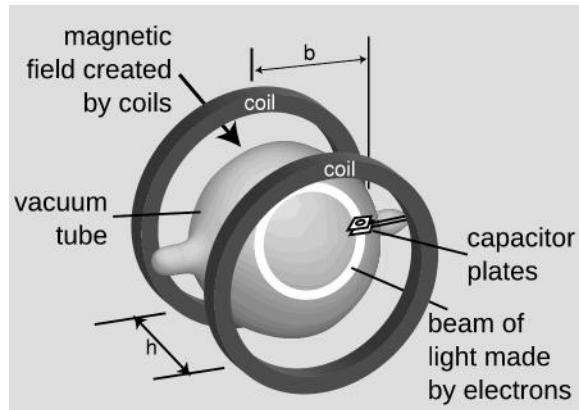
Apparatus

vacuum tube with Helmholtz coils
high-voltage power supply
DC power supply
multimeter

Goal: Measure the q/m ratio of the electron.

Why should you believe electrons exist? By the turn of the twentieth century, not all scientists believed in the literal reality of atoms, and few could imagine smaller objects from which the atoms themselves were constructed. Over two thousand years had elapsed since the Greeks first speculated that atoms existed based on philosophical arguments without experimental evidence. During the Middle Ages in Europe, “atomism” had been considered highly suspect, and possibly heretical. Finally by the Victorian era, enough evidence had accumulated from chemical experiments to make a persuasive case for atoms, but subatomic particles were not even discussed.

If it had taken two millennia to settle the question of atoms, it is remarkable that another, subatomic level of structure was brought to light over a period of only about five years, from 1895 to 1900. Most of the crucial work was carried out in a series of experiments by J.J. Thomson, who is therefore often considered the discoverer of the electron.



The vacuum tube apparatus used in this lab.

In this lab, you will carry out a variation on a crucial experiment by Thomson, in which he measured the ratio of the charge of the electron to its mass, q/m . The basic idea is to observe a beam of electrons in a region of space where there is an approximately uniform magnetic field, B . The electrons are emitted perpendicular to the field, and, it turns out, travel in a circle in a plane perpendicular to it. The force of the magnetic field on the electrons is

$$F = qvB \quad , \quad (1)$$

directed towards the center of the circle. Their acceleration is

$$a = \frac{v^2}{r} \quad , \quad (2)$$

so using $F = ma$, we can write

$$qvB = \frac{mv^2}{r} \quad . \quad (3)$$

If the initial velocity of the electrons is provided by accelerating them through a voltage difference V , they have a kinetic energy equal to qV , so

$$\frac{1}{2}mv^2 = qV \quad . \quad (4)$$

From equations 3 and 4, you can determine q/m . Note that since the force of a magnetic field on a moving charged particle is always perpendicular to the direction of the particle’s motion, the magnetic field can never do any work on it, and the particle’s KE and speed are therefore constant.

You will be able to see where the electrons are going, because the vacuum tube is filled with a hydrogen gas at a low pressure. Most electrons travel large distances through the gas without ever colliding with a hydrogen atom, but a few do collide, and the atoms then give off blue light, which you can see. Although I will loosely refer to “seeing the beam,” you are really seeing the light from the collisions, not the beam of electrons itself. The manufacturer of the tube has put in just enough gas to make the beam visible; more gas would make a brighter beam, but would cause it to spread out and become too broad to measure it precisely.

The field is supplied by an electromagnet consisting of two circular coils, each with 130 turns of wire (the same on all the tubes we have). The coils are placed on the same axis, with the vacuum tube at the center. A pair of coils arranged in this type of geometry are called Helmholtz coils. Such a setup

provides a nearly uniform field in a large volume of space between the coils, and that space is more accessible than the inside of a solenoid.

Setup

Heater circuit: As with all vacuum tubes, the cathode is heated to make it release electrons more easily. There is a separate low-voltage power supply built into the high-voltage supply. It has a set of green plugs that, in different combinations, allow you to get various low voltage values. Use it to supply 6 V to the terminals marked “heater” on the vacuum tube. The tube should start to glow.

Electromagnet circuit: Connect the other DC power supply, in series with an ammeter, to the terminals marked “coil.” The current from this power supply goes through both coils to make the magnetic field. Verify that the magnet is working by using it to deflect a nearby compass.

High-voltage circuit: Connect the high voltage supply to the terminals marked “anode.” Ask your instructor to check your circuit. Now plug in the HV supply and turn up the voltage to 300 V. You should see the electron beam. If you don’t see anything, try it with the lights dimmed.

Observations

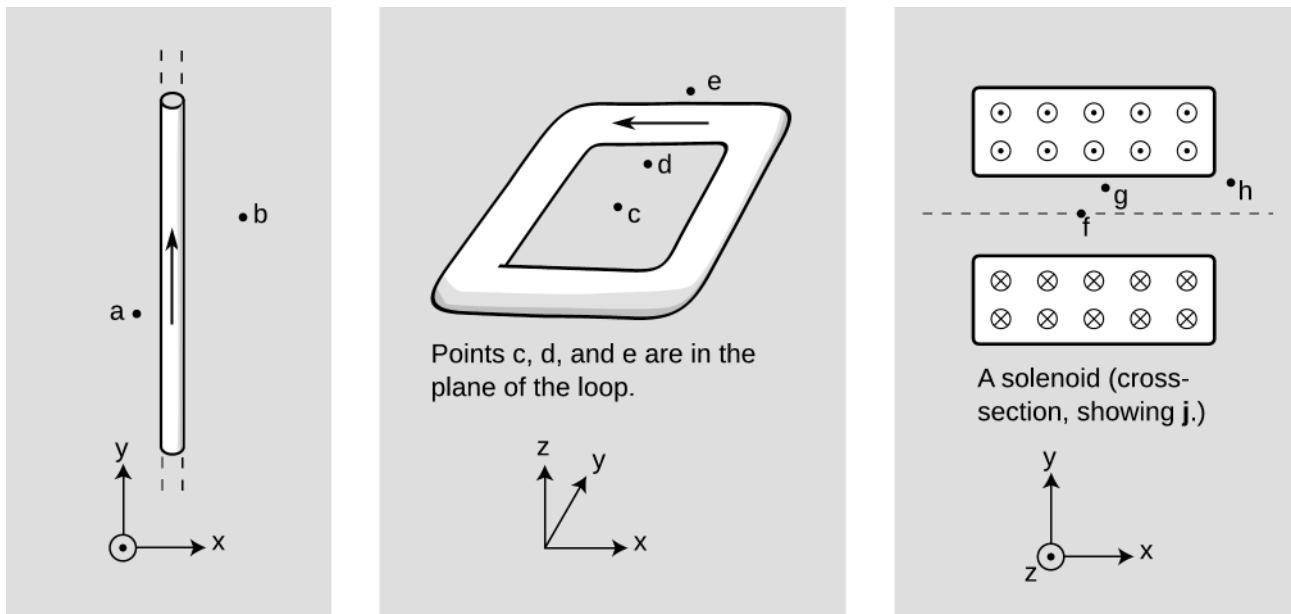
Make the necessary observations in order to find q/m , carrying out your plan to deal with the effects of the Earth’s field. The high voltage is supposed to be 300 V, but to get an accurate measurement of what it really is you’ll need to use a multimeter rather than the poorly calibrated meter on the front of the high voltage supply.

The beam can be measured accurately by using the glass rod inside the tube, which has a centimeter scale marked on it.

Be sure to compute q/m before you leave the lab. That way you’ll know you didn’t forget to measure something important, and that your result is reasonable compared to the currently accepted value.

Exercise 11A: Currents and magnetic fields

1. Find the directions of the magnetic fields at points a-h. Describe them using the given coordinate systems, e.g., “+x.”



2. The on-axis field of a circular current loop is shown in example 12, p. 276, to be

$$B = \frac{2\pi k I a^2}{c^2(a^2 + z^2)^{3/2}}.$$

- (a) Show by comparing with the field of a long, straight wire that the units make sense.
 (b) Show that the field for $z \gg a$ is that of a dipole, including the correct constant factor.

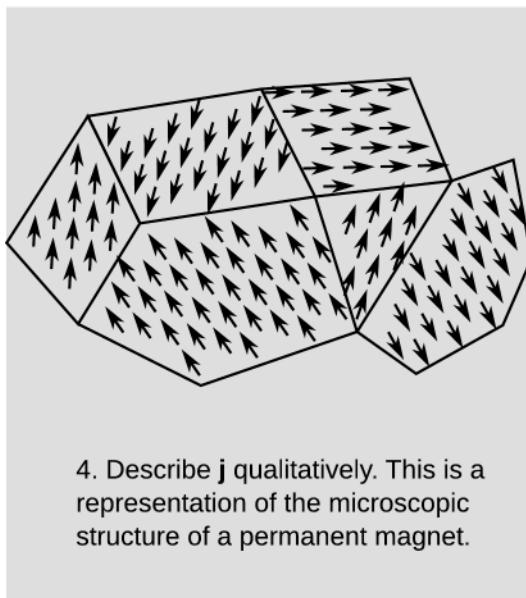
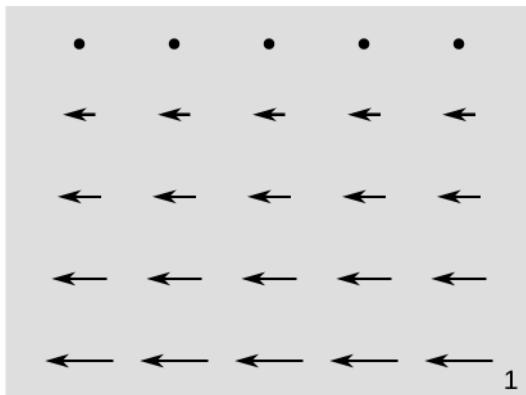
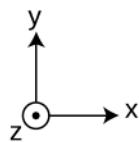
3. The diagrams below show side views of a square current loop immersed in an external magnetic field. The first one (reproduced from fig. j, p. 268) shows the loop's dipole moment. The second one labels the currents along the four sides of the square.



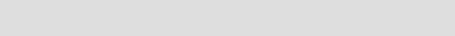
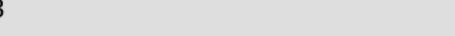
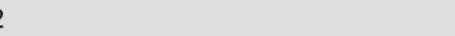
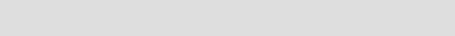
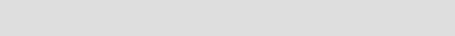
- (a) Find the orientation that would minimize the energy $U = -\mathbf{m} \cdot \mathbf{B}$.
 (b) Use the right-hand rule to find the force on each of the four edges.
 (c) Verify that the total force is zero.
 (d) Find the four torques, and verify that the direction of the total torque is such that it would tend to align the loop's dipole moment with the field.

Turn the page.

4. Each figure shows a static \mathbf{B} field in the x - y plane. The field is independent of z . Describe the current density \mathbf{j} in each case. A coordinate system is provided for convenience of description.



4. Describe \mathbf{j} qualitatively. This is a representation of the microscopic structure of a permanent magnet.

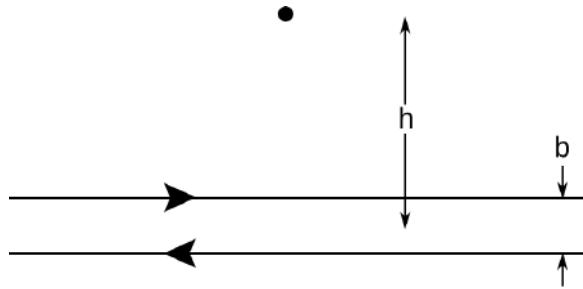


2

3

Exercise 11B: The magnetic field of twin-lead cable

In section 11.5, p. 271, we used a trick to find the field of a circular loop of twin-lead cable, and then the field of a single circular loop of current. In this exercise you will use the same technique to find the field of a long, straight piece of twin-lead cable, and then the field of a single, long, straight wire. This can be checked against the result of example 6, p. 181.



A piece of twin lead cable.

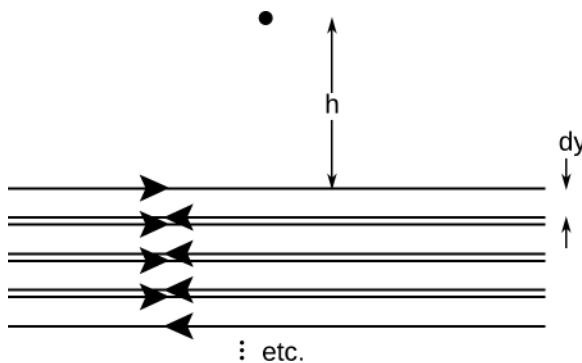
The figure shows the setup. The two conductors are separated by a distance b which is small compared to h , and they extend infinitely far to the left and right. We wish to find the field B at the point indicated in the figure. To start out with, we imagine, as in section 11.5, dividing up the space between the conductors into little rectangles, and we then set up an integral for the field in which we integrate over these tiny rectangles:

$$B = - \int \frac{k}{c^2 r^3} dm.$$

1. Use the relation $dm = I dA$ to make this into an integral over the area between the conductors. Then let the width of each little rectangle be dx , and make this into an integral over x .

2. This integral is not yet really a definite integral, because it has the parameter h inside. Change to a new, unitless variable $u = x/h$ and rewrite the result in terms of an actual definite integral, with other parameters appearing outside the integral as a single multiplicative factor.

3. You should find that the definite integral has the form $\int_{-\infty}^{\infty} (1+u^2)^{-3/2} du$. It's a waste of time to do an integral like this by hand. There is a nice piece of open-source software called Maxima that can do this kind of thing. You can download it for free, but for now we'll find it more convenient to use it through a web interface on a server, at maxima.cesga.es. Your definite integral will look like this: `integrate((1+u^2)^(-3/2),u,-inf,inf);` (note the semicolon). Find its value and complete the calculation of the field of the twin lead cable.



Superimposing infinitely many twin-lead wires to fill the entire half-plane $y < 0$.

4. Now, as shown in the second figure, we superimpose infinitely many twin-lead wires, with the result that the only current that doesn't cancel out with a neighbor is the one on the x axis. The role previously played by b is now played by dy , and h is now to be replaced with $h - y$, so that the result from part 3 is now dB , the infinitesimal field contributed by one of the strips. Integrate to find the field of a wire on the x axis, and compare with the result of example 6, p. 181.

AC circuits

Chapter 12

Review of oscillations, resonance, and complex numbers

The long road leading from the light bulb to the computer started with one very important step: the introduction of feedback into electronic circuits. Although the principle of feedback has been understood and applied to mechanical systems for centuries, and to electrical ones since the early twentieth century, for most of us the word 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.

12.1 Review of complex numbers

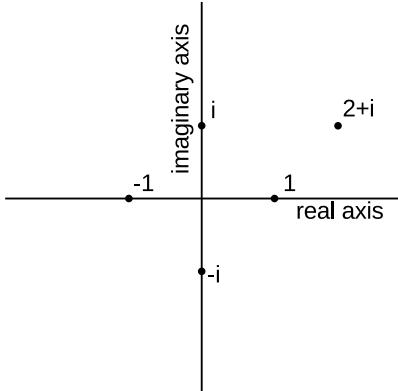
Positive feedback causes exponential behavior, while negative feedback causes oscillations. The complex number system makes it possible to describe all of these phenomena in a simple and unified way. For a more detailed treatment of complex numbers, see ch. 3 of James Nearing's free book at physics.miami.edu/~nearing/mathmethods.

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

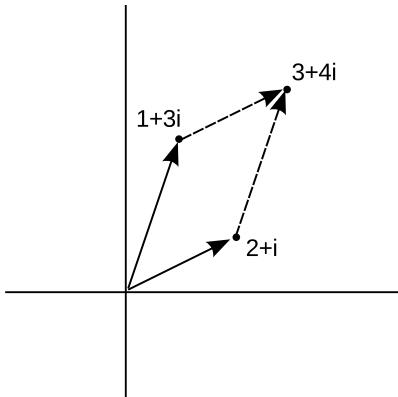
real numbers as a subset.

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

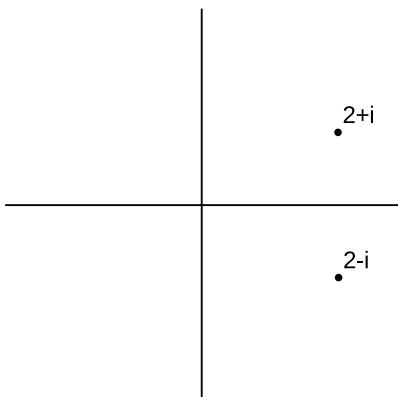
$$\begin{aligned} i^0 &= 1 && [\text{true for any base}] \\ i^1 &= i \\ i^2 &= -1 && [\text{definition of } i] \\ i^3 &= -i \\ i^4 &= 1. \end{aligned}$$



a / Visualizing complex numbers as points in a plane.



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



c / A complex number and its conjugate.

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.

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 1

▷ 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 1 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 A

Using this interpretation of multiplication, how could you find the square roots of a complex number? ▷ Answer, p. 432

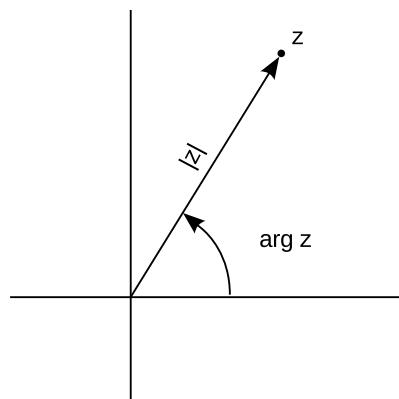
An identity

example 2

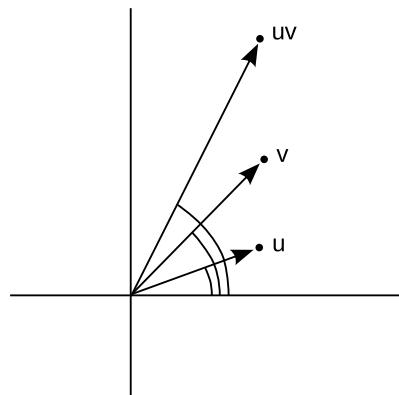
The magnitude $|z|$ of a complex number z obeys the identity $|z|^2 = z\bar{z}$. To prove this, we first note that \bar{z} has the same magnitude as z , since flipping it to the other side of the real axis doesn't change its distance from the origin. Multiplying z by \bar{z} gives a result whose magnitude 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

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

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.



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



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

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.

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

F From the results of question E, find the real and imaginary parts of $1/(1+i)$.

12.2 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 f.

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 3*

- ▷ 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 $\sin(-\theta) = -\sin \theta$.

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

A hard integral made easy

example 4

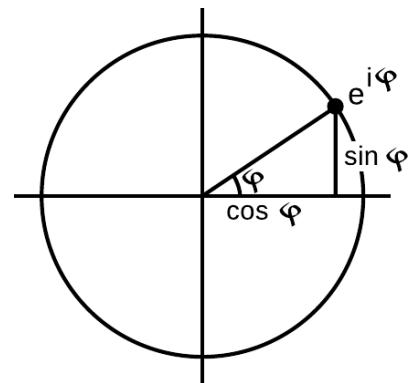
- ▷ Evaluate

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

- ▷ 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



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



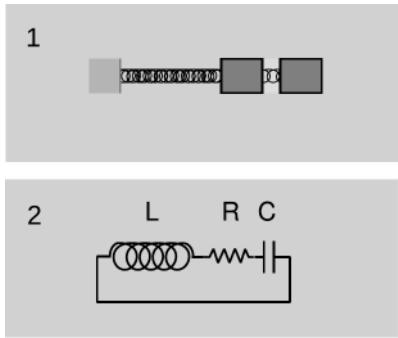
g / Leonhard Euler (1707-1783)

²In general, the use of complex number techniques to do an integral could

12.3 Simple harmonic motion

The simple harmonic oscillator should already be familiar to you. Here we show how complex numbers apply to the topic.

Figure h/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.



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

As a preview of ch. 13, figure h/2 shows the electrical analog of the mass on the spring. The resistor, marked R, is a familiar circuit element. The capacitor, C, is also familiar, and the symbol on the schematic is obviously meant to evoke a parallel-plate capacitor. The only unfamiliar circuit element is the one marked “L.” As suggested by the symbol on the schematic, think of this as a coil of wire, which generates a magnetic field. 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. 12.4.

The system of analogous variables is as follows:

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

Since this system of analogies is perfect, we’ll discuss the behavior of the more familiar 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 frequency³ is $\omega = \sqrt{k/m}$. It makes sense that there

result in a complex number, but that complex number would be a constant, which could be subsumed within the usual constant of integration.

³We use the word “frequency” to mean either f or $\omega = 2\pi f$ when the context

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 i 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 $c_1 \sin + c_2 \cos$ 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 $i\omega$

Sine compared to cosine

example 5

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 6

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

A function's first and second derivative

example 7

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 $1(3i)(3i) = -9$.

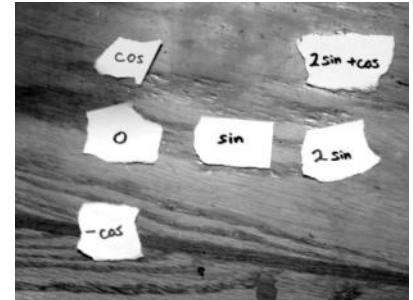
self-check B

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

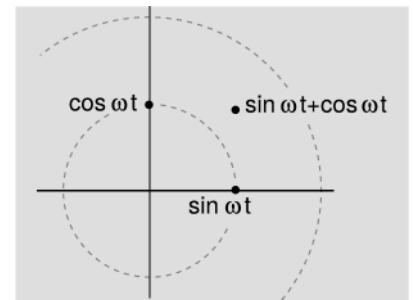
▷ Answer, p. 433

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

makes it clear which is being referred to.



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



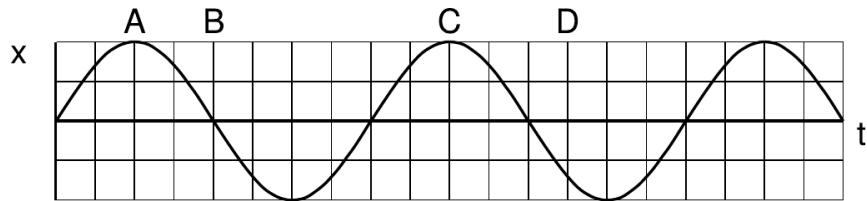
j / Example 6.

$k)A = 0$, 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.)

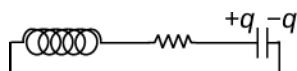
Discussion questions



A 1. The graph above shows the position as a function of time for a mass vibrating freely on a spring, with no driving force.

- How would we interpret the derivative of this function?
- List the forces.
- List the types of energy, and compare at times A, B, C, and D.

2. Suppose instead that this graph represents the charge q on one plate of the capacitor in the following circuit:



- Give a similar interpretation of the derivative and energy analysis.
- Given this graph, what can you infer about the resistance R ?

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

$$(\sin t)' = \cos t$$

$$(\cos t)' = -\sin t$$

$$(-\sin t)' = -\cos t$$

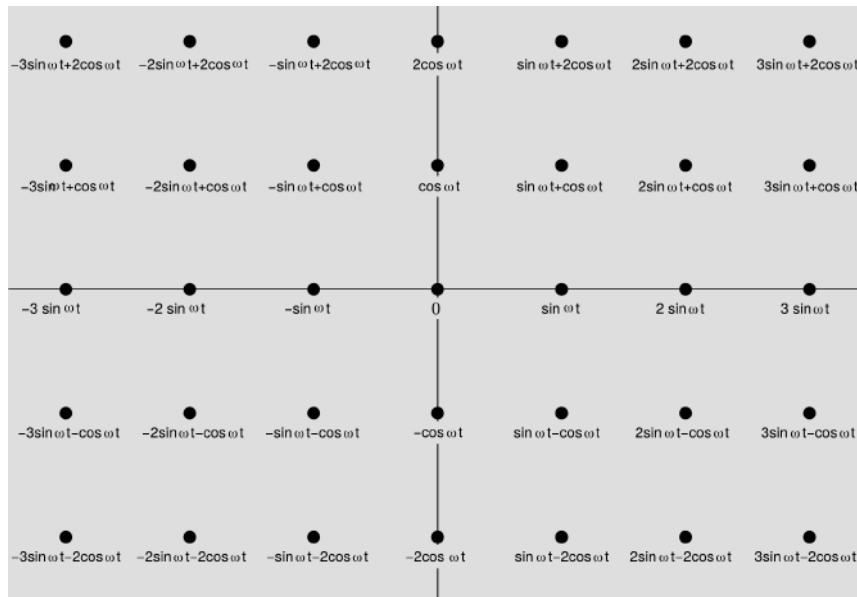
$$(-\cos t)' = \sin t$$

$$(2 \sin t)' = 2 \cos t$$

$$0' = 0$$

$$(\cos 2t)' = -2 \sin 2t$$

$$(\cos 3t)' = -3 \sin 3t$$



12.4 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 h 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 m shows an example.



m / The amplitude is halved with each cycle.

self-check C

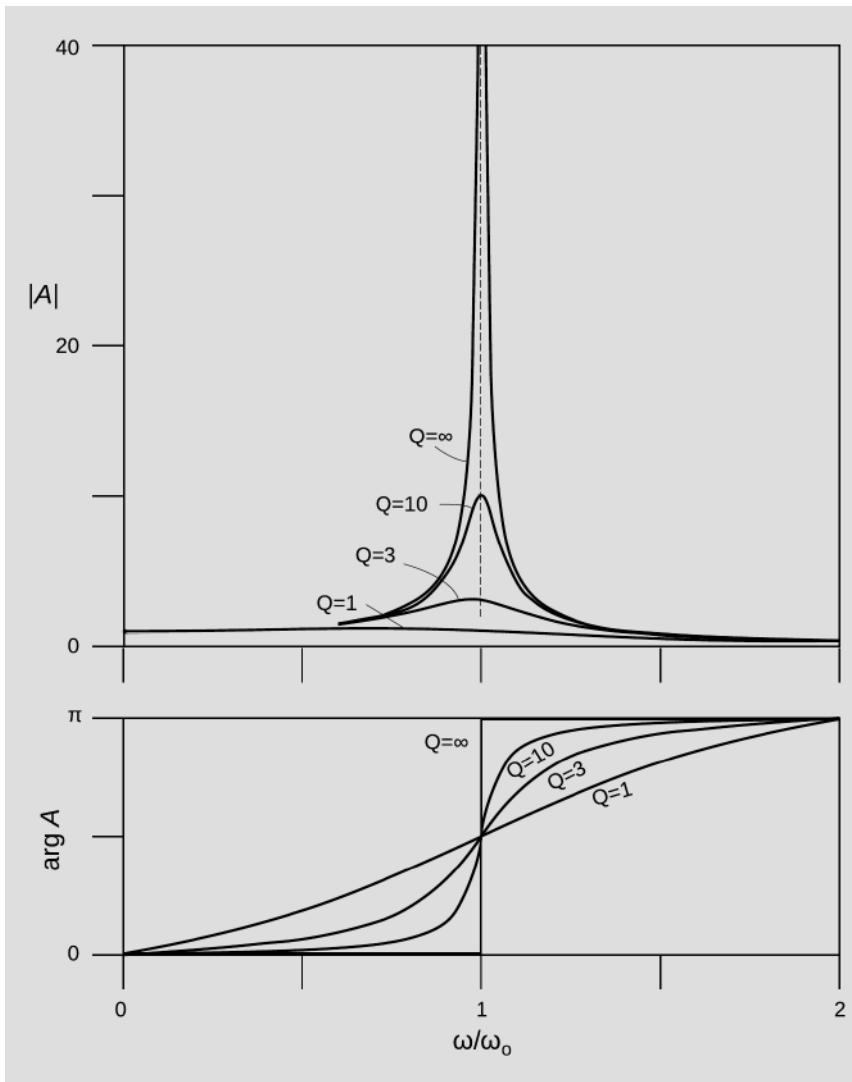
Figure m shows an x-t graph for a strongly damped vibration, which loses half of its amplitude with every cycle. What fraction of the energy is lost in each cycle?

▷ Answer, p. 433

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 $e^{2\pi} \approx 535$.

12.5 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.



n / 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.

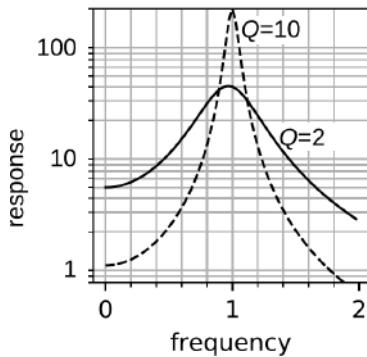
With the addition of a driving force F , the equation of motion for the damped oscillator becomes

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

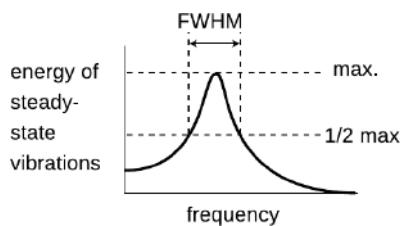
where F is a function of time. In terms of complex amplitudes, this is $(-\omega^2 m + i\omega b + k)A = \tilde{F}$. Here we introduce the notation \tilde{F} , which looks like a little sine wave above the F , to mean the complex number representing F 's amplitude. The result for the steady-state response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

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

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



o / 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 $Q = 10$.



p / Definition of the FWHM of the resonance peak.

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.

Figure n 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 o. The width of the resonance peak is often described using the full width at half-maximum, or FWHM, defined in figure p. The FWHM is approximately equal to $1/Q$ times the resonant frequency, the approximation being a good one when Q is large.

Dispersion of light in glass

example 8

A surprising and cool application is the explanation of why electromagnetic waves traveling through matter are *dispersive* (section 6.4), i.e., their speed depends on their frequency. Figure q/1 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

$$\text{original wave} = 1.$$

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 n. 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 is another distraction best left until after sec. 15.4. On the right side of the glass, we observe the superposition of the original wave and the secondary wave,

$$\text{transmitted wave} = 1 - iA.$$

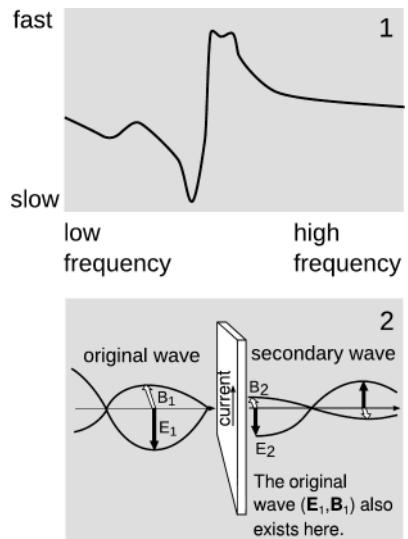
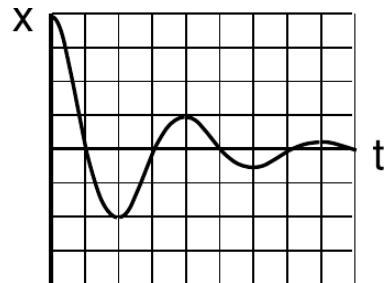
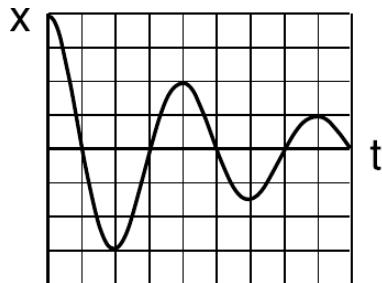
Consulting figure n, we see that for frequencies somewhat below the resonance, A is small and its phase approximately real-positive. Therefore $1 - iA$ is in the fourth quadrant, somewhat below the real axis. This represents a transmitted wave that is behind the original wave in terms of phase. The effect 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 .

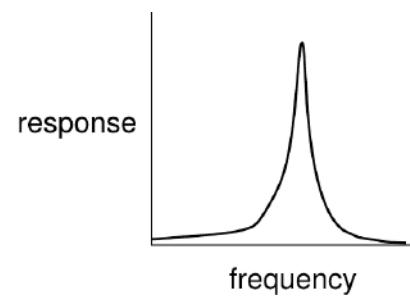
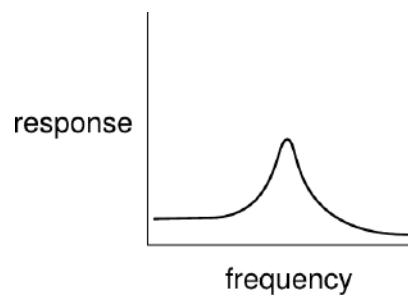
Discussion questions

- A** Compare the Q values of the two oscillators in the figures below.



q / Example 8. In the top panel, 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). The bottom panel shows 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>.

B Match the $x-t$ graphs in discussion question A with the amplitude-frequency graphs below.



Problems

Key

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

1 (a) Use complex number techniques to rewrite the function $f(t) = 4 \sin \omega t + 3 \cos \omega t$ in the form $A \sin(\omega t + \delta)$. ✓
(b) Verify the result using the trigonometric identity $\sin(\alpha + \beta) = \sin \alpha \cos \beta + \sin \beta \cos \alpha$.

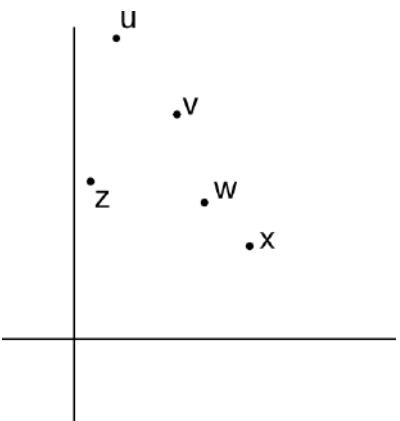
2 Use Euler's theorem to derive the addition theorems that express $\sin(a + b)$ and $\cos(a + b)$ in terms of the sines and cosines of a and b . ▷ Solution, p. 430

3 Find every complex number z such that $z^3 = 1$.
▷ Solution, p. 430

4 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

$$\frac{1+i}{(-2+2i)^{1/3}}.$$



Problem 4.

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.)

✓

5 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. 297 to explain why this is a solution to the differential equation.

▷ Solution, p. 431

6 Find the 100th derivative of $e^x \cos x$, evaluated at $x = 0$.
[Based on a problem by T. Needham.] ✓

7 Factor the expression $x^3 - y^3$ into factors of the lowest possible order, using complex coefficients. (Hint: use the result of problem 3.) Then do the same using real coefficients.

8 Calculate the quantity i^i (i.e., find its real and imaginary parts). ▷ Hint, p. 425 ✓ *

9 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? ✓

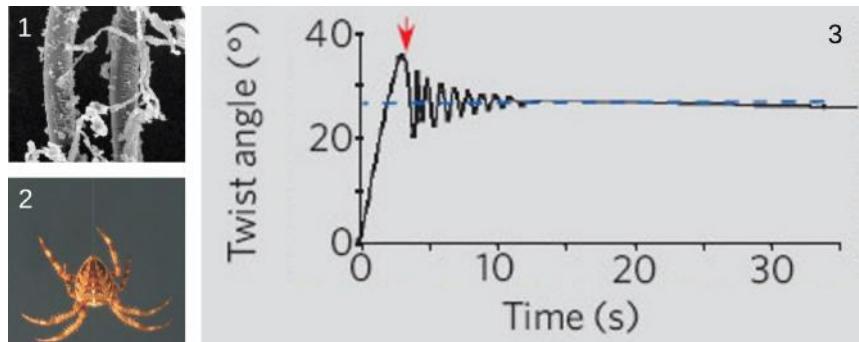
10 (a) We observe that the amplitude of a certain free oscillation decreases from A_0 to A_0/Z after n oscillations. Find its Q . ✓

(b) The figure is from *Shape memory in Spider draglines*, Emile, Le Floch, and Vollrath, *Nature* 440:621 (2006). Panel 1 shows an electron microscope's image of a thread of spider silk. In 2, a spider is hanging from such a thread. From an evolutionary point of view, it's probably a bad thing for the spider if it twists back and forth while hanging like this. (We're referring to a back-and-forth rotation about the axis of the thread, not a swinging motion like a pendulum.) The authors speculate that such a vibration could make the spider easier for predators to see, and it also seems to me that it would be a bad thing just because the spider wouldn't be able to control its orientation and do what it was trying to do. Panel 3 shows a graph of such an oscillation, which the authors measured using a video camera and a computer, with a 0.1 g mass hung from it in place of a spider. Compared to human-made fibers such as kevlar or copper wire, the spider thread has an unusual set of properties:

1. It has a low Q , so the vibrations damp out quickly.

- It doesn't become brittle with repeated twisting as a copper wire would.
- When twisted, it tends to settle in to a new equilibrium angle, rather than insisting on returning to its original angle. You can see this in panel 2, because although the experimenters initially twisted the wire by 35 degrees, the thread only performed oscillations with an amplitude much smaller than ± 35 degrees, settling down to a new equilibrium at 27 degrees.
- Over much longer time scales (hours), the thread eventually resets itself to its original equilibrium angle (shown as zero degrees on the graph). (The graph reproduced here only shows the motion over a much shorter time scale.) Some human-made materials have this "memory" property as well, but they typically need to be heated in order to make them go back to their original shapes.

Focusing on property number 1, estimate the Q of spider silk from the graph. ✓



Problem 10.

- 11** (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. ✓

- 12** Simplify $\arg(1/\bar{z})$.

Chapter 13

AC circuits

13.1 Capacitance and inductance

In a mechanical oscillation, energy is exchanged repetitively between potential and kinetic forms, and may also be siphoned off in the form of heat dissipated by friction. In an electrical circuit, resistors are the circuit elements that dissipate heat. In a circuit like a radio receiver, what are the electrical analogs of storing and releasing the potential and kinetic energy of a vibrating object? When you think of energy storage in an electrical circuit, you are likely to imagine a battery, but even rechargeable batteries can only go through 10 or 100 cycles before they wear out. In addition, batteries are not able to exchange energy on a short enough time scale for most applications. The circuit in a musical synthesizer may be called upon to oscillate thousands of times a second, and your microwave oven operates at gigahertz frequencies. Instead of batteries, we generally use capacitors and inductors to store energy in oscillating circuits. Capacitors, which you've already encountered, store energy in electric fields. An inductor does the same with magnetic fields.

13.1.1 Capacitors

A capacitor's energy exists in the electric fields near its electrodes. The energy is proportional to the square of the field strength, which is proportional to the charges on the plates. If we assume the plates carry charges that are the same in magnitude, $+q$ and $-q$, then the energy stored in the capacitor must be proportional to q^2 . For historical reasons, we write the constant of proportionality as $1/2C$,

$$U_C = \frac{1}{2C}q^2.$$

The constant C is a geometrical property of the capacitor, called its capacitance.

Based on this definition, the units of capacitance must be coulombs squared per joule, and this combination is more conveniently abbreviated as the farad, $1 \text{ F} = 1 \text{ C}^2/\text{J}$. “Condenser” is a less formal term for a capacitor. Note that the labels printed on capacitors often use MF to mean μF , even though MF should really be the symbol for megafarads, not microfarads. Confusion doesn't result from this nonstandard notation, since picofarad and microfarad values are the most common, and it wasn't until the 1990's that even



a / The symbol for a capacitor.



b / Some capacitors.

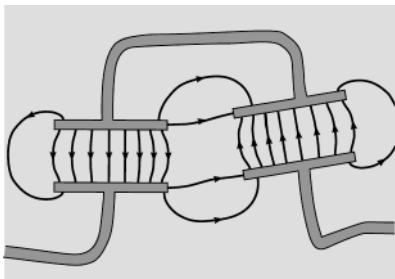
millifarad and farad values became available in practical physical sizes. Figure a shows the symbol used in schematics to represent a capacitor.

A parallel-plate capacitor

example 1

▷ Suppose a capacitor consists of two parallel metal plates with area A , and the gap between them is h . The gap is small compared to the dimensions of the plates. What is the capacitance?

▷ Using the result of problem 11, p. 69, $E = 4\pi k\sigma = 4\pi kq/A$. The energy stored in the field is $U_e = (1/8\pi k)E^2 Ah$. Substituting the first expression into the second, we find $U_e = 2\pi kq^2 h/A$. Comparing this to the definition of capacitance, we end up with $C = A/4\pi kh$.



c / A configuration of capacitors for which the lumped circuit approximation might be bad.

If you look at the printed circuit board in a typical piece of consumer electronics, there are many capacitors, often placed fairly close together. Life is made a lot simpler by the fact that each capacitor's field tends to be concentrated on its own interior. If their exterior fields are not so small, then each capacitor interacts with its neighbors in a complicated way, and the behavior of the circuit depends on the exact physical layout, since the interaction would be stronger or weaker depending on distance. The resulting behavior of the circuit then depends in detail on the three-dimensional geometry of the circuit, as in figure c. In reality, a capacitor does create weak external electric fields, but their effects are often negligible, and we can then use the *lumped-circuit approximation*, which states that each component's behavior depends only on the currents that flow in and out of it, not on the interaction of its fields with the other components.

13.1.2 Inductors

Any current will create a magnetic field, so in fact every current-carrying wire in a circuit acts as an inductor! However, this type of “stray” inductance is typically negligible, just as we can usually ignore the stray resistance of our wires and only take into account the actual resistors. To store any appreciable amount of magnetic energy, one usually uses a coil of wire designed specifically to be an inductor. All the loops' contribution to the magnetic field add together to make a stronger field. Unlike capacitors and resistors, practical inductors are easy to make by hand. One can for instance spool some wire around a short wooden dowel. An inductor like this, in the form cylindrical coil of wire, is called a solenoid, d, and a stylized solenoid, e, is the symbol used to represent an inductor in a circuit regardless of its actual geometry.

How much energy does an inductor store? The energy density is proportional to the square of the magnetic field strength, which is in turn proportional to the current flowing through the coiled wire, so the energy stored in the inductor must be proportional to I^2 . We

write $L/2$ for the constant of proportionality, giving

$$U_L = \frac{L}{2} I^2.$$

As in the definition of capacitance, we have a factor of $1/2$, which is purely a matter of definition. The quantity L is called the *inductance* of the inductor, and we see that its units must be joules per ampere squared. This clumsy combination of units is more commonly abbreviated as the henry, $1 \text{ H} = 1 \text{ J/A}^2$. Rather than memorizing this definition, it makes more sense to derive it when needed from the definition of inductance.

Many people know inductors simply as “coils,” or “chokes,” and will not understand you if you refer to an “inductor,” but they will still refer to L as the “inductance,” not the “coilance” or “chokeance!” The term “choke” is derived from the fact that, as we’ll see in sec. 14.1, an inductance tends to pass low frequencies while “choking off” high frequencies. This is the opposite of a capacitor: a capacitor has vacuum or an insulator between the plates, and therefore will not pass DC. Therefore capacitors and inductors are often used as filters to get rid of certain frequencies in a signal (example 1, p. 332). If you’re listening to music on a pair of stereo speakers while reading this, then there is probably a set of filters, called the cross-over filter, that send low frequencies to the woofers and highs to the tweeters.

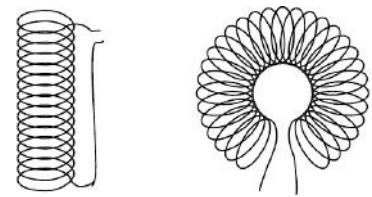
The type of air-core coil described above is a good one for many applications, such as a stereo speaker’s cross-over filter. For other applications, however, this technology would be too expensive, too bulky to use in miniaturized electronics, or too hard to integrate into a printed-circuit board. For some applications, this can be remedied by adding an iron core, while for others the behavior of an inductor can be achieved using a combination of transistors (an op-amp) and a capacitor. When we consider the behavior of these things as circuit elements, the same equations apply regardless of the implementation.

There is a lumped circuit approximation for inductors, just like the one for capacitors (p. 310). For a capacitor, this means assuming that the electric fields are completely internal, so that components only interact via currents that flow through wires, not due to the physical overlapping of their fields in space. Similarly for an inductor, the lumped circuit approximation is the assumption that the magnetic fields are completely internal.

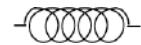
Inductance of a solenoid

example 2

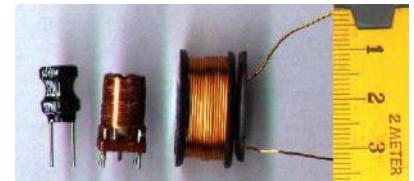
The magnetic field of an ideal solenoid is $B = 4\pi k c^{-2} I N / \ell$ (p. 273). By combining this with the definition of inductance $U = LI^2/2$ and the expression $(c^2/8\pi k)B^2$ for the energy density, we find (problem 10, p. 327) that the inductance of a solenoid is $L = (4\pi k/c^2)N^2 A / \ell$, where A is the cross-sectional area.



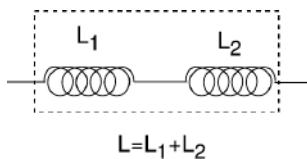
d / Two common geometries for inductors. The cylindrical shape on the left is called a solenoid.



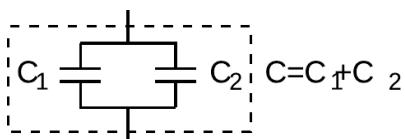
e / The symbol for an inductor.



f / Some inductors.



g / Inductances in series add.



h / Capacitances in parallel add.

Identical inductances in series

example 3

If two inductors are placed in series, any current that passes through the combined double inductor must pass through both its parts. If we assume the lumped circuit approximation, the two inductors' fields don't interfere with each other, so the energy is doubled for a given current. Thus by the definition of inductance, the inductance is doubled as well. In general, inductances in series add, just like resistances. The same kind of reasoning also shows that the inductance of a solenoid is approximately proportional to its length, assuming the number of turns per unit length is kept constant. (This is only approximately true, because putting two solenoids end-to-end causes the fields just outside their mouths to overlap and add together in a complicated manner. In other words, the lumped-circuit approximation may not be very good.)

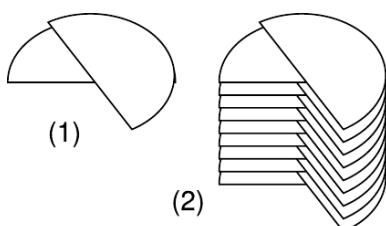
Identical capacitances in parallel

example 4

When two identical capacitances are placed in parallel, any charge deposited at the terminals of the combined double capacitor will divide itself evenly between the two parts. The electric fields surrounding each capacitor will be half the intensity, and therefore store one quarter the energy. Two capacitors, each storing one quarter the energy, give half the total energy storage. Since capacitance is inversely related to energy storage, this implies that identical capacitances in parallel give double the capacitance. In general, capacitances in parallel add. This is unlike the behavior of inductors and resistors, for which series configurations give addition.

This is consistent with the result of example 1, which had the capacitance of a single parallel-plate capacitor proportional to the area of the plates. If we have two parallel-plate capacitors, and we combine them in parallel and bring them very close together side by side, we have produced a single capacitor with plates of double the area, and it has approximately double the capacitance, subject to any violation of the lumped-circuit approximation due to the interaction of the fields where the edges of the capacitors are joined together.

Inductances in parallel and capacitances in series are explored in sec. 14.5, p. 340.



i / A variable capacitor.

A variable capacitor

example 5

Figure i/1 shows the construction of a variable capacitor out of two parallel semicircles of metal. One plate is fixed, while the other can be rotated about their common axis with a knob. The opposite charges on the two plates are attracted to one another, and therefore tend to gather in the overlapping area. This overlapping area, then, is the only area that effectively contributes to the capacitance, and turning the knob changes the capacitance.

The simple design can only provide very small capacitance values, so in practice one usually uses a bank of capacitors, wired in parallel, with all the moving parts on the same shaft.

Discussion questions

A Suppose that two parallel-plate capacitors are wired in parallel, and are placed very close together, side by side, so that the lumped circuit approximation is not very accurate. Will the resulting capacitance be too small, or too big? Could you twist the circuit into a different shape and make the effect be the other way around, or make the effect vanish? How about the case of two inductors in series?

B Most practical capacitors do not have an air gap or vacuum gap between the plates; instead, they have an insulating substance called a dielectric. We can think of the molecules in this substance as dipoles that are free to rotate (at least a little), but that are not free to move around, since it is a solid. The figure shows a highly stylized and unrealistic way of visualizing this. We imagine that all the dipoles are initially turned sideways, (1), and that as the capacitor is charged, they all respond by turning through a certain angle, (2). (In reality, the scene might be much more random, and the alignment effect much weaker.)

For simplicity, imagine inserting just one electric dipole into the vacuum gap. For a given amount of charge on the plates, how does this affect the amount of energy stored in the electric field? How does this affect the capacitance?

Now redo the analysis in terms of the mechanical work needed in order to charge up the plates.

13.2 Oscillations

Figure k shows the simplest possible oscillating circuit. For any useful application it would actually need to include more components. For example, if it was a radio tuner, it would need to be connected to an antenna and an amplifier. Nevertheless, all the essential physics is there.

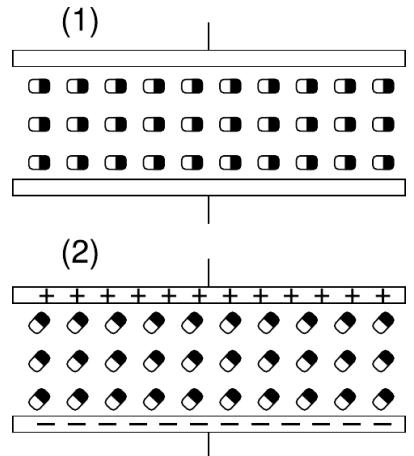
We can analyze it without any sweat or tears whatsoever, simply by filling in the rest of the system of analogies introduced in sections 12.3-12.4, pp. 296-298, with a mechanical oscillator. In figure l, we have two forms of stored energy,

$$U_{spring} = \frac{1}{2}kx^2 \quad (1)$$

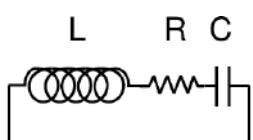
$$K = \frac{1}{2}mv^2. \quad (2)$$

In the circuit, the dissipation of energy into heat occurs via the resistor, with no mechanical force involved, so in order to make the analogy, we need to restate the role of the friction force in terms of energy. The power dissipated by friction equals the mechanical work it does in a time interval dt , divided by dt , $P = W/dt = F dx/dt = Fv = -bv^2$, so

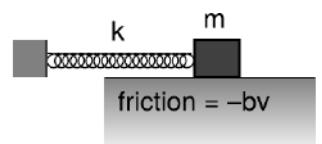
$$\text{rate of heat dissipation} = -bv^2. \quad (3)$$



j / Discussion question B.



k / A series LRC circuit.



l / A mechanical analogy for the LRC circuit.

self-check A

Equation (1) has x squared, and equations (2) and (3) have v squared. Because they're squared, the results don't depend on whether these variables are positive or negative. Does this make physical sense? ▷ Answer, p. 433

In the circuit, the stored forms of energy are

$$U_C = \frac{1}{2C}q^2 \quad (1')$$

$$U_L = \frac{1}{2}LI^2, \quad (2')$$

and the rate of heat dissipation in the resistor is

$$\text{rate of heat dissipation} = -RI^2. \quad (3')$$

We've previously discussed the analogies $x \leftrightarrow q$ and $v \leftrightarrow I$. Next we relate the variables that describe the system's permanent characteristics:

$$k \leftrightarrow 1/C$$

$$m \leftrightarrow L$$

$$b \leftrightarrow R$$

Since the mechanical system naturally oscillates with a frequency $\omega \approx \sqrt{k/m}$, we can immediately solve the electrical version by analogy, giving

$$\omega \approx \frac{1}{\sqrt{LC}}.$$

Since the resistance R is analogous to the friction parameter b in the mechanical case, we find that the Q (quality factor, not charge) of the resonance is inversely proportional to R , and the width of the resonance is directly proportional to R .

Tuning a radio receiver

example 6

A radio receiver uses this kind of circuit to pick out the desired station. Since the receiver resonates at a particular frequency, stations whose frequencies are far off will not excite any response in the circuit. The value of R has to be small enough so that only one station at a time is picked up, but big enough so that the tuner isn't too touchy. The resonant frequency can be tuned by adjusting either L or C , but variable capacitors are easier to build than variable inductors.

A numerical calculation

example 7

The phone company sends more than one conversation at a time over the same wire, which is accomplished by shifting each voice signal into different range of frequencies during transmission. The number of signals per wire can be maximized by making each

range of frequencies (known as a bandwidth) as small as possible. It turns out that only a relatively narrow range of frequencies is necessary in order to make a human voice intelligible, so the phone company filters out all the extreme highs and lows. (This is why your phone voice sounds different from your normal voice.)

▷ If the filter consists of an LRC circuit with a broad resonance centered around 1.0 kHz, and the capacitor is 1 μF (microfarad), what inductance value must be used?

▷ Solving for L , we have

$$\begin{aligned} L &= \frac{1}{C\omega^2} \\ &= \frac{1}{(10^{-6} \text{ F})(2\pi \times 10^3 \text{ s}^{-1})^2} \\ &= 2.5 \times 10^{-3} \text{ F}^{-1}\text{s}^2 \end{aligned}$$

Checking that these really are the same units as henries is a little tedious, but it builds character:

$$\begin{aligned} \text{F}^{-1}\text{s}^2 &= (\text{C}^2/\text{J})^{-1}\text{s}^2 \\ &= \text{J} \cdot \text{C}^{-2}\text{s}^2 \\ &= \text{J}/\text{A}^2 \\ &= \text{H} \end{aligned}$$

The result is 25 mH (millihenries). (An inductance value this large would probably be implemented with an op-amp, not a coil.)

13.3 Voltage and current

What is physically happening in one of these oscillating circuits? Let's first look at the mechanical case, and then draw the analogy to the circuit. For simplicity, let's ignore the existence of damping, so there is no friction in the mechanical oscillator, and no resistance in the electrical one.

Suppose we take the mechanical oscillator and pull the mass away from equilibrium, then release it. Since friction tends to resist the spring's force, we might naively expect that having zero friction would allow the mass to leap instantaneously to the equilibrium position. This can't happen, however, because the mass would have to have infinite velocity in order to make such an instantaneous leap. Infinite velocity would require infinite kinetic energy, but the only kind of energy that is available for conversion to kinetic is the energy stored in the spring, and that is finite, not infinite. At each step on its way back to equilibrium, the mass's velocity is controlled exactly by the amount of the spring's energy that has so far been converted into kinetic energy. After the mass reaches equilibrium, it overshoots due to its own momentum. It performs identical oscillations on both

sides of equilibrium, and it never loses amplitude because friction is not available to convert mechanical energy into heat.

Now with the electrical oscillator, the analog of position is charge. Pulling the mass away from equilibrium is like depositing charges $+q$ and $-q$ on the plates of the capacitor. Since resistance tends to resist the flow of charge, we might imagine that with no friction present, the charge would instantly flow through the inductor (which is, after all, just a piece of wire), and the capacitor would discharge instantly. However, such an instant discharge is impossible, because it would require infinite current for one instant. Infinite current would create infinite magnetic fields surrounding the inductor, and these fields would have infinite energy. Instead, the rate of flow of current is controlled at each instant by the relationship between the amount of energy stored in the magnetic field and the amount of current that must exist in order to have that strong a field. After the capacitor reaches $q = 0$, it overshoots. The circuit has its own kind of electrical “inertia,” because if charge was to stop flowing, there would have to be zero current through the inductor. But the current in the inductor must be related to the amount of energy stored in its magnetic fields. When the capacitor is at $q = 0$, all the circuit’s energy is in the inductor, so it must therefore have strong magnetic fields surrounding it and quite a bit of current going through it.

The only thing that might seem spooky here is that we used to speak as if the current in the inductor caused the magnetic field, but now it sounds as if the field causes the current. Actually this is symptomatic of the elusive nature of cause and effect in physics. It’s equally valid to think of the cause and effect relationship in either way. This may seem unsatisfying, however, and for example does not really get at the question of what brings about a voltage difference across the resistor (in the case where the resistance is finite); there must be such a voltage difference, because without one, Ohm’s law would predict zero current through the resistor.

Voltage, then, is what is really missing from our story so far.

Let’s start by studying the voltage across a capacitor. Voltage is electrical potential energy per unit charge, so the voltage difference between the two plates of the capacitor is related to the amount by which its energy would increase if we increased the absolute values of the charges on the plates from q to $q + dq$:

$$\begin{aligned} V_C &= (U_{q+dq} - U_q) / dq \\ &= \frac{dU_C}{dq} \\ &= \frac{d}{dq} \left(\frac{1}{2C} q^2 \right) \\ &= \frac{q}{C} \end{aligned}$$

Many books use this as the definition of capacitance. This equation, by the way, probably explains the historical reason why C was defined so that the energy was *inversely* proportional to C for a given value of q : the people who invented the definition were thinking of a capacitor as a device for storing charge rather than energy, and the amount of charge stored for a fixed voltage (the charge “capacity”) is proportional to C .

In the case of an inductor, we know that if there is a steady, constant current flowing through it, then the magnetic field is constant, and so is the amount of energy stored; no energy is being exchanged between the inductor and any other circuit element. But what if the current is changing? The magnetic field is proportional to the current, so a change in one implies a change in the other. For concreteness, let’s imagine that the magnetic field and the current are both decreasing. The energy stored in the magnetic field is therefore decreasing, and by conservation of energy, this energy can’t just go away — some other circuit element must be taking energy from the inductor. The simplest example, shown in figure m, is a series circuit consisting of the inductor plus one other circuit element. It doesn’t matter what this other circuit element is, so we just call it a black box, but if you like, we can think of it as a resistor, in which case the energy lost by the inductor is being turned into heat by the resistor. The junction rule tells us that both circuit elements have the same current through them, so I could refer to either one.

Experience with the loop rule would also lead us to expect that $V_{\text{inductor}} + V_{\text{black box}} = 0$, so the two voltage drops have the same absolute value, which we could then notate simply as V . In real-world circuits, this is indeed what we would usually find, to a good approximation, using an AC voltmeter, although the loop rule can actually be violated in an AC circuit ([Z324](#)).

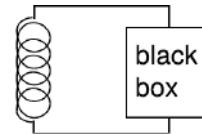
Whatever the black box is, the rate at which it is taking energy from the inductor is given by $P = IV$, so (ignoring signs)

$$\begin{aligned} IV &= \frac{dU_L}{dt} \\ &= \frac{d}{dt} \left(\frac{1}{2} LI^2 \right) \\ &= LI \frac{dI}{dt}, \end{aligned}$$

or

$$V = L \frac{dI}{dt}, \quad [\text{See below re sign.}]$$

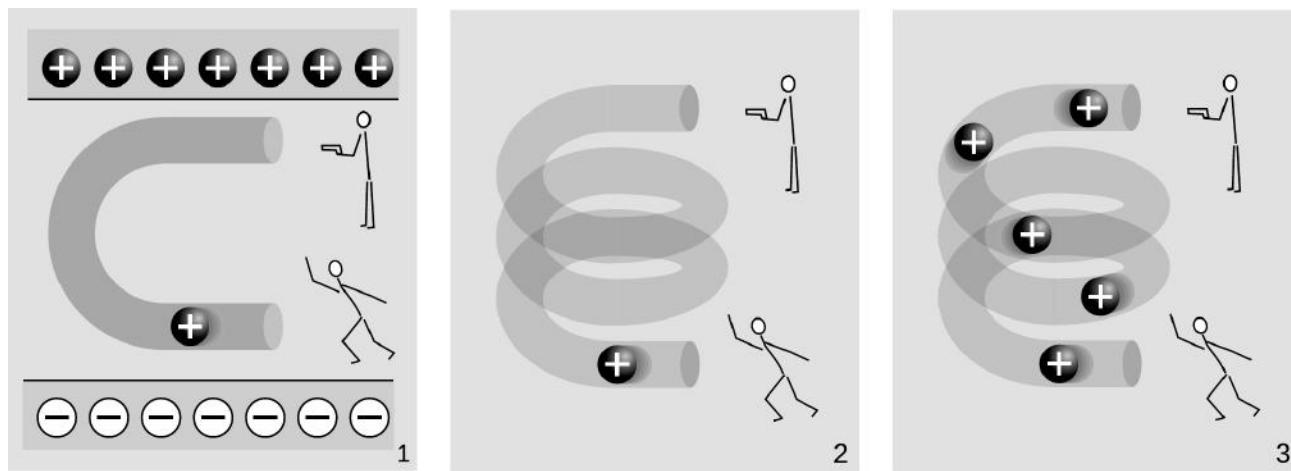
which in many books is taken to be the definition of inductance. When applying this in a circuit, we insert a sign in front that indicates that the inductor resists the change in current, i.e., that provides negative feedback. The choice of sign that accomplishes



m / The inductor releases energy and gives it to the black box.

this is partly a matter of arbitrary choice, since the real-world measurements that ultimately define V and I depend on how we hook up the voltmeter and ammeter.

To someone who knew only about DC circuits, $V = L dI/dt$ would be a surprising result. Suppose, for concreteness, that the black box in figure m is a resistor, and that the inductor's energy is decreasing, and being converted into heat in the resistor. The voltage drop across the resistor indicates that it has an electric field across it, which is driving the current. But where is this electric field coming from? There are no charges anywhere that could be creating it! The answer, of course, is that this is an example of electromagnetic induction. One of Maxwell's equations (sec. 6.7, p. 163) is that $\text{curl } \mathbf{E} = -\partial \mathbf{B}/\partial t$, which tells us that the change in the magnetic field will cause a curly electric field. If we hadn't already known about induction, this example would have forced us to invent it.



n / Electric fields made by charges, 1, and by changing magnetic fields, 2 and 3.

The cartoons in figure n compares electric fields made by charges, 1, to electric fields made by changing magnetic fields, 2-3. In n/1, two physicists are in a room whose ceiling is positively charged and whose floor is negatively charged. The physicist on the bottom throws a positively charged bowling ball into the curved pipe. The physicist at the top uses a radar gun to measure the speed of the ball as it comes out of the pipe. They find that the ball has slowed down by the time it gets to the top. By measuring the change in the ball's kinetic energy, the two physicists are acting just like a voltmeter. They conclude that the top of the tube is at a higher voltage than the bottom of the pipe. A difference in voltage indicates an electric field, and this field is clearly being caused by the charges in the floor and ceiling.

In $n/2$, there are no charges anywhere in the room except for the charged bowling ball. Moving charges make magnetic fields, so there is a magnetic field surrounding the helical pipe while the ball is moving through it. A magnetic field has been created where there was none before, and that field has energy. Where could the energy have come from? It can only have come from the ball itself, so the ball must be losing kinetic energy. The two physicists working together are again acting as a voltmeter, and again they conclude that there is a voltage difference between the top and bottom of the pipe. This indicates an electric field, but this electric field can't have been created by any charges, because there aren't any in the room. This electric field was created by the change in the magnetic field.

The bottom physicist keeps on throwing balls into the pipe, until the pipe is full of balls, $n/3$, and finally a steady current is established. While the pipe was filling up with balls, the energy in the magnetic field was steadily increasing, and that energy was being stolen from the balls' kinetic energy. But once a steady current is established, the energy in the magnetic field is no longer changing. The balls no longer have to give up energy in order to build up the field, and the physicist at the top finds that the balls are exiting the pipe at full speed again. There is no voltage difference any more. Although there is a current, dI/dt is zero.

Ballasts

example 8

In a gas discharge tube, such as a neon sign, enough voltage is applied to a tube full of gas to ionize some of the atoms in the gas. Once ions have been created, the voltage accelerates them, and they strike other atoms, ionizing them as well and resulting in a chain reaction. This is a spark, like a bolt of lightning. But once the spark starts up, the device begins to act as though it has no resistance: more and more current flows, without the need to apply any more voltage. The power, $P = IV$, would grow without limit, and the tube would burn itself out.

The simplest solution is to connect an inductor, known as the "ballast," in series with the tube, and run the whole thing on an AC voltage. During each cycle, as the voltage reaches the point where the chain reaction begins, there is a surge of current, but the inductor resists such a sudden change of current, and the energy that would otherwise have burned out the bulb is instead channeled into building a magnetic field.

A common household fluorescent lightbulb consists of a gas discharge tube in which the glass is coated with a fluorescent material. The gas in the tube emits ultraviolet light, which is absorbed by the coating, and the coating then glows in the visible spectrum.

Until recently, it was common for a fluorescent light's ballast to be



o / Ballasts for fluorescent lights. Top: a big, heavy inductor used as a ballast in an old-fashioned fluorescent bulb. Bottom: a small solid-state ballast, built into the base of a modern compact fluorescent bulb.

a simple inductor, and for the whole device to be operated at the 60 Hz frequency of the electrical power lines. This caused the lights to flicker annoyingly at 120 Hz, and could also cause an audible hum, since the magnetic field surrounding the inductor could exert mechanical forces on things. Modern compact fluorescent bulbs have ballasts built into their bases that use a frequency in the kilohertz range, eliminating the flicker and hum.

In section 13.2 we analyzed the oscillations of a series LRC circuit simply by appealing to the mechanical analog, which we had already solved, but it's also interesting to write down the actual differential equation for the circuit. The loop rule is approximately valid under the lumped circuit approximation ([2324](#)), so

$$V_L + V_R + V_C = 0.$$

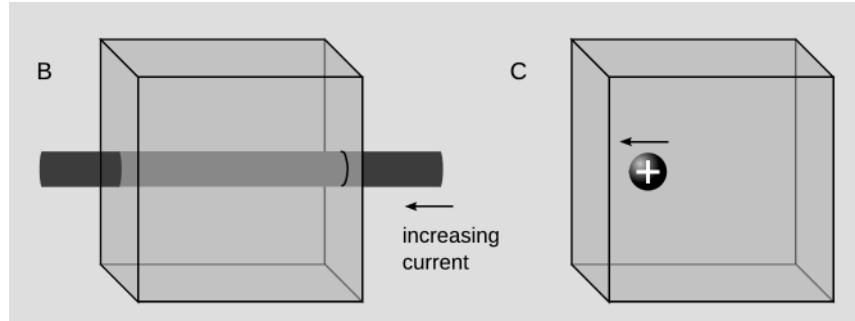
We now want to express this in terms of q and its derivatives, so we use $V_C = q/C$, $V_R = IR = Rq'$, and $|V_L| = L dI/dt = Lq''$, resulting in

$$Lq'' + Rq' + \frac{1}{C}q = 0.$$

If we look for solutions of the form $q = e^{st}$, we find that $Ls^2 + Rs + 1/C = 0$. When $R = 0$, the solutions are $s = \pm i/\sqrt{LC}$, or $\omega = \mp 1/\sqrt{LC}$, as expected.

Discussion questions

A What happens when the physicist at the bottom in figure n/3 starts getting tired, and decreases the current?



Discussion questions B and C.

B The wire passes in through one side of the cube and out through the other. If the current through the wire is increasing, then the wire will act like an inductor, and there will be a voltage difference between its ends. (The inductance will be relatively small, since the wire isn't coiled up, and the ΔV will therefore also be fairly small, but still not zero.) The ΔV implies the existence of electric fields, and yet Gauss's law says the flux must be zero, since there is no charge inside the cube. Why isn't Gauss's law violated?

C The charge has been loitering near the edge of the cube, but is then suddenly hit with a mallet, causing it to fly off toward the left side of the cube. Disturbances in the electric and magnetic fields ripple outward through space at the speed of light. Because the charge is closer to the left side of the cube, the change in the electric field occurs there before the information reaches the right side. This would seem certain to lead to a violation of Gauss's law. How can the ideas explored in discussion question B show the resolution to this paradox?

13.4 Decay

Up until now I've avoided talking about the fact that by changing the characteristics of an oscillator, it is possible to produce non-oscillatory behavior. For example, imagine taking the mass-on-a-spring system and making the spring weaker and weaker. In the limit of small k , it's as though there was no spring whatsoever, and the behavior of the system is that if you kick the mass, it simply starts slowing down. Although the limit $k \rightarrow 0$ is intuitively easy to imagine in the mechanical case, we get the non-oscillatory behavior whenever $Q < 1/2$ ([2324](#)). In sections 13.4.1 and 13.4.2 we will analyze only the cases where either the capacitor or the inductor is completely absent, giving $Q = 0$.

13.4.1 The RC circuit

We first analyze the RC circuit, p. In reality one would have to "kick" the circuit, for example by briefly inserting a battery, in order to get any interesting behavior. Setting $L = 0$ in our analysis from p. 320, we have

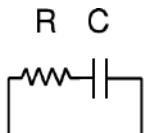
$$I = -\frac{1}{RC}q$$

The negative sign tells us that the current tends to reduce the charge on the capacitor, i.e., to discharge it. It makes sense that the current is proportional to q : if q is large, then the attractive forces between the $+q$ and $-q$ charges on the plates of the capacitor are large, and charges will flow more quickly through the resistor in order to reunite. If there was zero charge on the capacitor plates, there would be no reason for current to flow. Since amperes, the unit of current, are the same as coulombs per second, it appears that the quantity RC must have units of seconds, and you can check for yourself that this is correct. RC is therefore referred to as the time constant of the circuit.

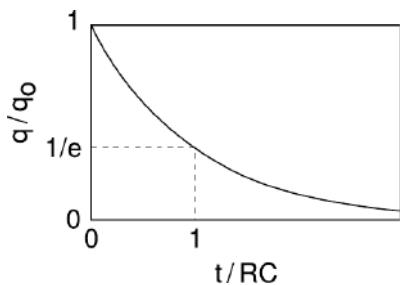
How exactly do I and q vary with time? Rewriting I as dq/dt , we have

$$\frac{dq}{dt} = -\frac{1}{RC}q.$$

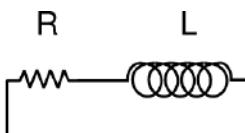
We need a function $q(t)$ whose derivative equals itself, but multiplied by a negative constant. A function of the form ae^{bt} , where $e = 2.718\dots$ is the base of natural logarithms, is the only one that has its derivative equal to itself, and ae^{bt} has its derivative equal to itself



p / An RC circuit.



q / Over a time interval RC , the charge on the capacitor is reduced by a factor of e .



r / An RL circuit.

multiplied by b . Thus our solution is

$$q = q_0 \exp\left(-\frac{t}{RC}\right).$$

13.4.2 The RL circuit

The RL circuit, r, can be attacked by similar methods, and it can easily be shown that it gives

$$I = I_0 \exp\left(-\frac{R}{L}t\right).$$

The RL time constant equals L/R .

Death by solenoid; spark plugs

example 9

When we suddenly break an RL circuit, what will happen? It might seem that we're faced with a paradox, since we only have two forms of energy, magnetic energy and heat, and if the current stops suddenly, the magnetic field must collapse suddenly. But where does the lost magnetic energy go? It can't go into resistive heating of the resistor, because the circuit has now been broken, and current can't flow!

The way out of this conundrum is to recognize that the open gap in the circuit has a resistance which is large, but not infinite. This large resistance causes the RL time constant L/R to be very small. The current thus continues to flow for a very brief time, and flows straight across the air gap where the circuit has been opened. In other words, there is a spark!

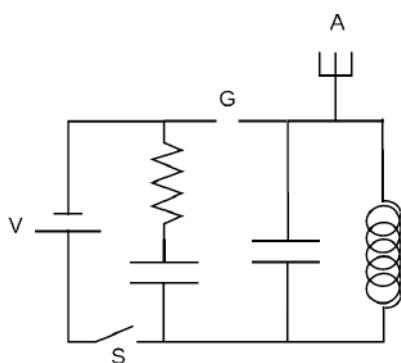
We can determine based on several different lines of reasoning that the voltage drop from one end of the spark to the other must be very large. First, the air's resistance is large, so $V = IR$ requires a large voltage. We can also reason that all the energy in the magnetic field is being dissipated in a short time, so the power dissipated in the spark, $P = IV$, is large, and this requires a large value of V . (I isn't large — it is decreasing from its initial value.) Yet a third way to reach the same result is to consider the equation $V_L = dI/dt$: since the time constant is short, the time derivative dI/dt is large.

This is exactly how a car's spark plugs work. Another application is to electrical safety: it can be dangerous to break an inductive circuit suddenly, because so much energy is released in a short time. There is also no guarantee that the spark will discharge across the air gap; it might go through your body instead, since your body might have a lower resistance.

A spark-gap radio transmitter

example 10

Figure s shows a primitive type of radio transmitter, called a spark gap transmitter, used to send Morse code around the turn of the



s / Example 10.

twentieth century. The high voltage source, V , is typically about 10,000 volts. When the telegraph switch, S , is closed, the RC circuit on the left starts charging up. An increasing voltage difference develops between the electrodes of the spark gap, G . When this voltage difference gets large enough, the electric field in the air between the electrodes causes a spark, partially discharging the RC circuit, but charging the LC circuit on the right. The LC circuit then oscillates at its resonant frequency (typically about 1 MHz), but the energy of these oscillations is rapidly radiated away by the antenna, A , which sends out radio waves.

Discussion questions

- A** Carry out the analysis of the RL circuit explicitly.
- B** A gopher gnaws through one of the wires in the DC lighting system in your front yard, and the lights turn off. At the instant when the circuit becomes open, we can consider the bare ends of the wire to be like the plates of a capacitor, with an air gap (or gopher gap) between them. What kind of capacitance value are we talking about here? What would this tell you about the RC time constant?

Notes for chapter 13

≥317 The loop rule in an AC circuit

Why does the loop rule work at all in AC circuits?

When we take measurements with an AC voltmeter, we find in practice that the loop rule often holds for AC circuits. But this is a little bit like the story of the centipede who is asked how he keeps all his legs coordinated and replies, “You know, I never thought about that before,” after which he is no longer able to walk. The loop rule is essentially a statement that the electric field has $\text{curl } \mathbf{E} = 0$, but this is false when we have induction. One of Maxwell’s equations (sec. 6.7, p. 163) is that $\text{curl } \mathbf{E} = -\partial \mathbf{B}/\partial t$, which tells us that the change in the magnetic field will cause a curly electric field. There is thus no well-defined electric potential, and the loop rule should be invalid. So why does it seem to work here when it shouldn’t? The answer is that when the lumped circuit approximation is valid, the external magnetic fields of circuit elements like inductors are negligible, and therefore it is approximately true, in the empty space surrounding the circuit where we place our voltmeters, that $\text{curl } \mathbf{E} = 0$. Because of these issues, many writers will define a symbol such as \mathcal{E} , referred to as “emf,”¹ which is minus the work done by an electric field between one point in space and another. This is essentially the same as the definition of a voltage difference, but without the connotation that there is a well-defined electric potential.

≥321 Overdamping

Overdamping occurs for $Q < 1/2$.

As shown on p. 320, the solutions to the equation of motion for a series LRC circuit are of the form $q = e^{st}$, where $Ls^2 + Rs + 1/C = 0$. The roots of this polynomial equation are given by the quadratic formula, which contains the square root of the discriminant $R^2 - 4L/C$.

¹standing for “electromotive force,” a horrifically bad piece of terminology, since it doesn’t even have units of force

By analogy with the mechanical case (p. 300), $Q = C^{-1/2}L^{1/2}R^{-1}$, so the discriminant is positive when $Q < 1/2$, which is referred to as the case of overdamping.

≥340 Fourier analysis

Any linear circuit element’s behavior can be fully characterized by its behavior on sine waves.

The idea is that according to a set of related mathematical results called Fourier’s theorem and the Fourier inversion theorem, we expect to be able to write any sufficiently well-behaved function as a sum of sine waves. For periodic functions, the sum is in general an infinite series. For example, a square wave can be written as $\widehat{\mathcal{M}} + \frac{1}{3}\widehat{\mathcal{M}} + \frac{1}{5}\widehat{\mathcal{M}} + \dots$ For nonperiodic functions, the sum will usually have to be an integral, i.e., a continuous sum. Although the mathematical theorems require the functions to have certain properties, physicists and engineers normally don’t worry about them, or if necessary we generalize the notion of a function to try to make them work. An example of such a generalization is the Dirac delta function.

Problems

Key

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

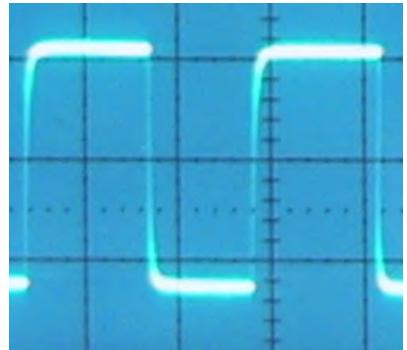
1 Suppose that an FM radio tuner for the US commercial broadcast band (88-108 MHz) consists of a series LRC circuit. If the inductor is $1.0 \mu\text{H}$, what range of capacitances should the variable capacitor be able to provide? ✓

2 If we take the henry to be defined by the relation $U = (1/2)LI^2$ in SI units, show that L/R has units of time.

▷ Solution, p. 431

3 The figure shows an oscilloscope trace for a signal that is meant to be a square wave. Of course a real-life signal can never be a perfect, idealized shape such as a square wave. Here we see that both the leading and trailing edges look like exponential functions. Suppose that the time constant of these exponentials is 5 ms (about 1/10 of a division), the resistance in the circuit is 1.0Ω , and the exponential behavior is due to a stray capacitance in series with the resistance.

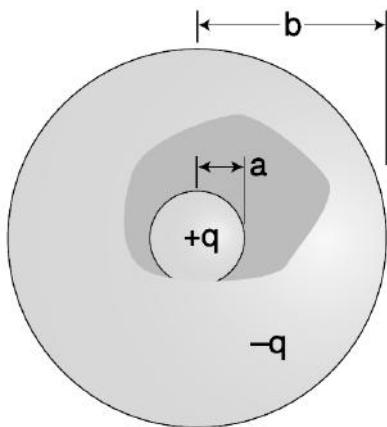
- (a) Estimate the stray capacitance. ✓
- (b) Verify directly that the units of your answer work out to be farads, if the farad is defined by the definition of capacitance $U = q^2/2C$.



Problem 3.

4 Find an expression for the energy stored in a capacitor in terms of C and V . ✓

5 Suppose that a parallel-plate capacitor is constructed so that the gap h between the plates can be changed. Find proportionalities describing what happens to the energy stored in the capacitor if this is done (a) at fixed voltage, and (b) at fixed charge.



Problem 6. Part of the outside sphere has been drawn as if it is transparent, in order to show the inside sphere.

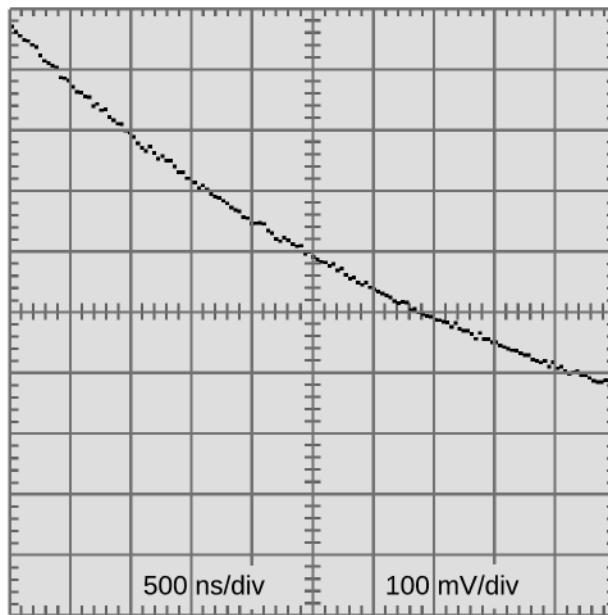
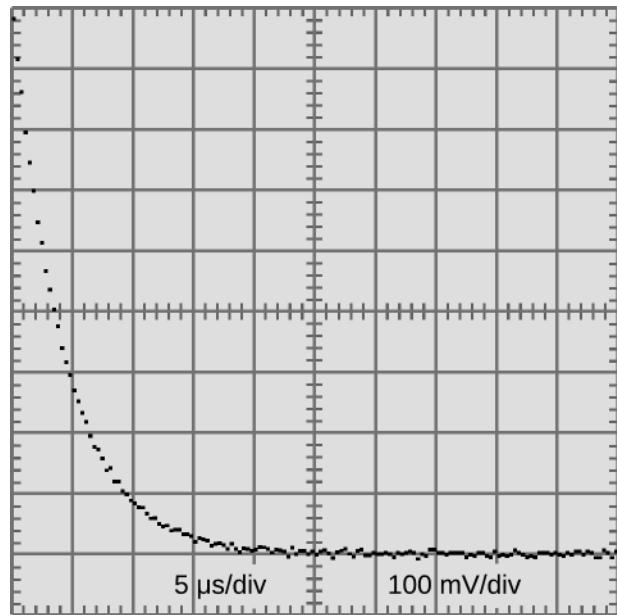
6 (a) In example 6, p. 99, we analyzed a spherical capacitor. Building on that analysis, let's find its capacitance. There are two ways that might occur to us to approach this, using either one of the relations $U = q^2/2C$ and $V = q/C$. Complete the calculation by whichever method seems like it will be easier. ✓

(b) In the limit where the gap between a and b is very small, show that you can recover the result of example 1, p. 310, for a parallel-plate capacitor, i.e., the curvature doesn't matter in this limit.

(c) Find the capacitance of the surface of the earth, assuming there is an outer spherical "plate" at infinity. (In reality, this outer plate would just represent some distant part of the universe to which we carried away some of the earth's charge in order to charge up the earth.)

✓

7 Resistors have a standard color code on them, but capacitors often have only cryptic part numbers printed on them that don't say the value of the component. Suppose we have an unknown capacitor, and we want to find its value. We take a known resistance $R = 4.7 \text{ k}\Omega$ and form a series RC circuit. Hooking it up to a square-wave generator, and observing the results on a digital oscilloscope, we see the results shown in the figures. These are two versions of the same electrical signal on the scope. Note the scales shown at the bottom. The only difference between the two traces is time scale. Find the unknown capacitance. Why would it be necessary to look at both scales in order to get a good measurement? ✓



Problem 7.

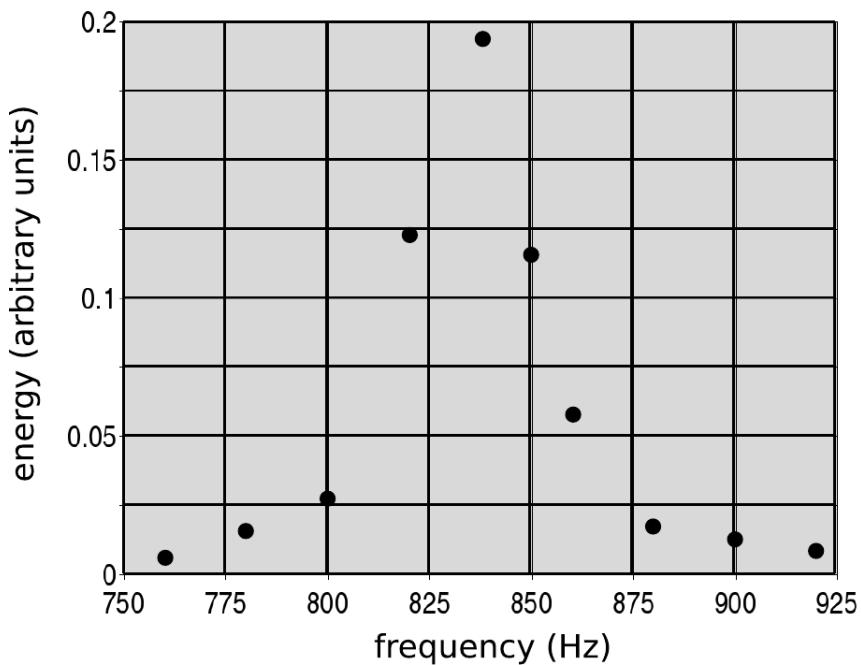
8 In a series LR circuit, find the time required for the *energy* in the inductor's field (not the current or voltage) to fall off by a factor of e . \checkmark

9 If the inductance in an LC circuit is increased by a factor of 100, what will happen to the resonant frequency?

▷ Solution, p. 431

10 (a) Carry out the algebra in example 2, p. 311, to show that the inductance of an ideal solenoid is $(4\pi k/c^2)N^2A/\ell$. (For a real solenoid, the interior field will be smaller and nonuniform, and the exterior field will be nonzero.)

(b) A particular solenoid used in my lab classes has roughly $N = 3000$, $\ell = 10$ cm, and a diameter of about 10 cm. Estimate its inductance. This will be only a rough estimate, because the solenoid is not very long compared to its diameter, and is therefore very nonideal.



Problem 11.

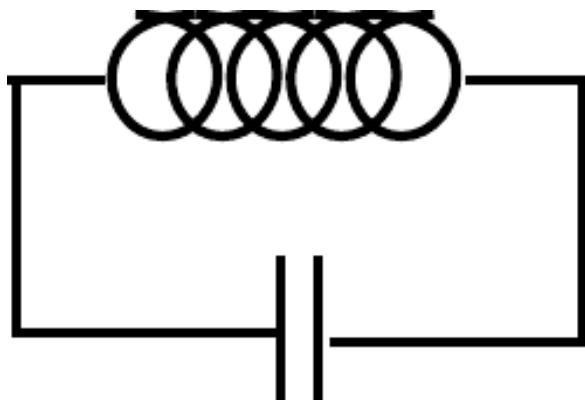
11 The graph, measured by a Fullerton College student, shows the steady-state response of an LRC circuit as a function of frequency. Find the Q of the resonance. \checkmark

Minilab 13: Energy in electric and magnetic fields

Apparatus

coil ($\sim 1 \text{ H}$)
 $0.01 \mu\text{F}$ capacitor
oscilloscope
sine wave generator

Goal: Observe how the energy content of the electric and magnetic fields relates to the fields' magnitudes.

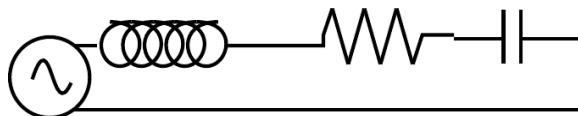


a / A simplified version of the circuit.

The basic idea of this lab is to observe a circuit like the one shown in figure a, consisting of a capacitor and an inductor. Imagine that we first deposit positive and negative charges on the plates of the capacitor. If we imagined that the universe was purely mechanical, obeying Newton's laws of motion, we would expect that the attractive force between these charges would cause them to come back together and reestablish a stable equilibrium in which there was zero net charge everywhere in the circuit.

However, the capacitor in its initial, charged, state has an electric field between its plates, and this field possesses energy. This energy can't just go away, because energy is conserved. What really happens is that as charge starts to flow off of the capacitor plates, a current is established in the coil. This current creates a magnetic field in the space inside and

around the coil. The electric energy doesn't just evaporate; it turns into magnetic energy. We end up with an oscillation in which the capacitor and the coil trade energy back and forth. Your goal is to monitor this energy exchange, and to use it to deduce a power-law relationship between each field and its energy.



b / The actual circuit.

The practical realization of the circuit involves some further complications, as shown in figure b.

The wires are not superconductors, so the circuit has some nonzero resistance, and the oscillations would therefore gradually die out, as the electric and magnetic energies were converted to heat. The sine wave generator serves both to initiate the oscillations and to maintain them, replacing, in each cycle, the energy that was lost to heat.

Furthermore, the circuit has a resonant frequency at which it prefers to oscillate, and when the resistance is very small, the width of the resonance is very narrow. To make the resonance wider and less finicky, we intentionally insert a $10 \text{ k}\Omega$ resistor. The inductance of the coil is about a henry, giving a resonant frequency in the low kilohertz range.

Observations

Let E be the magnitude of the electric field between the capacitor plates, and let \tilde{E} be the maximum value of this quantity. It is then convenient to define $x = E/\tilde{E}$, a unitless quantity ranging from -1 to 1 . Similarly, let $y = B/\tilde{B}$ for the corresponding magnetic quantities. The electric field is proportional to the voltage difference across the capacitor plates, which is something we can measure di-

rectly using the oscilloscope:

$$x = \frac{E}{\tilde{E}} = \frac{V_C}{\tilde{V}_C}$$

The magnetic field in the coil is proportional to the current. Unfortunately, an oscilloscope doesn't measure current, so there's no equally direct way to get a handle on the magnetic field. However, all the current that goes through the coil must also go through the resistor, and Ohm's law relates the current through the resistor to the voltage drop across it. This voltage drop is something we can measure with the oscilloscope, so we have

$$y = \frac{B}{\tilde{B}} = \frac{I}{\tilde{I}} = \frac{V_R}{\tilde{V}_R}$$

To measure x and y , you need to connect channels 1 and 2 of the oscilloscope across the resistor and the capacitor. Since both channels of the scope are grounded on one side (the side with the ground tab on the banana-to-bnc connector), you need to make sure that their grounded sides both go to the piece of wire between the resistor and the capacitor. Furthermore, one output of the sine wave generator is normally grounded, which would mess everything up: two different points in the circuit would be grounded, which would mean that there would be a short across some of the circuit elements. To avoid this, loosen the banana plug connectors on the sine wave generator, and swing away the piece of metal that normally connects one of the output plugs to the ground.

Tune the sine wave generator's frequency to resonance, and take the data you'll need in order to determine x and y at a whole bunch of different places over one cycle.

The quality of the results can depend a lot on the quality of the connections. If the display on the scope changes noticeably when you wiggle the wires, you have a problem.

Analysis

Plot y versus x on a piece of graph paper. Let's assume that the energy in a field depends on

the field's strength raised to some power p . Conservation of energy then gives

$$|x|^p + |y|^p = 1$$

Use your graph to determine p , and interpret your result.

Prelab

P1. Sketch what your graph would look like for $p = 0.1$, $p = 1$, $p = 2$, and $p = 10$. (You should be able to do $p = 1$ and $p = 2$ without any computations. For $p = 0.1$ and $p = 10$, you can either run some numbers on your calculator or use your mathematical knowledge to sketch what they would turn out like.)

Chapter 14

Impedance

14.1 Impedance

When a resistor, capacitor, or inductor is driven by an oscillating voltage with amplitude V and frequency ω , current will flow through it (once the steady state has been achieved) with the same frequency and with some amplitude and phase. If we use the idea introduced in sec. 12.3, p. 296 of representing a sine wave with a complex number, then we have a complex number \tilde{V} that completely embodies the voltage we apply, and another complex number \tilde{I} describing the outcome. As a notational convention, we write these symbols with tildes on top. The tilde looks like a little sine wave and reminds us that we're representing a sine wave. We can then define a complex number

$$Z = \frac{\tilde{V}}{\tilde{I}}$$

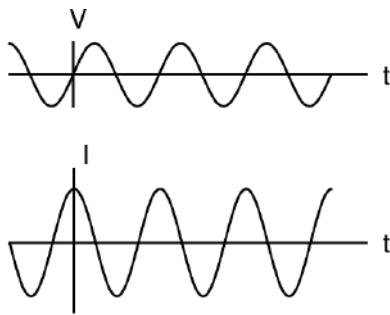
called the *impedance*. The impedance has units of ohms, and in the case of a resistor, Z is a real number that equals the resistance.

In the purely capacitive case, the relation $V = q/C$ lets us calculate $I = \frac{dq}{dt} = C \frac{dV}{dt}$. A capacitor does not follow Ohm's law, and it is not true that $I = V/R$, because taking the derivative of a sinusoidal function shifts its phase by 90 degrees. But we can still define an impedance. If the voltage varies as, for example, $V(t) = V_0 \sin(\omega t)$, then the current will be $I(t) = \omega C V_0 \cos(\omega t)$. We use the complex numbers 1 and i represent the functions $\sin \omega t$ and $\cos \omega t$. In our example, $V(t)$ is a sine wave multiplied by a number that gives its amplitude, so we associate that function with a number \tilde{V} lying on the real axis. Its magnitude, $|\tilde{V}|$, gives the amplitude in units of volts, while its argument $\arg \tilde{V}$, gives its phase angle, which is zero. The current is a multiple of the cosine, so we identify it with a number \tilde{I} lying on the imaginary axis. We have $\arg \tilde{I} = 90^\circ$, and $|\tilde{I}|$ is the amplitude of the current, in units of amperes. But comparing with our result above, we have $|\tilde{I}| = \omega C |\tilde{V}|$. Bringing together the phase and magnitude information, we have $\tilde{I} = i\omega C \tilde{V}$. This looks very much like Ohm's law, so we write

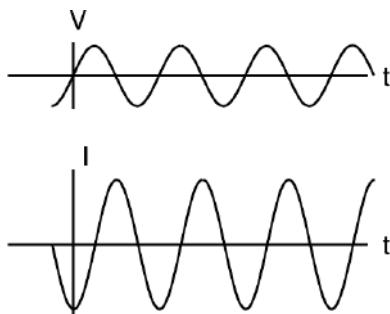
$$\tilde{I} = \frac{\tilde{V}}{Z_C},$$

where the quantity

$$Z_C = -\frac{i}{\omega C}, \quad [\text{impedance of a capacitor}]$$



a / In a capacitor, the current is 90° ahead of the voltage in phase.



b / The current through an inductor lags behind the voltage by a phase angle of 90° .

having units of ohms, is the *impedance* of the capacitor at this frequency.

It makes sense that the impedance becomes infinite at zero frequency. Zero frequency means that it would take an infinite time before the voltage would change by any amount. In other words, this is like a situation where the capacitor has been connected across the terminals of a battery and been allowed to settle down to a state where there is constant charge on both terminals. Since the electric fields between the plates are constant, there is no energy being added to or taken out of the field. A capacitor that can't exchange energy with any other circuit component is nothing more than a broken (open) circuit.

self-check A

Why can't a capacitor have its impedance printed on it along with its capacitance?

▷ Answer, p. 433

Similar math (but this time with an integral instead of a derivative) gives

$$Z_L = i\omega L \quad [\text{impedance of an inductor}]$$

for an inductor. It makes sense that the inductor has lower impedance at lower frequencies, since at zero frequency there is no change in the magnetic field over time. No energy is added to or released from the magnetic field, so there are no induction effects, and the inductor acts just like a piece of wire with negligible resistance. The term "choke" for an inductor refers to its ability to "choke out" high frequencies.

The phase relationships shown in figures a and b can be remembered using the mnemonic "eVIL," which shows that the voltage (V) leads the current (I) in an inductive circuit, while the opposite is true in a capacitive one.

Summarizing, the impedances of resistors, capacitors, and inductors are

$$\begin{aligned} Z_R &= R \\ Z_C &= -\frac{i}{\omega C} \\ Z_L &= i\omega L. \end{aligned}$$

Low-pass and high-pass filters

example 1

An LRC circuit only responds to a certain range (band) of frequencies centered around its resonant frequency. As a filter, this is known as a bandpass filter. If you turn down both the bass and the treble on your stereo, you have created a bandpass filter.

To create a high-pass or low-pass filter, we only need to insert a capacitor or inductor, respectively, in series. For instance, a

very basic surge protector for a computer could be constructed by inserting an inductor in series with the computer. The desired 60 Hz power from the wall is relatively low in frequency, while the surges that can damage your computer show much more rapid time variation. Even if the surges are not sinusoidal signals, we can think of a rapid “spike” qualitatively as if it was very high in frequency — like a high-frequency sine wave, it changes very rapidly.

Inductors tend to be big, heavy, expensive circuit elements, so a simple surge protector would be more likely to consist of a capacitor in *parallel* with the computer. (In fact one would normally just connect one side of the power circuit to ground via a capacitor.) The capacitor has a very high impedance at the low frequency of the desired 60 Hz signal, so it siphons off very little of the current. But for a high-frequency signal, the capacitor’s impedance is very small, and it acts like a zero-impedance, easy path into which the current is diverted.

The main things to be careful about with impedance are that (1) the concept only applies to a circuit that is being driven sinusoidally, (2) the impedance of an inductor or capacitor is frequency-dependent.

Discussion question

A Figure a on page 332 shows the voltage and current for a capacitor. Sketch the q - t graph, and use it to give a physical explanation of the phase relationship between the voltage and current. For example, why is the current zero when the voltage is at a maximum or minimum?

B Figure b on page 332 shows the voltage and current for an inductor. The power is considered to be positive when energy is being put into the inductor’s magnetic field. Sketch the graph of the power, and then the graph of U , the energy stored in the magnetic field, and use it to give a physical explanation of the P - t graph. In particular, discuss why the frequency is doubled on the P - t graph.

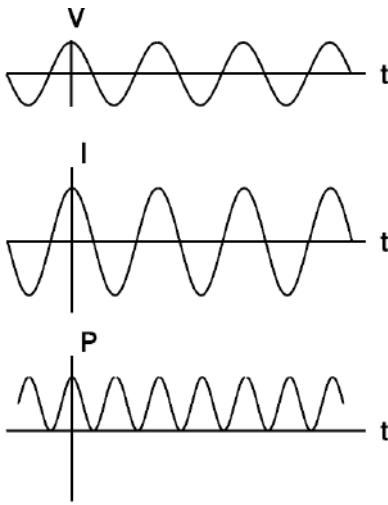
C Relate the features of the graph in figure b on page 332 to the story told in cartoons in figure n/2-3 on page 318.

14.2 Power

How much power is delivered when an oscillating voltage is applied to an impedance? The equation $P = IV$ is generally true, since voltage is defined as energy per unit charge, and current is defined as charge per unit time: multiplying them gives energy per unit time. In a DC circuit, all three quantities were constant, but in an oscillating (AC) circuit, all three display time variation.

14.2.1 A resistor

First let’s examine the case of a resistor. For instance, you’re probably reading this book from a piece of paper illuminated by



c / Power in a resistor: the rate at which electrical energy is being converted into heat.

a glowing lightbulb, which is driven by an oscillating voltage with amplitude V_0 . In the special case of a resistor, we know that I and V are in phase. For example, if V varies as $V_0 \cos \omega t$, then I will be a cosine as well, $I_0 \cos \omega t$. The power is then $I_0 V_0 \cos^2 \omega t$, which is always positive,¹ and varies between 0 and $I_0 V_0$. Even if the time variation was $\cos \omega t$ or $\sin(\omega t + \pi/4)$, we would still have a maximum power of $I_0 V_0$, because both the voltage and the current would reach their maxima at the same time. In a lightbulb, the moment of maximum power is when the circuit is most rapidly heating the filament. At the instant when $P = 0$, a quarter of a cycle later, no current is flowing, and no electrical energy is being turned into heat. Throughout the whole cycle, the filament is getting rid of energy by radiating light. Since the circuit oscillates at a frequency² of 60 Hz, the temperature doesn't really have time to cycle up or down very much over the 1/60 s period of the oscillation, and we don't notice any significant variation in the brightness of the light, even with a short-exposure photograph.

Thus, what we really want to know is the average power, “average” meaning the average over one full cycle. Since we’re covering a whole cycle with our average, it doesn’t matter what phase we assume. Let’s use a cosine. The total amount of energy transferred over one cycle is

$$\begin{aligned} E &= \int dE \\ &= \int_0^T \frac{dE}{dt} dt, \end{aligned}$$

where $T = 2\pi/\omega$ is the period.

$$\begin{aligned} E &= \int_0^T P dt \\ &= \int_0^T I_0 V_0 \cos^2 \omega t dt \\ &= I_0 V_0 \int_0^T \cos^2 \omega t dt \\ &= I_0 V_0 \int_0^T \frac{1}{2} (1 + \cos 2\omega t) dt \end{aligned}$$

¹A resistor always turns electrical energy into heat. It never turns heat into electrical energy!

²Note that this time “frequency” means f , not ω ! Physicists and engineers generally use ω because it simplifies the equations, but electricians and technicians always use f . The 60 Hz frequency is for the U.S.

The reason for using the trig identity $\cos^2 x = (1 + \cos 2x)/2$ in the last step is that it lets us get the answer without doing a hard integral. Over the course of one full cycle, the quantity $\cos 2\omega t$ goes positive, negative, positive, and negative again, so the integral of it is zero. We then have

$$\begin{aligned} E &= I_o V_o \int_0^T \frac{1}{2} dt \\ &= \frac{I_o V_o T}{2} \end{aligned}$$

The average power is

$$\begin{aligned} P_{av} &= \frac{\text{energy transferred in one full cycle}}{\text{time for one full cycle}} \\ &= \frac{I_o V_o T / 2}{T} \\ &= \frac{I_o V_o}{2}, \end{aligned}$$

i.e., the average is half the maximum. The power varies from 0 to $I_o V_o$, and it spends equal amounts of time above and below the maximum, so it isn't surprising that the average power is half-way in between zero and the maximum. Summarizing, we have

$$P_{av} = \frac{I_o V_o}{2} \quad [\text{average power in a resistor}]$$

for a resistor.

14.2.2 RMS quantities

Suppose one day the electric company decided to start supplying your electricity as DC rather than AC. How would the DC voltage have to be related to the amplitude V_o of the AC voltage previously used if they wanted your lightbulbs to have the same brightness as before? The resistance of the bulb, R , is a fixed value, so we need to relate the power to the voltage and the resistance, eliminating the current. In the DC case, this gives $P = IV = (V/R)V = V^2/R$. (For DC, P and P_{av} are the same.) In the AC case, $P_{av} = I_o V_o / 2 = V_o^2 / 2R$. Since there is no factor of $1/2$ in the DC case, the same power could be provided with a DC voltage that was smaller by a factor of $1/\sqrt{2}$. Although you will hear people say that household voltage in the U.S. is 110 V, its amplitude is actually $(110 \text{ V}) \times \sqrt{2} \approx 160 \text{ V}$. The reason for referring to $V_o/\sqrt{2}$ as "the" voltage is that people who are naive about AC circuits can plug $V_o/\sqrt{2}$ into a familiar DC equation like $P = V^2/R$ and get the right *average* answer. The quantity $V_o/\sqrt{2}$ is called the "RMS" voltage, which stands for "root mean square." The idea is that if you square the function $V(t)$, take its average (mean) over one cycle, and then take the square root of that average, you get $V_o/\sqrt{2}$. Many digital meters provide RMS readouts for measuring AC voltages and currents.

14.2.3 A capacitor

For a capacitor, the calculation starts out the same, but ends up with a twist. If the voltage varies as a cosine, $V_o \cos \omega t$, then the relation $I = C dV/dt$ tells us that the current will be some constant multiplied by minus the sine, $-V_o \sin \omega t$. The integral we did in the case of a resistor now becomes

$$E = \int_0^T -I_o V_o \sin \omega t \cos \omega t dt,$$

and based on figure d, you can easily convince yourself that over the course of one full cycle, the power spends two quarter-cycles being negative and two being positive. In other words, the average power is zero!

Why is this? It makes sense if you think in terms of energy. A resistor converts electrical energy to heat, never the other way around. A capacitor, however, merely stores electrical energy in an electric field and then gives it back. For a capacitor,

$$P_{av} = 0 \quad [\text{average power in a capacitor}]$$

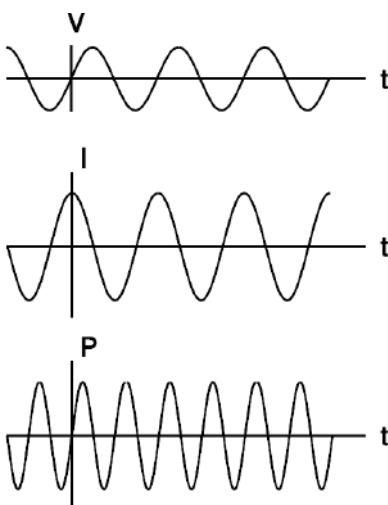
Notice that although the average power is zero, the power at any given instant is *not* typically zero, as shown in figure d. The capacitor *does* transfer energy: it's just that after borrowing some energy, it always pays it back in the next quarter-cycle.

14.2.4 An inductor

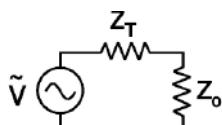
The analysis for an inductor is similar to that for a capacitor: the power averaged over one cycle is zero. Again, we're merely storing energy temporarily in a field (this time a magnetic field) and getting it back later.

14.3 Impedance matching

Figure e shows a commonly encountered situation: we wish to maximize the average power, P_{av} , delivered to the load for a fixed value of V_o , the amplitude of the oscillating driving voltage. We assume that the impedance of the transmission line, Z_T is a fixed value, over which we have no control, but we are able to design the load, Z_o , with any impedance we like. For now, we'll also assume that both impedances are resistive. For example, Z_T could be the resistance of a long extension cord, and Z_o could be a lamp at the end of it. The result generalizes immediately, however, to any kind of impedance. For example, the load could be a stereo speaker's magnet coil, which displays both inductance and resistance. (For a purely inductive or capacitive load, P_{av} equals zero, so the problem isn't very interesting!)



d / Power in a capacitor: the rate at which energy is being stored in (+) or removed from (-) the electric field.



e / We wish to maximize the power delivered to the load, Z_o , by adjusting its impedance.

Since we're assuming both the load and the transmission line are resistive, their impedances add in series, and the amplitude of the current is given by

$$I_o = \frac{V_o}{Z_o + Z_T},$$

so

$$\begin{aligned} P_{av} &= I_o V_o / 2 \\ &= I_o^2 Z_o / 2 \\ &= \frac{V_o^2 Z_o}{(Z_o + Z_T)^2} / 2. \end{aligned}$$

The maximum of this expression occurs where the derivative is zero,

$$\begin{aligned} 0 &= \frac{1}{2} \frac{d}{dZ_o} \left[\frac{V_o^2 Z_o}{(Z_o + Z_T)^2} \right] \\ 0 &= \frac{1}{2} \frac{d}{dZ_o} \left[\frac{Z_o}{(Z_o + Z_T)^2} \right] \\ 0 &= (Z_o + Z_T)^{-2} - 2Z_o (Z_o + Z_T)^{-3} \\ 0 &= (Z_o + Z_T) - 2Z_o \\ Z_o &= Z_T \end{aligned}$$

In other words, to maximize the power delivered to the load, we should make the load's impedance the same as the transmission line's. This result may seem surprising at first, but it makes sense if you think about it. If the load's impedance is too high, it's like opening a switch and breaking the circuit; no power is delivered. On the other hand, it doesn't pay to make the load's impedance too small. Making it smaller does give more current, but no matter how small we make it, the current will still be limited by the transmission line's impedance. As the load's impedance approaches zero, the current approaches this fixed value, and the the power delivered, $I_o^2 Z_o$, decreases in proportion to Z_o .

Maximizing the power transmission by matching Z_T to Z_o is called *impedance matching*. For example, an 8-ohm home stereo speaker will be correctly matched to a home stereo amplifier with an internal impedance of 8 ohms, and 4-ohm car speakers will be correctly matched to a car stereo with a 4-ohm internal impedance. You might think impedance matching would be unimportant because even if, for example, we used a car stereo to drive 8-ohm speakers, we could compensate for the mismatch simply by turning the volume knob higher. This is indeed one way to compensate for any impedance mismatch, but there is always a price to pay. When the impedances are matched, half the power is dissipated in the transmission line and half in the load. By connecting a 4-ohm

amplifier to an 8-ohm speaker, however, you would be setting up a situation in which two watts were being dissipated as heat inside the amp for every watt being delivered to the speaker. In other words, you would be wasting energy, and perhaps burning out your amp when you turned up the volume to compensate for the mismatch.

14.4 Impedances in series and parallel

How do impedances combine in series and parallel? The beauty of treating them as complex numbers is that they simply combine according to the same rules you've already learned for resistances.

Series impedance

example 2

- ▷ A capacitor and an inductor in series with each other are driven by a sinusoidally oscillating voltage. At what frequency is the current maximized?
- ▷ Impedances in series, like resistances in series, add. The capacitor and inductor act as if they were a single circuit element with an impedance

$$\begin{aligned} Z &= Z_L + Z_C \\ &= i\omega L - \frac{i}{\omega C}. \end{aligned}$$

The current is then

$$\tilde{I} = \frac{\tilde{V}}{i\omega L - i/\omega C}.$$

We don't care about the phase of the current, only its amplitude, which is represented by the absolute value of the complex number \tilde{I} , and this can be maximized by making $|i\omega L - i/\omega C|$ as small as possible. But there is some frequency at which this quantity is zero —

$$\begin{aligned} 0 &= i\omega L - \frac{i}{\omega C} \\ \frac{1}{\omega C} &= \omega L \\ \omega &= \frac{1}{\sqrt{LC}} \end{aligned}$$

At this frequency, the current is infinite! What is going on physically? This is an LRC circuit with $R = 0$. It has a resonance at this frequency, and because there is no damping, the response at resonance is infinite. Of course, any real LRC circuit will have some damping, however small.

Resonance with damping

example 3

- ▷ What is the amplitude of the current in a series LRC circuit?

▷ Generalizing from example 2, we add a third, real impedance:

$$\begin{aligned}\tilde{|I|} &= \frac{|\tilde{V}|}{|Z|} \\ &= \frac{|\tilde{V}|}{|R + i\omega L - i/\omega C|} \\ &= \frac{|\tilde{V}|}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}}\end{aligned}$$

This result would have taken pages of algebra without the complex number technique!

A second-order stereo crossover filter *example 4*

A stereo crossover filter ensures that the high frequencies go to the tweeter and the lows to the woofer. This can be accomplished simply by putting a single capacitor in series with the tweeter and a single inductor in series with the woofer. However, such a filter does not cut off very sharply. Suppose we model the speakers as resistors. (They really have inductance as well, since they have coils in them that serve as electromagnets to move the diaphragm that makes the sound.) Then the power they draw is $I^2 R$. Putting an inductor in series with the woofer, $f/1$, gives a total impedance that at high frequencies is dominated by the inductor's, so the current is proportional to ω^{-1} , and the power drawn by the woofer is proportional to ω^{-2} .

A second-order filter, like $f/2$, is one that cuts off more sharply: at high frequencies, the power goes like ω^{-4} . To analyze this circuit, we first calculate the total impedance:

$$Z = Z_L + (Z_C^{-1} + Z_R^{-1})^{-1}$$

All the current passes through the inductor, so if the driving voltage being supplied on the left is \tilde{V}_d , we have

$$\tilde{V}_d = \tilde{I}_L Z,$$

and we also have

$$\tilde{V}_L = \tilde{I}_L Z_L.$$

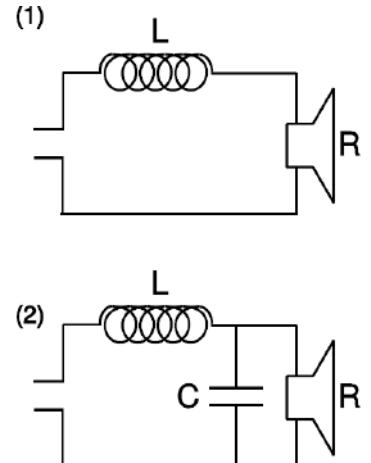
The loop rule, applied to the outer perimeter of the circuit, gives

$$\tilde{V}_d = \tilde{V}_L + \tilde{V}_R.$$

Straightforward algebra now results in

$$\tilde{V}_R = \frac{\tilde{V}_d}{1 + Z_L/Z_C + Z_L/Z_R}.$$

At high frequencies, the Z_L/Z_C term, which varies as ω^2 , dominates, so \tilde{V}_R and \tilde{I}_R are proportional to ω^{-2} , and the power is proportional to ω^{-4} .



f / Example 4.

14.5 Capacitors and inductors in series and parallel

In examples 3 and 4 on p. 312, we found that inductance values add in series (just like resistors), while capacitances add in parallel. What about inductances in parallel and capacitances in series? A simple way to get at this more generally is to use our knowledge of impedances. If, for example, a linear circuit element acts like a capacitor for all sinusoidal signals, then it follows that it acts like a capacitor in all cases. That is, a linear circuit element's behavior is fully characterized by its behavior with sine waves ([Z324](#)). Since the behavior on a sine wave is completely characterized by the element's impedance, it follows that we can deduce the rules for parallel and series elements by looking at how the impedances combine. Since the impedance of resistors, capacitors, and inductors is proportional to R , $1/C$, and L , respectively, we find the following rules:

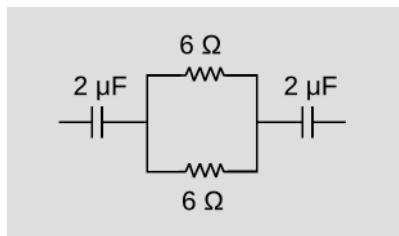
Resistances add in series. Inverse resistances add in parallel. These facts are already familiar from secs. 9.2-9.3, p. 210).

Inductances have the same parallel and series behavior as resistances.

Capacitances add in parallel. Inverse capacitances add in series.

Discussion question

- A Simplify the circuit element shown in the figure.



Discussion question A.

Problems

Key

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

1 Each of the following circuit elements or combination of elements is driven with a sinusoidal voltage at $\omega = 1 \text{ Hz}$. In each case, determine the phase angle, in degrees, of the voltage relative to the current. Define the sign of this phase to be positive if the voltage leads the current, i.e., the phase is the same as the argument of the impedance. Ignore significant figures, i.e., assume that all data are exact.

- (a) A 1 H inductor. ✓
- (b) A 1 F capacitor. ✓
- (c) A 2 H inductor in series with a 1 F capacitor. ✓
- (d) A 1 H inductor in series with a 1 Ω resistor. ✓
- (e) A 2 H inductor in parallel with a 1 F capacitor.

✓

2 (a) Find the parallel impedance of a $37 \text{ k}\Omega$ resistor and a 1.0 nF capacitor at $f = 1.0 \times 10^4 \text{ Hz}$. ✓

(b) A voltage with an amplitude of 1.0 mV drives this impedance at this frequency. What is the amplitude of the current drawn from the voltage source, what is the current's phase angle with respect to the voltage, and does it lead the voltage, or lag behind it? ✓

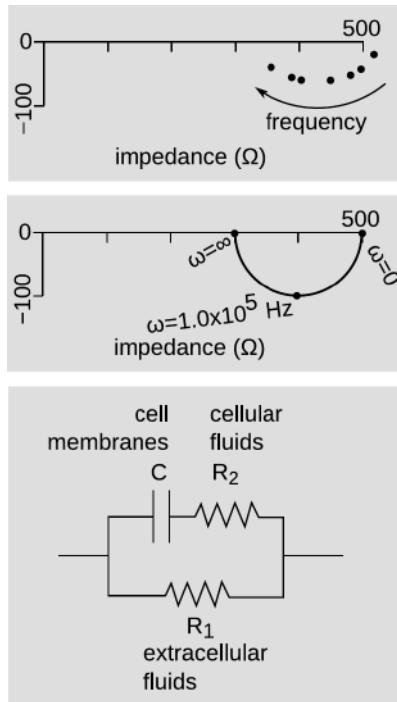
3 A series LRC circuit consists of a 1.000Ω resistor, a 1.000 F capacitor, and a 1.000 H inductor. (These are not particularly easy values to find on the shelf!)

- (a) Plot its impedance as a point in the complex plane for each of the following frequencies: $\omega=0.250, 0.500, 1.000, 2.000,$ and 4.000 Hz .
- (b) What is the resonant angular frequency, ω_{res} , and how does this relate to your plot? ✓
- (c) What is the resonant frequency f_{res} corresponding to your answer in part b? ✓

4 At a frequency of 53.1 kHz , a certain series LR circuit has an impedance of $1.6 \text{ k}\Omega + (1.2 \text{ k}\Omega)i$. Suppose that instead we want to achieve the same impedance using two circuit elements in parallel. What must the elements be? As a check on your answer, you should find that both values are round numbers when rounded off to the correct number of significant figures.

5 (a) A capacitance C and an inductance L are connected in parallel. Show that there is a frequency at which no current at all flows through this system as a whole in response to an externally applied voltage. Find the frequency. This is a “notch” or “band reject” filter. ✓

(b) There will inevitably be some stray resistance. How does this affect the situation described above?



Problem 6.

6 The top panel of the figure shows measurements (Settle *et al.*, 1980) of the impedance of a human body, from an ankle to the opposite wrist. The impedance is shown in the complex plane, at a series of increasing frequencies. This technique can be used as a cheap, noninvasive way of estimating a person’s body composition. The idea is that fat is effectively an insulator, while muscle contains both fluids, which act like a resistance, and cell membranes, which act like capacitors. The middle panel is a simplified version of the graph, cooked up so as to provide round numbers that are easier to work with in a textbook exercise. All the real and imaginary parts of the impedances are multiples of 100, in units of ohms, for the three dots. The bottom panel shows a model used to explain the graph. Use the model to determine the parameters R_1 , R_2 , and C . ✓

7 Resolve the following paradox. A capacitance C is initially charged, and is then connected to another capacitance C , forming a loop. With the charge now shared equally, the energy is halved. If the connection is made using wires that have finite resistance, then this energy loss could be explained through resistive heating. But how is conservation of energy satisfied if the resistance of the wires is zero?

8 (a) In a series LC circuit driven by a DC voltage ($\omega = 0$), compare the energy stored in the inductor to the energy stored in the capacitor.

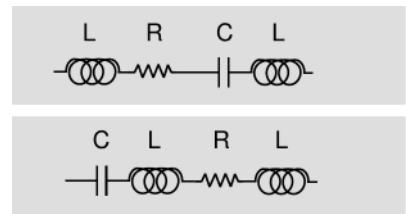
(b) Carry out the same comparison for an LC circuit that is oscillating freely (without any driving voltage).

(c) Now consider the general case of a series LC circuit driven by an oscillating voltage at an arbitrary frequency. Let $\overline{U_L}$ and be the average energy stored in the inductor, and similarly for $\overline{U_C}$. Define a quantity $u = \overline{U_C}/(\overline{U_L} + \overline{U_C})$, which can be interpreted as the capacitor’s average share of the energy, while $1 - u$ is the inductor’s average share. Find u in terms of L , C , and ω , and sketch a graph of u and $1 - u$ versus ω . What happens at resonance? Make sure your result is consistent with your answer to part a. ✓

*

9 Compare the impedances in the figures.

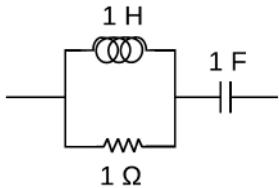
▷ Solution, p. 431



Problem 9.

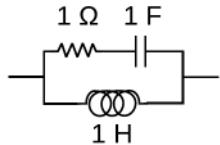
Exercise 14: Impedance: see one, do one

1. Warm-up
 - a) Instructor: compute $(1 - i)^{-1}$.
 - b) Students: compute $(1 + i)^{-1}$.
2. a) Instructor:



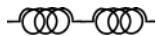
Compute the equivalent impedance at $\omega = 0$, ∞ , and 1 Hz. What units does it have? Interpret the phase of the result at 1 Hz.

- b) Students:



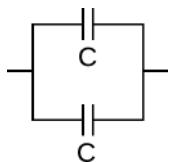
Do the same for this circuit.

3. a) Instructor:



Compute the equivalent impedance in two ways.

- b) Students:



- c) Repeat for L and L in parallel.
d) Repeat for C and C in series.

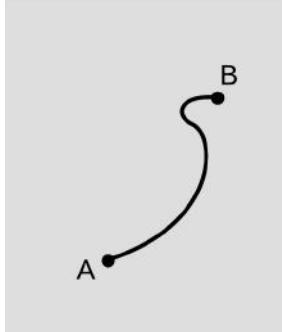
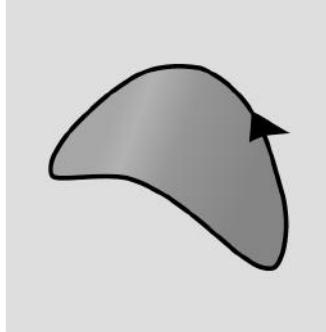
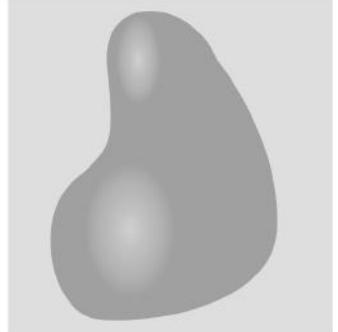
Stokes's theorem

Chapter 15

Stokes's theorem

15.1 A round-up of vector calculus

The table below is a round-up of the basics of vector calculus from this book. All of this has been presented earlier, with the exception of the full statement of Stokes's theorem, which was given in sec. 4.1, p. 85, only for the special case of a “conservative” field, i.e., one whose curl is zero. If you’re a science or engineering major in the US, then most likely you are now at the end of a semester in which you’re taking this electricity and magnetism course concurrently with a vector calculus course, in which this material has been presented in more depth.

		
<i>region</i>	1-dimensional	2-dimensional (oriented surface)
<i>boundary</i>	0-dimensional (two points)	1-dimensional (oriented curve)
<i>derivative operator</i>	grad	curl
<i>linearity</i>	$\text{grad}(\phi_1 + \phi_2) = \text{grad } \phi_1 + \text{grad } \phi_2$	$\text{curl}(\mathbf{F} + \mathbf{G}) = \text{curl } \mathbf{F} + \text{curl } \mathbf{G}$
<i>fundamental theorem:</i>		
<i>integral over interior</i> <i>=stuff on boundary</i>	$\int \text{grad } \phi \cdot d\ell = \phi_B - \phi_A$	$\int \text{curl } \mathbf{F} \cdot d\mathbf{A} = \int \mathbf{F} \cdot d\ell$ (Stokes's theorem)
		$\int \text{div } \mathbf{F} dv = \int \mathbf{F} \cdot d\mathbf{A}$ (Gauss's theorem)

The big picture is that we have three different derivative operators, which take as input either a scalar-valued function or a vector-valued one, and similarly for their outputs. For each of these, there is a theorem that has the same structure as the fundamental theorem of calculus. The familiar version of the fundamental theorem,

from a first-semester calculus course, says that if you integrate a function over some interval, the result is related to the function's properties on the end-points of that interval (it equals the difference between the values of the antiderivative at the end-points). These two end-points constitute the *boundary* of the line segment representing this interval on the real-number line, i.e., if an ant stands on the x axis at some point inside this interval, and it wants to get out, it is going to have to walk past one of the end-points. For each of the three operators of vector calculus, we have a fundamental theorem with a similar structure: the integral over a region is related to what's going on at the boundary of the region.

In the following three examples we demonstrate how this works for each of the three operators when applied to Maxwell's equations, which are recapitulated in the side-bar for convenient reference.

Maxwell's equations
 $\operatorname{div} \mathbf{E} = 4\pi k\rho$

$$\operatorname{div} \mathbf{B} = 0$$

$$\operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\operatorname{curl} \mathbf{B} = \frac{1}{c^2} \frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi k}{c^2} \mathbf{j}$$

The potential and the electric field

example 1

The fundamental theorem of calculus for the gradient is

$$\int \operatorname{grad} \phi \cdot d\ell = \phi_B - \phi_A,$$

where the integral is along a path from A to B. If we let ϕ be an electric potential, then the corresponding electric field is $\mathbf{E} = -\operatorname{grad} \phi$. We then have

$$-\int \mathbf{E} \cdot d\ell = \phi_B - \phi_A,$$

which is a statement that the difference in potential read by a voltmeter equals the work per unit charge that would have to be done to move the charge from A to B, against the field. The field on the interior of the path is related to the potential on the boundary (end-points A and B).

A static electric field

example 2

The fundamental theorem of calculus for the curl, i.e., Stokes's theorem, is

$$\int \operatorname{curl} \mathbf{F} \cdot d\mathbf{A} = \int \mathbf{F} \cdot d\ell,$$

where the integral on the left is over a surface, and the one on the right is taken as we travel around the boundary, i.e., the edge of the surface. One of Maxwell's equations states that $\operatorname{curl} \mathbf{E} = -\partial \mathbf{B}/\partial t$. For a static field, there is no time variation, so $\operatorname{curl} \mathbf{E} = 0$. We then have

$$0 = \int \mathbf{E} \cdot d\ell.$$

This is a statement that for static fields, the work done by the electric field around a closed loop is always zero, as required by conservation of energy.

Electric field and charge density

example 3

The fundamental theorem of calculus for the divergence, i.e., Gauss's theorem, is

$$\int \operatorname{div} \mathbf{F} \, dv = \int \mathbf{F} \cdot d\mathbf{A},$$

where the left-hand integral is over a volume of space, and the one on the right is over the surface that constitutes the boundary of that volume. One of Maxwell's equations is $\operatorname{div} \mathbf{E} = 4\pi k\rho$, so we have

$$4\pi k \int \rho \, dv = \int \mathbf{E} \cdot d\mathbf{A}.$$

This is just Gauss's law, which relates the total charge inside the region to the flux out through its boundary.

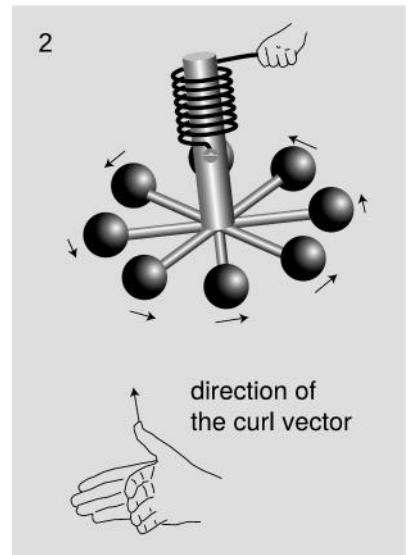
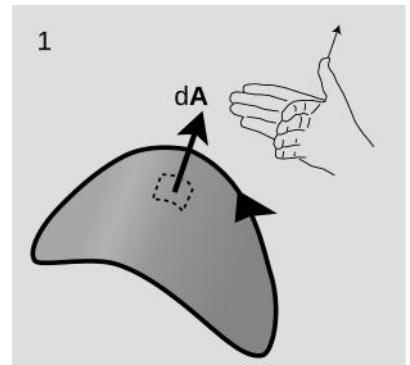
15.2 Stokes's theorem

The only new thing here is Stokes's theorem for the case where the curl is not zero,

$$\int \operatorname{curl} \mathbf{F} \cdot d\mathbf{A} = \int \mathbf{F} \cdot d\ell$$

The integral on the right is sometimes referred to as the circulation of the field \mathbf{F} around a certain closed loop, and in order to define the sign of this circulation, we have to set an orientation for the loop. For example, in figure a/1, the arrowhead on the edge of the potato-chip shape says that we're defining counterclockwise to be the positive direction.

In the integral on the left-hand side, the vector $d\mathbf{A}$ is perpendicular to the surface, as before. This is ambiguous, since there are two opposite directions, each of which is perpendicular to the surface. In our previous applications of this type of surface integral, to Gauss's law, we were using a closed surface, so we resolved this ambiguity just by making our area vectors point outward. Now we're dealing with an open surface, i.e., one with an edge, which doesn't enclose a volume, so there is no longer an "outward" direction. We therefore define the direction of $d\mathbf{A}$ using the right-hand rule shown in figure a/1. This is the same as the right-hand relationship we previously used to define the direction of the curl, figure a/2; the relationship has to be the same, because we want Stokes's theorem to work in the limit of infinitely small curves.



a / 1. The right-hand relationships used in Stokes's theorem.
2. A reminder of the right-hand relationship used to define the direction of the curl.

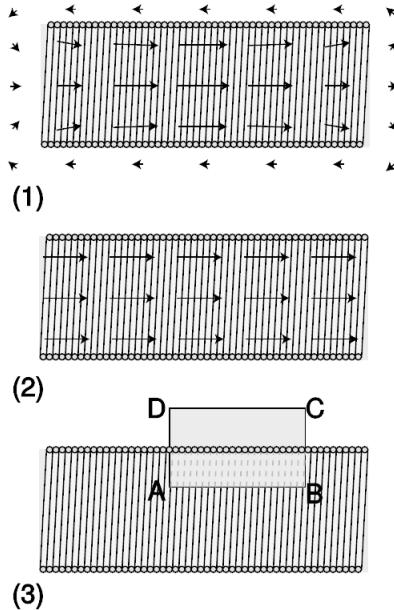
15.3 Ampère's law

If we apply Stokes's theorem to the static magnetic field made by a current, Maxwell's equation $\text{curl } \mathbf{B} = (4\pi k/c^2)\mathbf{j}$ gives

$$\frac{4\pi k}{c^2} \int \mathbf{j} \cdot d\mathbf{A} = \int \mathbf{B} \cdot d\ell.$$

The integral on the left is just the amount of current passing through the loop, so

$$\frac{4\pi k}{c^2} I_{\text{through}} = \int \mathbf{B} \cdot d\ell [\text{static fields}].$$



b / Example 4: a cutaway view of a solenoid.

This result is known as Ampère's law.

A solenoid

example 4

- ▷ What is the field inside a long, straight solenoid of length ℓ and radius a , and having N loops of wire evenly wound along it, which carry a current I ?
- ▷ This is an interesting example, because it allows us to get a very good approximation to the field, but without some experimental input it wouldn't be obvious what approximation to use. Figure b/1 shows what we'd observe by measuring the field of a real solenoid. The field is nearly constant inside the tube, as long as we stay far away from the mouths. The field outside is much weaker. For the sake of an approximate calculation, we can idealize this field as shown in figure b/2. Of the edges of the Ampèrean surface shown in b/3, only AB contributes to the flux — there is zero field along CD, and the field is perpendicular to edges BC and DA. Ampère's law gives

$$\begin{aligned} \int \mathbf{B} \cdot d\ell &= \frac{4\pi k}{c^2} I_{\text{through}} \\ (B)(\text{length of AB}) &= \frac{4\pi k}{c^2} NI \left(\frac{\text{length of AB}}{\ell} \right) \\ B &= \frac{4\pi k NI}{c^2 \ell} \end{aligned}$$

self-check A

What direction is the current in figure b?

▷ Answer, p. 433

15.4 Faraday's law

One of Maxwell's equations is $\text{curl } \mathbf{E} = -\partial \mathbf{B} / \partial t$. This states that a time-varying magnetic field induces a curly electric field (sec. 6.7, p. 163), and we've used it in the analysis of electromagnetic waves as well as in the qualitative description of an electric generator. Application of Stokes's theorem to this equation gives

$$-\int \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{A} = \int \mathbf{E} \cdot d\ell.$$

We're now going to rewrite the left-hand side to put it in a form where it talks about the magnetic flux through the surface. The idea here is that we can interchange the order of the integral and the time derivative, like $\int \frac{\partial}{\partial t} = \frac{\partial}{\partial t} \int$. This is a valid step because an integral is a kind of sum, and the integral of a sum is the sum of the integrals. This makes the left-hand side into $-\frac{\partial}{\partial t} \int \mathbf{B} \cdot d\mathbf{A}$, but $\int \mathbf{B} \cdot d\mathbf{A}$ is just the magnetic flux Φ_B through the surface. We therefore have

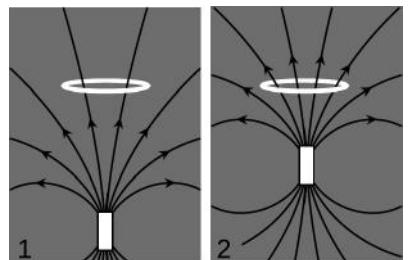
$$-\frac{\partial \Phi_B}{\partial t} = \int \mathbf{E} \cdot d\ell.$$

That is, when a changing magnetic field induces a curly electric field, the circulation of the electric field around a closed loop equals minus the rate at which the magnetic flux through the loop is changing. This is called Faraday's law.

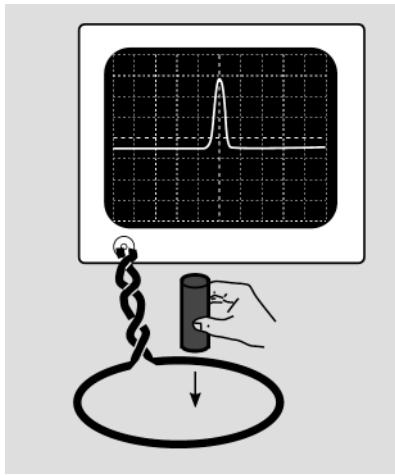
Michael Faraday, whom we met on p. 26, worked before Maxwell, so he didn't derive his result from Maxwell's equations — it was the other way around, with Maxwell putting together Faraday's work on induction along with some other known ingredients to create Maxwell's equations. Faraday never learned calculus, trigonometry, or any math beyond the most basic algebra, so he probably visualized what we now call Faraday's law as shown in figure c.

As the magnet moves from 1 to 2, the flux through the imaginary disk increases. Faraday's law predicts that a curly electric field will be produced, and it predicts its circulation $\int \mathbf{E} \cdot d\ell$ around the edge of the disk to be equal to minus the rate of change of the flux. This electric field exists regardless of whether there is anything physically present at our imaginary disk, but if we happen to have placed a circular loop of wire there, and we connect it to a voltmeter or an oscilloscope, we will measure a voltage, given by the integral defining the circulation of the electric field. The figure only shows a two-dimensional picture of a small number of magnetic field lines, but subject to this crude approximation, we could say that the number of field lines passing through the disk has gone up from 2 to 4. This tells us how much the magnetic flux has increased.

The only way for this number of field lines to change is if the field lines and the loop cut through each other, as in the stage magic trick where a magician passes one steel ring through another, linking



c / The bar magnet makes some magnetic flux through an imaginary disk (white).



d / A real-world setup for measuring the scenario from figure c.

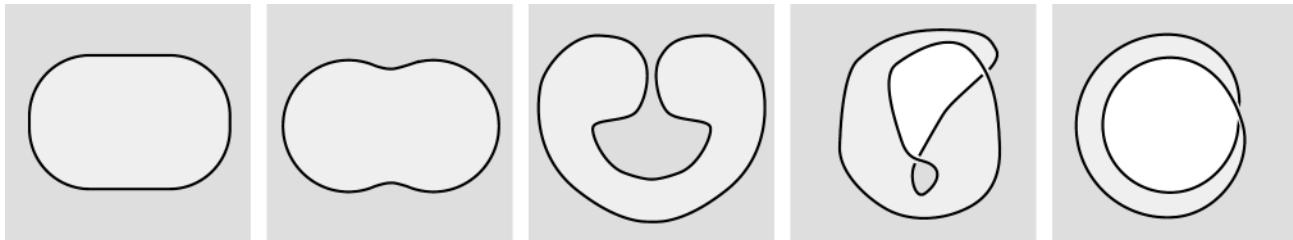


e / Doubling the flux linkage.

them like a chain. For this reason, the flux passing through the disk can be referred to as the *flux linkage*. The induction effect predicted by Faraday's law is proportional to the rate at which the field lines are cutting through the loop.

For ease of visualization, I drew the surface in figure c as a simple disk, whose boundary is a circle. If we were to form this circle from a loop of wire and connect it to an oscilloscope, as in figure d, the signal would probably be easily detectable. But as an electric generator, this is still somewhat impractical. The strength of the effect can be increased greatly by replacing the single loop with a coil. In example 4 on p. 350, we saw that when we use a current in a coil of wire to *create* a magnetic field, the effect was proportional to the number of turns of wire per unit length. Something similar is true when we *expose* a coil of wire to a *changing* magnetic field, in order to produce a voltage. The effect is again proportional to the number of turns of wire.

The easy way to visualize this is in terms of flux linkage and the cutting of the magnetic field lines across the wire. In figure e, the finger represents the magnetic field, and the cord represents the boundary of the imaginary surface used in Faraday's law. In applications, the cord could be a piece of wire. It is obvious that if the finger (magnetic field line) is to pass through the cord like a ghost, it will have to make twice as many cuts through in the case where the cord is wrapped in two loops. Since the induction effect is proportional to the number of cuts, it is doubled.



f / Transforming a single loop into a double loop.

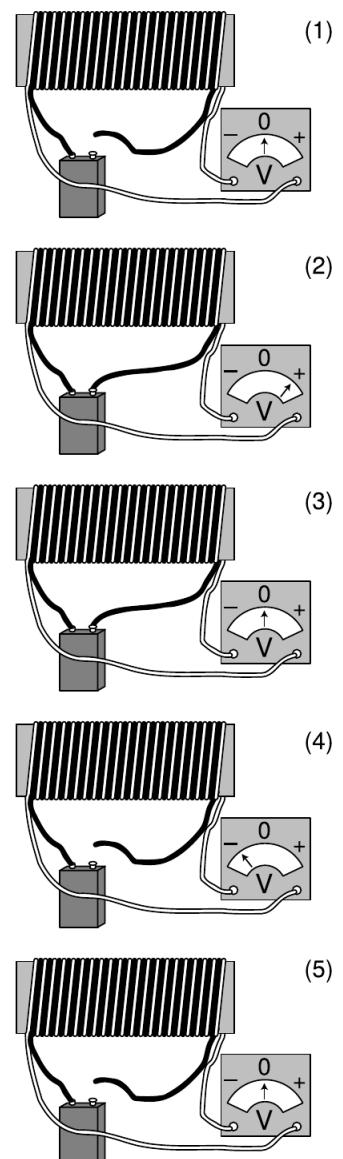
The proportionality to the number of loops is not an exception or modification to Faraday's law, and we don't have to appeal to the "cutting" interpretation to explain it if we don't want to. Figure f shows the transformation of a single loop to a double one. During this process, the surface is like a soap bubble suspended across the wire that forms the boundary. The surface is never torn or glued together. Although I find it a little difficult to visualize the final "soap bubble" surface, it is clearly true that each field line perpendicular to the page will now pass through the surface twice, producing twice the flux compared to a single circular loop with the same circumference.

When we obtained Faraday's law from Maxwell's equations and Stokes's theorem, we assumed implicitly that once we picked a surface, it stayed the same. The only change over time was the change in the magnetic field. But clearly if we viewed the situation in another frame of reference, the surface would be moving, and we expect our laws of physics to be valid regardless of the frame of reference. It is OK for the surface to move, and it can even change size or shape. This is intuitively appealing if we think of the "cutting" interpretation.

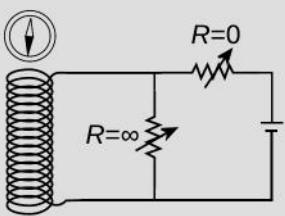
Faraday arrived at what we now call Faraday's law completely empirically in 1831. It's fascinating to read about the original trail of evidence that he followed. Figure g is a simplified drawing of a crucial experiment, as described in his original paper: "Two hundred and three feet of copper wire ... were passed round a large block of wood; [another] two hundred and three feet of similar wire were interposed as a spiral between the turns of the first, and metallic contact everywhere prevented by twine [insulation]. One of these [coils] was connected with a galvanometer [voltmeter], and the other with a battery... When the contact was made, there was a sudden and very slight effect at the galvanometer, and there was also a similar slight effect when the contact with the battery was broken. But whilst the ... current was continuing to pass through the one [coil], no ... effect ... upon the other [coil] could be perceived, although the active power of the battery was proved to be great, by its heating the whole of its own coil [through ordinary resistive heating] ..."

From Faraday's notes and publications, it appears that the situation in figure g/3 was a surprise to him, and he probably thought it would be a surprise to his readers, as well. That's why he offered evidence that the current was still flowing: to show that the battery hadn't just died. The induction effect occurred during the short time it took for the black coil's magnetic field to be established, g/2. Even more counterintuitively, we get an effect, equally strong but in the opposite direction, when the circuit is *broken*, g/4. The effect occurs only when the magnetic field is changing, and it appears to be proportional to the *rate of change* of the magnetic flux through the block (or actually through the corkscrew-shaped surface formed by the white wire), which has one sign when the field is being established, and in the opposite direction when it collapses.

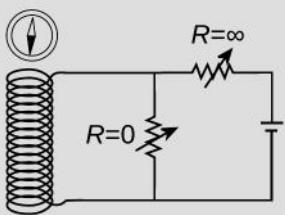
Although Faraday's discovery was empirical, Faraday's law has close logical interrelationships with other principles of physics. In example 5 we see that induction is necessary based on conservation of energy, and in example 6 that it is necessary based on the fact that motion is relative.



g / Faraday's experiment, simplified and shown with modern equipment.



1



2

h / Shorting across an inductor, example 5. Panel 2 shows what is observed immediately after the resistances are changed.

Conservation of energy

example 5

Figure h/1 shows a solenoid with a current being driven through it by a battery. The wire that the solenoid is made of has some finite resistance, so this is not a short circuit, but current is flowing. There is a magnetic field in and around the solenoid. The magnetic compass near the mouth of the solenoid is nearly aligned with the solenoid's axis, showing that this field is much stronger than any ambient field such as the earth's.

So far there has been no obvious reason for having the two variable resistors. The one set to $R = 0$ might as well be a piece of wire, and the one set to $R = \infty$ could just be air. But now suppose that we rapidly change the resistors so that they have the values shown in h/2. In fact, I have a power supply in my lab that seems to do essentially this when I flip its switch to the "off" position.

If our intuition is based solely on experience with DC circuits, then we would expect that the current would instantly cease. The resistance that is now infinite is an open circuit, which means that the power supply is disconnected from the circuit. We have shorted across the inductor, and we know that if we short across a light-bulb, it just winks out.

But that is not at all what happens. The compass stays aligned with the solenoid, showing that the field still exists. Only very slowly does it relax to alignment with the direction of the ambient field.

Although this is a little surprising, it becomes easier to understand when we consider that the field in circuit 1 had energy. Therefore the field can't just go poof. It has to have some mechanism for transforming that energy into some other form. The only mechanism available for doing that is resistive heating in the coil. But this will take some time, as measured by the RL time constant of circuit 2. During this time, the field and the current just gradually die out.

To push this current through the circuit, which has some resistance, we will need an electric field, even though the battery has been taken out of action. This electric field exists due to Faraday's law, and the minus sign in Faraday's law is interpreted as saying that the field is in the direction that tends to resist the change in the magnetic field.

Frames of reference

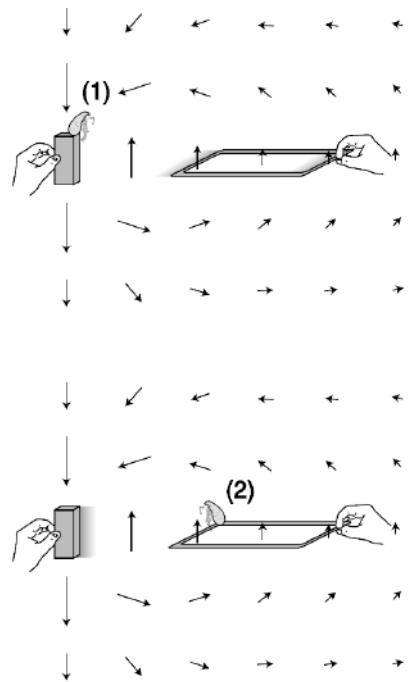
example 6

In figure i, flea 1 doesn't believe in this modern foolishness about induction. She's sitting on the bar magnet, which to her is obviously at rest. As the square wire loop is dragged away from her and the magnet, its protons experience a force out of the page, because the cross product $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$ is out of the page. The electrons, which are negatively charged, feel a force into the

page. The conduction electrons are free to move, but the protons aren't. In the front and back sides of the loop, this force is perpendicular to the wire. In the right and left sides, however, the electrons are free to respond to the force. Note that the magnetic field is weaker on the right side. It's as though we had two pumps in a loop of pipe, with the weaker pump trying to push in the opposite direction; the weaker pump loses the argument.¹ We get a current that circulates around the loop.² There is no induction going on in this frame of reference; the forces that cause the current are just the ordinary magnetic forces experienced by any charged particle moving through a magnetic field.

Flea 2 is sitting on the loop, which she considers to be at rest. In her frame of reference, it's the bar magnet that is moving. Like flea 1, she observes a current circulating around the loop, but unlike flea 1, she cannot use magnetic forces to explain this current. As far as she is concerned, the electrons were initially at rest. Magnetic forces are forces between moving charges and other moving charges, so a magnetic field can never accelerate a charged particle starting from rest. A force that accelerates a charge from rest can only be an *electric* force, so she is forced to conclude that there is an electric field in her region of space. This field drives electrons around and around in circles, so it is apparently violating the loop rule — it is a curly field. What reason can flea 2 offer for the existence of this electric field pattern? Well, she's been noticing that the magnetic field in her region of space has been changing, possibly because that bar magnet over there has been getting farther away. She observes that a changing magnetic field creates a curly electric field.

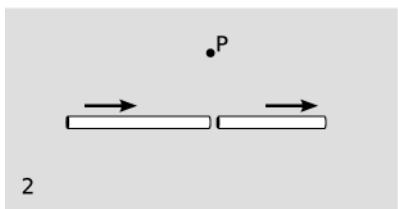
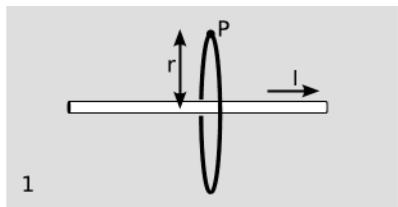
We therefore conclude that induction effects *must* exist based on the fact that motion is relative. If we didn't want to admit induction effects, we would have to outlaw flea 2's frame of reference, but the whole idea of relative motion is that all frames of reference are created equal, and there is no way to determine which one is really at rest.



i / A generator that works with linear motion.

¹If the pump analogy makes you uneasy, consider what would happen if all the electrons moved into the page on both sides of the loop. We'd end up with a net negative charge at the back side, and a net positive charge on the front. This actually would happen in the first nanosecond after the loop was set in motion. This buildup of charge would start to quench both currents due to electrical forces, but the current in the right side of the wire, which is driven by the weaker magnetic field, would be the first to stop. Eventually, an equilibrium will be reached in which the same amount of current is flowing at every point around the loop, and no more charge is being piled up.

²The wire is not a perfect conductor, so this current produces heat. The energy required to produce this heat comes from the hands, which are doing mechanical work as they separate the magnet from the loop.



Discussion question

- A** 1. The figure shows a line of charges moving to the right, creating a current I . An Ampèrean surface in the form of a disk has been superimposed. Use Maxwell's equations to find the field B at point P .
2. A tiny gap is chopped out of the line of charge. What happens when this gap is directly underneath the point P ?

Discussion question A.

Problems

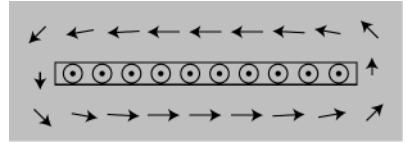
Key

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

1 The equation $B = (k/c^2) \cdot 2I/r$ for the magnetic field of a long, straight wire was derived in examples 2, p. 121, (the form of the equation) and 6, p. 181 (the factor of 2). Derive the equation using Ampère's law. (Cf. also problem 9, p. 282, using the Biot-Savart law.)

2 The figure shows a sheet of current coming out of the page. Such a sheet can be characterized by a linear current density η , which has units of A/m . (The letter η is Greek eta, which makes the "ee" sound in modern Greek) The figure shows magnetic field vectors in the $\pm x$ directions, with equal magnitudes above and below the sheet. In general, however, this symmetry need not exist. We could always add a constant magnetic field to the whole field pattern and get another equally valid solution of Maxwell's equations. The only thing we can actually determine is ΔB_x , the difference in the horizontal field between the top and bottom of the sheet.

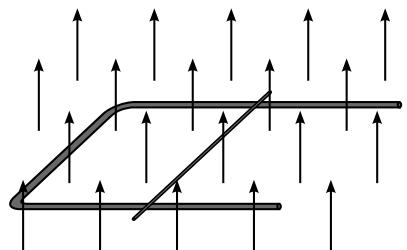
- (a) Find ΔB_x in terms of η . ✓
(b) In the symmetric case shown in the figure, what can you say about the pressure or tension experienced by the sheet, using the visual modes of reasoning from sec. 5.2.2, p. 126? If the sheet has no structural strength, what will it do?



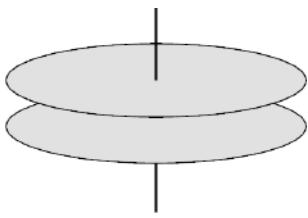
Problem 2.

3 A U-shaped wire makes electrical contact with a second, straight wire, of length ℓ , which rolls along it to the right at velocity v , as shown in the figure. The whole thing is immersed in a uniform magnetic field B , which is perpendicular to the plane of the circuit. The resistance R of the rolling wire is much greater than that of the U.

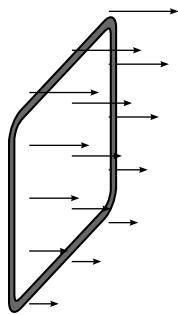
- (a) Use Faraday's law to find the amount of current through the wire, and its direction. ✓
(b) Use conservation of energy to find the direction of the force on the wire.
(c) Verify the direction of the force using right-hand rules.
(d) Find the magnitude of the force acting on the wire. ✓
(e) Consider how the answer to part a would have changed if the direction of the field had been reversed, and also do the case where the direction of the rolling wire's motion is reversed. Verify that this is in agreement with your answer to part b.



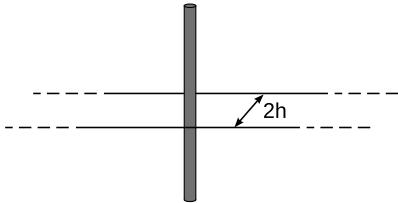
Problem 3.



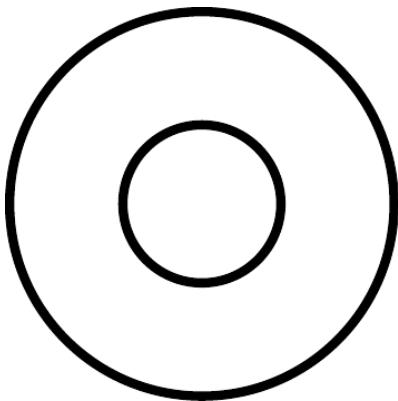
Problem 4.



Problem 5.



Problem 6.



Problem 7.

4 The circular parallel-plate capacitor shown in the figure is being charged up over time, with the voltage difference across the plates varying as $V = st$, where s is a constant. The plates have radius b , and the distance between them is d . We assume $d \ll b$, so that the electric field between the plates is uniform, and parallel to the axis. Find the induced magnetic field at a point between the plates, at a distance R from the axis. \triangleright Hint, p. 425 ✓

Problems 5-6 require enough knowledge of vector calculus to evaluate line and surface integrals with integrands that aren't constant.

5 A wire loop of resistance R and area A , lying in the $y - z$ plane, falls through a nonuniform magnetic field $\mathbf{B} = kz\hat{\mathbf{x}}$, where k is a constant. The z axis is vertical.

- (a) Find the direction of the force on the wire based on conservation of energy.
- (b) Verify the direction of the force using right-hand rules.
- (c) Find the magnetic force on the wire. ✓

6 Verify Ampère's law in the case shown in the figure, assuming the known equation for the field of a wire. A wire carrying current I passes perpendicularly through the center of the rectangular Ampèrean surface. The length of the rectangle is infinite, so it's not necessary to compute the contributions of the ends.

7 A certain electrical transmission line, shown in cross-section, consists of two hollow, coaxial pipes. Let r be the distance from the axis. The inner pipe is at $r = a$ and the outer at $r = b$. The inner pipe carries current I , and the outer $-I$, i.e., the transmission line is part of a complete circuit, and the current flows out through one conductor and back through the other. Find the magnitude of the magnetic field (a) for $r < a$, (b) for $a < r < b$, and (c) for $r > b$.

8 A cylindrical wire carries a current that is not uniformly distributed across its cross-section. The current density is a function $j(r)$ of the distance from the axis. Show that if the magnetic field is known as a function $B(r)$, then the current density can be determined.

Electromagnetic properties of materials

Chapter 16

Electromagnetic properties of materials

Different types of matter have a variety of useful electrical and magnetic properties. Some are conductors, and some are insulators. Some, like iron and nickel, can be magnetized, while others have useful electrical properties, e.g., dielectrics, discussed qualitatively in the discussion question on page 313, which allow us to make capacitors with much higher values of capacitance than would otherwise be possible. We need to organize our knowledge about the properties that materials can possess, and see whether this knowledge allows us to calculate anything useful with Maxwell's equations.

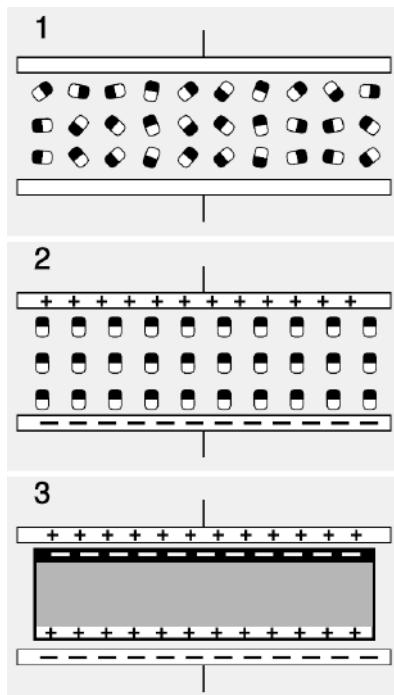
16.1 Conductors

A perfect conductor, such as a superconductor, has no DC electrical resistance. It is not possible to have a static electric field inside it, because then charges would move in response to that field, and the motion of the charges would tend to reduce the field, contrary to the assumption that the field was static. Things are a little different at the surface of a perfect conductor than on the interior. We expect that any net charges that exist on the conductor will spread out under the influence of their mutual repulsion, and settle on the surface. Gauss's law requires that the fields on the two sides of a sheet of charge have $|\mathbf{E}_{\perp,1} - \mathbf{E}_{\perp,2}|$ proportional to the surface charge density, and since the field inside the conductor is zero, we infer that there can be a field on or immediately outside the conductor, with a nonvanishing component perpendicular to the surface. The component of the field parallel to the surface must vanish, however, since otherwise it would cause the charges to move along the surface.

On a hot summer day, the reason the sun feels warm on your skin is that the oscillating fields of the light waves excite currents in your skin, and these currents dissipate energy by ohmic heating. In a perfect conductor, however, this could never happen, because there is no such thing as ohmic heating. Since electric fields can't penetrate a perfect conductor, we also know that an electromagnetic wave can never pass into one. By conservation of energy, we know that the wave can't just vanish, and if the energy can't be dissipated as heat, then the only remaining possibility is that all of the wave's energy is reflected. This is why metals, which are good

electrical conductors, are also highly reflective. They are not *perfect* electrical conductors, however, so they are not perfectly reflective. The wave enters the conductor, but immediately excites oscillating currents, and these oscillating currents dissipate the energy both by ohmic heating and by reradiating the reflected wave. Since the parts of Maxwell's equations describing radiation have time derivatives in them, the efficiency of this reradiation process depends strongly on frequency. When the frequency is high and the material is a good conductor, reflection predominates, and is so efficient that the wave only penetrates to a very small depth, called the skin depth. In the limit of poor conduction and low frequencies, absorption predominates, and the skin depth becomes much greater. In a high-frequency AC circuit, the skin depth in a copper wire is very small, and therefore the signals in such a circuit are propagated entirely at the surfaces of the wires. In the limit of low frequencies, i.e., DC, the skin depth approaches infinity, so currents are carried uniformly over the wires' cross-sections.

We can quantify how well a particular material conducts electricity. We know that the resistance of a wire is proportional to its length, and inversely proportional to its cross-sectional area. The constant of proportionality is $1/\sigma$, where σ (not the same σ as the surface charge density) is called the electrical conductivity. Exposed to an electric field \mathbf{E} , a conductor responds with a current per unit cross-sectional area $\mathbf{J} = \sigma\mathbf{E}$. The skin depth is proportional to $1/\sqrt{f\sigma}$, where f is the frequency of the wave.



a / A capacitor with a dielectric between the plates.

16.2 Dielectrics

A material with a very low conductivity is an insulator. Such materials are usually composed of atoms or molecules whose electrons are strongly bound to them; since the atoms or molecules have zero total charge, their motion cannot create an electric current. But even though they have zero charge, they may not have zero dipole moment. Imagine such a substance filling in the space between the plates of a capacitor, as in figure a. For simplicity, we assume that the molecules are oriented randomly at first, a/1, and then become completely aligned when a field is applied, a/2. The effect has been to take all of the negatively charged black ends of the molecules and shift them upward, and the opposite for the positively charged white ends. Where the black and white charges overlap, there is still zero net charge, but we have a strip of negative charge at the top, and a strip of positive charge at the bottom, a/3. The effect has been to cancel out part of the charge that was deposited on the plates of the capacitor. Now this is very subtle, because Maxwell's equations treat these charges on an equal basis, but in terms of practical measurements, they are completely different. The charge on the plates can be measured by inserting an ammeter in the circuit, and inte-

grating the current over time. But the charges in the layers at the top and bottom of the dielectric never flowed through any wires, and cannot be detected by an ammeter. In other words, the total charge, q , appearing in Maxwell's equations is actually $q = q_{\text{free}} - q_{\text{bound}}$, where q_{free} is the charge that moves freely through wires, and can be detected in an ammeter, while q_{bound} is the charge bound onto the individual molecules, which can't. We will, however, detect the presence of the bound charges via their electric fields. Since their electric fields partially cancel the fields of the free charges, a voltmeter will register a smaller than expected voltage difference between the plates. If we measure q_{free}/V , we have a result that is larger than the capacitance we would have expected.

Although the relationship $\mathbf{E} \leftrightarrow q$ between electric fields and their sources is unalterably locked in by Gauss's law, that's not what we see in practical measurements. In this example, we can measure the voltage difference between the plates of the capacitor and divide by the distance between them to find \mathbf{E} , and then integrate an ammeter reading to find q_{free} , and we will find that Gauss's law appears not to hold. We have $\mathbf{E} \leftrightarrow q_{\text{free}}/(\text{constant})$, where the constant fudge factor is greater than one. This constant is a property of the dielectric material, and tells us how many dipoles there are, how strong they are, and how easily they can be reoriented. The conventional notation is to incorporate this fudge factor into Gauss's law by defining an altered version of the electric field,

$$\mathbf{D} = \epsilon \mathbf{E},$$

and to rewrite Gauss's law as

$$\Phi_D = q_{\text{in, free}}.$$

The constant ϵ is a property of the material, known as its permittivity. In a vacuum, ϵ takes on a value known as ϵ_0 , defined as $1/(4\pi k)$. In a dielectric, ϵ is greater than ϵ_0 . When a dielectric is present between the plates of a capacitor, its capacitance is proportional to ϵ . The following table gives some sample values of the permittivities of a few substances.

substance	ϵ/ϵ_0 at zero frequency
vacuum	1
air	1.00054
water	80
barium titanate	1250

A capacitor with a very high capacitance is potentially a superior replacement for a battery, but until the 1990's this was impractical because capacitors with high enough values couldn't be made, even with dielectrics having the largest known permittivities. Such supercapacitors, some with values in the kilofarad range, are now available. Most of them do not use dielectric at all; the very high



b / A stud finder is used to locate the wooden beams, or studs, that form the frame behind the wallboard. It is a capacitor whose capacitance changes when it is brought close to a substance with a particular permittivity. Although the wall is external to the capacitor, a change in capacitance is still observed, because the capacitor has “fringing fields” that extend outside the region between its plates.

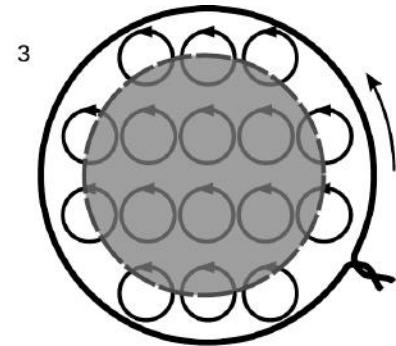
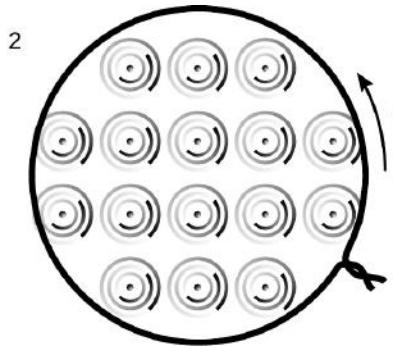
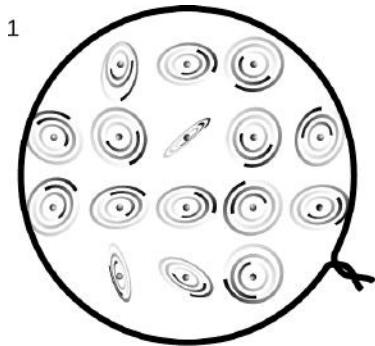
capacitance values are instead obtained by using electrodes that are not parallel metal plates at all, but exotic materials such as aerogels, which allows the spacing between the “electrodes” to be very small.

Although figure a/2 shows the dipoles in the dielectric being completely aligned, this is not a situation commonly encountered in practice. In such a situation, the material would be as polarized as it could possibly be, and if the field was increased further, it would not respond. In reality, a capacitor, for example, would normally be operated with fields that produced quite a small amount of alignment, and it would be under these conditions that the linear relationship $\mathbf{D} = \epsilon \mathbf{E}$ would actually be a good approximation. Before a material’s maximum polarization is reached, it may actually spark or burn up.

self-check A

Suppose a parallel-plate capacitor is built so that a slab of dielectric material can be slid in or out. (This is similar to the way the stud finder in figure b works.) We insert the dielectric, hook the capacitor up to a battery to charge it, and then use an ammeter and a voltmeter to observe what happens when the dielectric is withdrawn. Predict the changes observed on the meters, and correlate them with the expected change in capacitance. Discuss the energy transformations involved, and determine whether positive or negative work is done in removing the dielectric.

▷ Answer, p. 433



c / The magnetic version of figure a. A magnetically permeable material is placed at the center of a solenoid.

16.3 Magnetic materials

16.3.1 Magnetic permeability

Atoms and molecules may have magnetic dipole moments as well as electric dipole moments. Just as an electric dipole contains bound charges, a magnetic dipole has bound currents, which come from the motion of the electrons as they orbit the nucleus, c/1. Such a substance, subjected to a magnetic field, tends to align itself, c/2,

so that a sheet of current circulates around the externally applied field. Figure c/3 is closely analogous to figure a/3; in the central gray area, the atomic currents cancel out, but the atoms at the outer surface form a sheet of bound current. However, whereas like charges repel and opposite charges attract, it works the other way around for currents: currents in the same direction attract, and currents in opposite directions repel. Therefore the bound currents in a material inserted inside a solenoid tend to *reinforce* the free currents, and the result is to strengthen the field. The total current is $I = I_{\text{free}} + I_{\text{bound}}$, and we define an altered version of the magnetic field,

$$\mathbf{H} = \frac{\mathbf{B}}{\mu},$$

and rewrite Ampère's law as

$$\Gamma_H = I_{\text{through, free}}.$$

The constant μ is the permeability, with a vacuum value of $\mu_0 = 4\pi k/c^2$. Here are the magnetic permeabilities of some substances:

substance	μ/μ_0
vacuum	1
aluminum	1.00002
steel	700
transformer iron	4,000
mu-metal	20,000

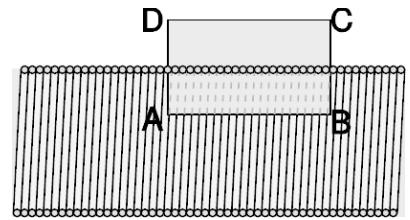
An iron-core electromagnet

example 1

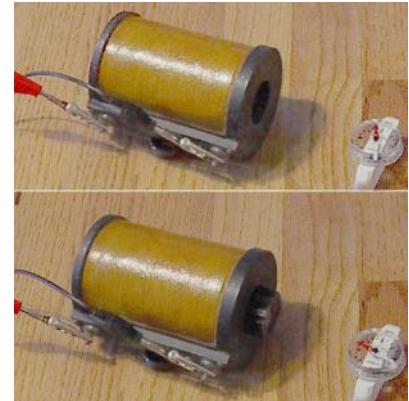
- ▷ A solenoid has 1000 turns of wire wound along a cylindrical core with a length of 10 cm. If a current of 1.0 A is used, find the magnetic field inside the solenoid if the core is air, and if the core is made of iron with $\mu/\mu_0 = 4,000$.
- ▷ Air has essentially the same permeability as vacuum, so using the result of example 4 on page 350, we find that the field is 0.013 T.

We now consider the case where the core is filled with iron. The original derivation in example 4 started from Ampère's law, which we now rewrite as $\Gamma_H = I_{\text{through, free}}$. As argued previously, the only significant contributions to the circulation come from line segment AB. This segment lies inside the iron, where $\mathbf{H} = \mathbf{B}/\mu$. The \mathbf{H} field is the same as in the air-core case, since the new form of Ampère's law only relates \mathbf{H} to the current in the wires (the free current). This means that $\mathbf{B} = \mu\mathbf{H}$ is greater by a factor of 4,000 than in the air-core case, or 52 T. This is an extremely intense field — so intense, in fact, that the iron's magnetic polarization would probably become saturated before we could actually get the field that high.

The electromagnet of example 1 could also be used as an inductor, and its inductance would be proportional to the permeability



d / Example 1: a cutaway view of a solenoid.



e / Example 1: without the iron core, the field is so weak that it barely deflects the compass. With it, the deflection is nearly 90°.

of the core. This makes it possible to construct high-value inductors that are relatively compact. Permeable cores are also used in transformers.

A transformer or inductor with a permeable core does have some disadvantages, however, in certain applications. The oscillating magnetic field induces an electric field, and because the core is typically a metal, these currents dissipate energy strongly as heat. This behaves like a fairly large resistance in series with the coil. Figure f shows a method for reducing this effect. The iron core of this transformer has been constructed out of laminated layers, which has the effect of blocking the conduction of the eddy currents.



f / A transformer with a laminated iron core. The input and output coils are inside the paper wrapper. The iron core is the black part that passes through the coils at the center, and also wraps around them on the outside.

A ferrite bead

example 2

Cables designed to carry audio signals are typically made with two adjacent conductors, such that the current flowing out through one conductor comes back through the other one. Computer cables are similar, but usually have several such pairs bundled inside the insulator. This paired arrangement is known as differential mode, and has the advantage of cutting down on the reception and transmission of interference. In terms of transmission, the magnetic field created by the outgoing current is almost exactly canceled by the field from the return current, so electromagnetic waves are only weakly induced. In reception, both conductors are bathed in the same electric and magnetic fields, so an emf that adds current on one side subtracts current from the other side, resulting in cancellation.

The opposite of differential mode is called common mode. In common mode, all conductors have currents flowing in the same direction. Even when a circuit is designed to operate in differential mode, it may not have exactly equal currents in the two conductors with $I_1 + I_2 = 0$, meaning that current is leaking off to ground at one end of the circuit or the other. Although paired cables are relatively immune to differential-mode interference, they do not have any automatic protection from common-mode interference.



g / Example 2: ferrite beads. The top panel shows a clip-on type, while the bottom shows one built into a cable.

Figure g shows a device for reducing common-mode interference called a ferrite bead, which surrounds the cable like a bead on a string. Ferrite is a magnetically permeable alloy. In this application, the ohmic properties of the ferrite actually turn out to be advantageous.

Let's consider common-mode transmission of interference. The bare cable has some DC resistance, but is also surrounded by a magnetic field, so it has inductance as well. This means that it behaves like a series L-R circuit, with an impedance that varies as $R + i\omega L$, where both R and L are very small. When we add the ferrite bead, the inductance is increased by orders of magnitude, but so is the resistance. Neither R nor L is actually constant with respect to frequency, but both are much greater than for the bare

cable.

Suppose, for example, that a signal is being transmitted from a digital camera to a computer via a USB cable. The camera has an internal impedance that is on the order of 10Ω , the computer's input also has a $\sim 10 \Omega$ impedance, and in differential mode the ferrite bead has no effect, so the cable's impedance has its low, designed value (probably also about 10Ω , for good impedance matching). The signal is transmitted unattenuated from the camera to the computer, and there is almost no radiation from the cable.

But in reality there will be a certain amount of common-mode current as well. With respect to common mode, the ferrite bead has a large impedance, with the exact value depending on frequency, but typically on the order of 100Ω for frequencies in the MHz range. We now have a series circuit consisting of three impedances: 10, 100, and 10Ω . For a given emf applied by an external radio wave, the current induced in the circuit has been attenuated by an order of magnitude, relative to its value without the ferrite bead.

Why is the ferrite necessary at all? Why not just insert ordinary air-core inductors in the circuit? We could, for example, have two solenoidal coils, one in the outgoing line and one in the return line, interwound with one another with their windings oriented so that their differential-mode fields would cancel. There are two good reasons to prefer the ferrite bead design. One is that it allows a clip-on device like the one in the top panel of figure g, which can be added without breaking the circuit. The other is that our circuit will inevitably have some stray capacitance, and will therefore act like an LRC circuit, with a resonance at some frequency. At frequencies close to the resonant frequency, the circuit would absorb and transmit common-mode interference very strongly, which is exactly the opposite of the effect we were hoping to produce. The resonance peak could be made low and broad by adding resistance in series, but this extra resistance would attenuate the differential-mode signals as well as the common-mode ones. The ferrite's resistance, however, is actually a purely magnetic effect, so it vanishes in differential mode.

Surprisingly, some materials have magnetic permeabilities less than μ_0 . This cannot be accounted for in the model above, and although there are semiclassical arguments that can explain it to some extent, it is fundamentally a quantum mechanical effect. Materials with $\mu > \mu_0$ are called paramagnetic, while those with $\mu < \mu_0$ are referred to as diamagnetic. Diamagnetism is generally a much weaker effect than paramagnetism, and is easily masked if there is any trace of contamination from a paramagnetic material. Diamagnetic materials have the interesting property that they are repelled



h / A frog is levitated diamagnetically by the nonuniform field inside a powerful magnet. Evidently frog has $\mu < \mu_0$.

$$\mathbf{H} = \frac{\mathbf{B}}{\mu}$$

$$\mathbf{D} = \epsilon \mathbf{E}$$

$$\mu_0 = \frac{4\pi k}{c^2}$$

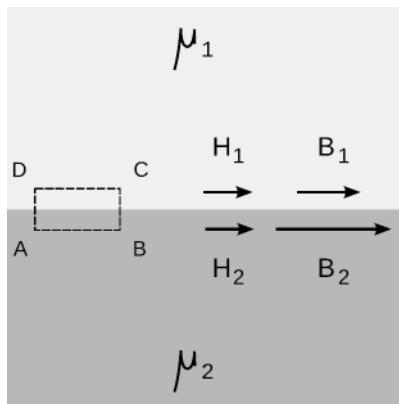
$$\epsilon_0 = \frac{1}{4\pi k}$$

$$\operatorname{div} \mathbf{D} = \rho$$

$$\operatorname{div} \mathbf{B} = 0$$

$$\operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\operatorname{curl} \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{j}$$



i / At a boundary between two substances with $\mu_2 > \mu_1$, the \mathbf{H} field has a continuous component parallel to the surface, which implies a discontinuity in the parallel component of the magnetic field \mathbf{B} .

from regions of strong magnetic field, and it is therefore possible to levitate a diamagnetic object above a magnet, as in figure h.

A complete statement of Maxwell's equations in the presence of electric and magnetic materials is as follows:

Suppose we have a boundary between two substances. By constructing a Gaussian or Ampèrian surface that extends across the boundary, we can arrive at various constraints on how the fields must behave as we move from one substance into the other, when there are no free currents or charges present, and the fields are static. An interesting example is the application of Faraday's law, $\Gamma_H = 0$, to the case where one medium — let's say it's air — has a low permeability, while the other one has a very high one. We will violate Faraday's law unless the component of the \mathbf{H} field parallel to the boundary is a continuous function, $\mathbf{H}_{\parallel,1} = \mathbf{H}_{\parallel,2}$. This means that if μ/μ_0 is very high, the component of $\mathbf{B} = \mu \mathbf{H}$ parallel to the surface will have an abrupt discontinuity, being much stronger inside the high-permeability material. The result is that when a magnetic field enters a high-permeability material, it tends to twist abruptly to one side, and the pattern of the field tends to be channeled through the material like water through a funnel. In a transformer, a permeable core functions to channel more of the magnetic flux from the input coil to the output coil. Figure j shows another example, in which the effect is to shield the interior of the sphere from the externally imposed field. Special high-permeability alloys, with trade names like Mu-Metal, are sold for this purpose.

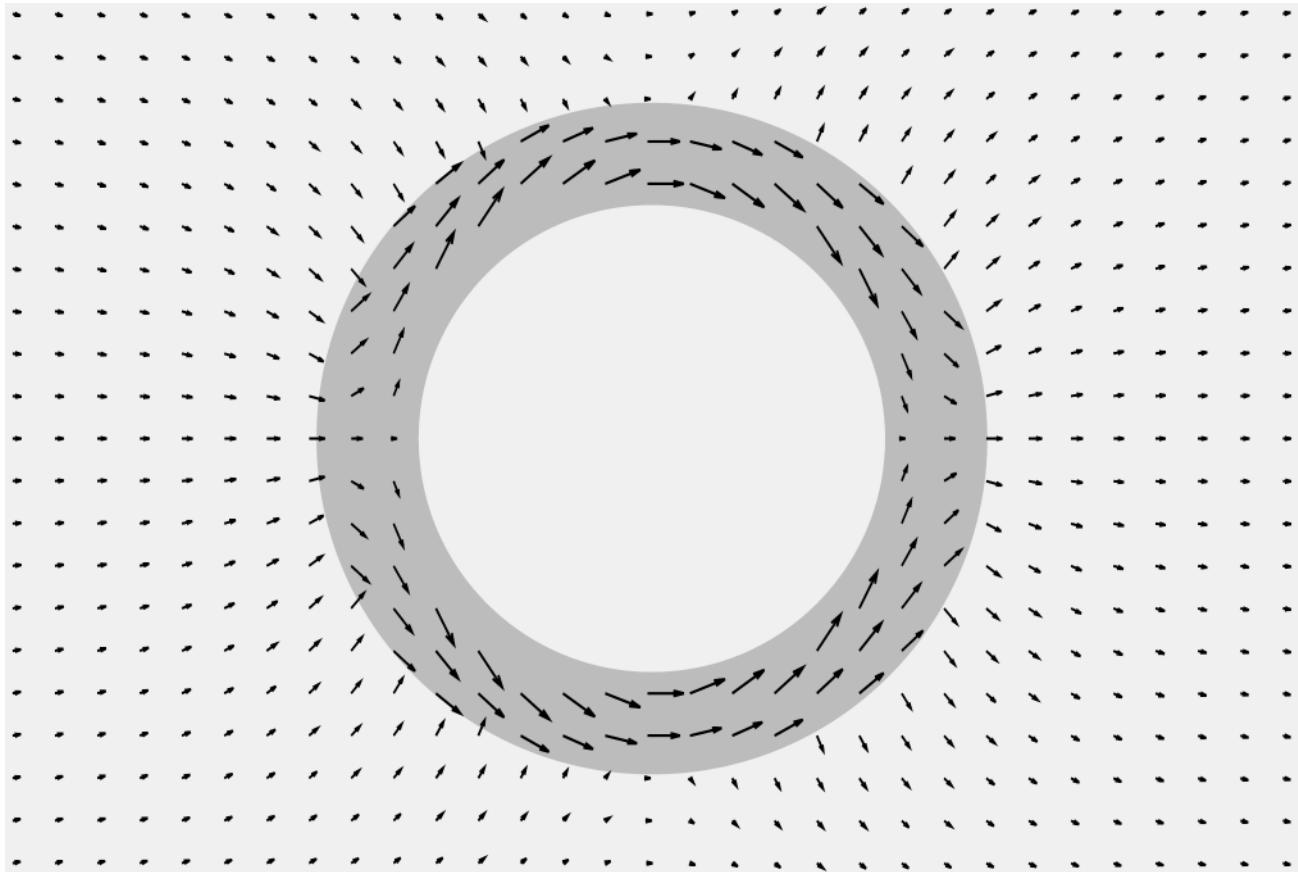
Variables that are continuous at a boundary

$$\begin{array}{ll} \mathbf{E}_\parallel & \mathbf{D}_\perp \\ \mathbf{H}_\parallel & \mathbf{B}_\perp \end{array}$$

16.3.2 Ferromagnetism

The very last magnetic phenomenon we'll discuss is probably the very first experience you ever had of magnetism. Ferromagnetism is a phenomenon in which a material tends to organize itself so that it has a nonvanishing magnetic field. It is exhibited strongly by iron and nickel, which explains the origin of the name.

Figure k/1 is a simple one-dimensional model of ferromagnetism. Each magnetic compass needle represents an atom. The compasses in the chain are stable when aligned with one another, because each

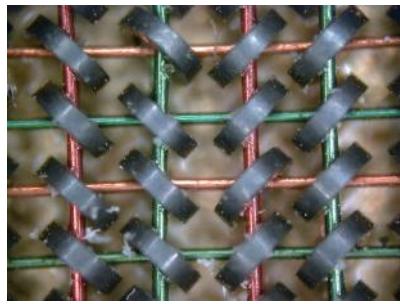


j / A hollow sphere with $\mu/\mu_0 = 10$, is immersed in a uniform, externally imposed magnetic field. The interior of the sphere is shielded from the field. The arrows map the magnetic field \mathbf{B} . (See homework problem nn, page nnn.)

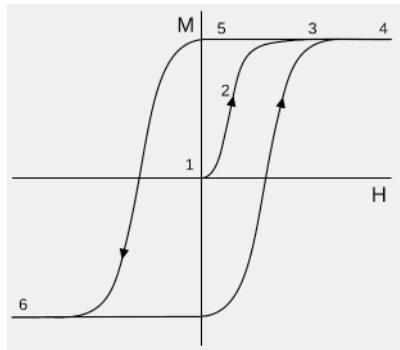


k / A model of ferromagnetism.

one's north end is attracted to its neighbor's south end. The chain can be turned around, k/2, without disrupting its organization, and the compasses do not realign themselves with the Earth's field, because their torques on one another are stronger than the Earth's torques on them. The system has a memory. For example, if I want to remind myself that my friend's address is 137 Coupling Ct., I can align the chain at an angle of 137 degrees. The model fails, however, as an explanation of real ferromagnetism, because in two or more dimensions, the most stable arrangement of a set of interacting



I / Magnetic core memory.



m / A hysteresis curve.

magnetic dipoles is something more like $k/3$, in which alternating rows point in opposite directions. In this two-dimensional pattern, every compass is aligned in the most stable way with all four of its neighbors. This shows that ferromagnetism, like diamagnetism, has no purely classical explanation; a full explanation requires quantum mechanics.

Because ferromagnetic substances “remember” the history of how they were prepared, they are commonly used to store information in computers. Figure I shows 16 bits from an ancient (ca. 1970) 4-kilobyte random-access memory, in which each doughnut-shaped iron “core” can be magnetized in one of two possible directions, so that it stores one bit of information. Today, RAM is made of transistors rather than magnetic cores, but a remnant of the old technology remains in the term “core dump,” meaning “memory dump,” as in “my girlfriend gave me a total core dump about her mom’s divorce.” Most computer hard drives today do store their information on rotating magnetic platters, but the platter technology may be obsoleted by flash memory in the near future.

The memory property of ferromagnets can be depicted on the type of graph shown in figure m, known as a hysteresis curve. The y axis is the magnetization of a sample of the material — a measure of the extent to which its atomic dipoles are aligned with one another. If the sample is initially unmagnetized, 1, and a field H is externally applied, the magnetization increases, 2, but eventually becomes saturated, 3, so that higher fields do not result in any further magnetization, 4. The external field can then be reduced, 5, and even eliminated completely, but the material will retain its magnetization. It is a permanent magnet. To eliminate its magnetization completely, a substantial field must be applied in the opposite direction. If this reversed field is made stronger, then the substance will eventually become magnetized just as strongly in the opposite direction. Since the hysteresis curve is nonlinear, and is not a function (it has more than one value of M for a particular value of B), a ferromagnetic material does not have a single, well-defined value of the permeability μ ; a value like 4,000 for transformer iron represents some kind of a rough average.

The fluxgate compass

example 3

The fluxgate compass is a type of magnetic compass without moving parts, commonly used on ships and aircraft. An AC current is applied in a coil wound around a ferromagnetic core, driving the core repeatedly around a hysteresis loop. Because the hysteresis curve is highly nonlinear, the addition of an external field such as the Earth’s alters the core’s behavior. Suppose, for example, that the axis of the coil is aligned with the magnetic north-south. The core will reach saturation more quickly when the coil’s field is in the same direction as the Earth’s, but will not saturate as early in the next half-cycle, when the two fields are

in opposite directions. With the use of multiple coils, the components of the Earth's field can be measured along two or three axes, permitting the compass's orientation to be determined in two or (for aircraft) three dimensions.

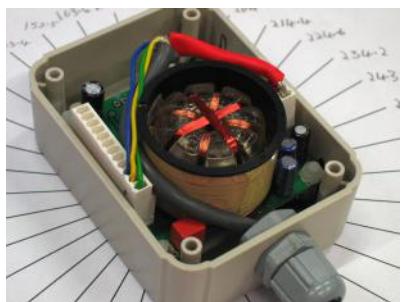
Sharp magnet poles

example 4

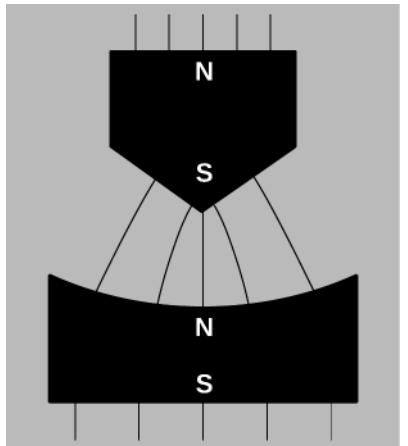
Although a ferromagnetic material does not really have a single value of the magnetic permeability, there is still a strong tendency to have $\mathbf{B}_{\parallel} \approx 0$ just outside the magnet's surface, for the same reasons as discussed above for high-permeability substances in general. For example, if we have a cylindrical bar magnet about the size and shape of your finger, magnetized lengthwise, then the field near the ends is nearly perpendicular to the surfaces, while the field near the sides, although it may be oriented nearly parallel to the surface, is very weak, so that we still have $\mathbf{B}_{\parallel} \approx 0$. This is in close analogy to the situation for the *electric* field near the surface of a conductor in equilibrium, for which $\mathbf{E}_{\parallel} = 0$. This analogy is close enough so that we can recycle much of our knowledge about electrostatics.

For example, we saw in example 2, p. 93, and problem 17, p. 108, that charge tends to collect on the most highly curved portions of a conductor, and therefore becomes especially dense near a corner or knife-edge. This gives us a way of making especially intense magnetic fields. Most people would imagine that a very intense field could be made simply by using a very large and bulky permanent magnet, but this doesn't actually work very well, because magnetic dipole fields fall off as $1/r^3$, so that at a point near the surface, nearly all the field is contributed by atoms near the surface. Our analogy with electrostatics suggests that we should instead construct a permanent magnet with a sharp edge.

Figure o shows the cross-sectional shapes of two magnet poles used in the historic Stern-Gerlach experiment that discovered the spin of the electron. The external magnetic field is represented using field lines. The field lines enter and exit the surfaces perpendicularly, and they are particularly dense near the corner of the upper pole, indicating a strong field. The spreading of the field lines indicates that the field is strongly nonuniform, becoming much weaker toward the bottom of the gap between the poles. This strong nonuniformity was crucial for the experiment, in which the magnets were used as part of a dipole spectrometer. See figure u on p. 136 for an electric version of such a spectrometer.



n / A fluxgate compass.



o / Example 4.

16.4 Electromagnetic waves in matter

In example 8, p. 302, we gave an explanation of why electromagnetic waves traveling through matter are *dispersive* (sec. 6.4), i.e., their speed depends on their frequency. The concept is that when an

electromagnetic wave enters a material such as a glass windowpane, charges inside the material oscillate in response to the wave. The charges have a resonant frequency (or, in real-world materials, several different resonant frequencies), so the amplitude and phase of their oscillation depends on the frequency of the driving wave. The oscillating charges in turn re-emit their own wave, which superposes on top of the original wave. The superposition may either lead or lag behind the original wave, so its crests arrive early or late. The effect is identical to what we would expect if the speed of the wave had some other value than c .

In that explanation, we focused on the phases and the trend of the effect with changing frequency, ignoring real constants. Fortunately, it turns out that the effect of all those ignored constants can be summarized in a simple way in terms of the bulk properties of the material. If we compare Maxwell's equations in matter with their vacuum version, we see that the speed of an electromagnetic wave moving through a substance described by permittivity and permeability ϵ and μ is $1/\sqrt{\epsilon\mu}$.

For most substances, we observe that ϵ is highly frequency-dependent, and this is well explained by our earlier analysis in terms of the excitation of oscillating charges, where the behavior changes dramatically as the frequency passes through a resonant frequency.

The possibility of $\mu \neq \mu_0$ corresponds microscopically to a picture in which dipoles flip their orientation back and forth in response to the wave. For most substances we find that this isn't a significant effect, and $\mu \approx \mu_0$.

Color interference

example 5

The colorful image on the cover of this book was created by taking two polarizing films, as in example 1, p. 157, and minilab 6, p. 173, and placing between them a calcium sulfate crystal. At the atomic level, the crystal is a lattice of atoms, and the lattice is asymmetric, so that it has two distinguishable axes. Let's call these axes x and y , and describe the polarization of electromagnetic waves by the direction of the electric fields. If a wave has a polarization in the x direction, the permittivity has some value ϵ_x , but a wave with its polarization in the y direction experiences some other ϵ_y . Both of these depend on frequency.

Let's analyze the simplest possible example that elucidates the physics. Suppose that we orient the first polarizing film at a 45° angle with respect to the axes, so that the light entering the crystal is (ignoring unitful constant factors)

$$\mathbf{E}_1 = \hat{x} + \hat{y}.$$

As the wave emerges from the crystal, the difference in velocity for the two components will put them out of phase. Let's say that the result is to reverse the phase of the y component compared

to the x ,

$$\mathbf{E}_2 = \hat{\mathbf{x}} - \hat{\mathbf{y}}.$$

This is a 90-degree rotation compared to \mathbf{E}_1 . Now suppose that the second polarizing film is oriented in the same direction as the first film. Then although \mathbf{E}_1 was in exactly the right direction to pass through, the direction of \mathbf{E}_2 is precisely wrong. The light is completely blocked.

But this phase relationship depends not just on the thickness of the crystal and the values of ϵ_x and ϵ_y , but also on frequency. For some other frequency of light, the phase of the x and y polarizations in \mathbf{E}_2 could end up the same as in \mathbf{E}_1 , in which case the light would be entirely transmitted through the second filter. Of course all of the intermediate possibilities occur as well. For this reason, some colors are more strongly transmitted through this setup and some more strongly absorbed.

Problems

Key

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

Relativity

Chapter 17

★Relativity (optional stand-alone chapter)

This optional chapter is a stand-alone presentation of special relativity. It can be read before, during, or after the rest of the book.

17.1 Time is not absolute

When Einstein first began to develop the theory of relativity, around 1905, the only real-world observations he could draw on were ambiguous and indirect. Today, the evidence is part of everyday life. For example, every time you use a GPS receiver, a, you're using Einstein's theory of relativity. Somewhere between 1905 and today, technology became good enough to allow conceptually *simple* experiments that students in the early 20th century could only discuss in terms like "Imagine that we could..." A good jumping-on point is 1971. In that year, J.C. Hafele and R.E. Keating brought atomic clocks aboard commercial airliners, b, and went around the world, once from east to west and once from west to east. Hafele and Keating observed that there was a discrepancy between the times measured by the traveling clocks and the times measured by similar clocks that stayed home at the U.S. Naval Observatory in Washington. The east-going clock lost time, ending up off by -59 ± 10 nanoseconds, while the west-going one gained 273 ± 7 ns.

17.1.1 The correspondence principle

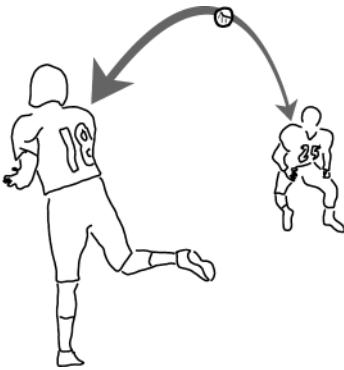
This establishes that time doesn't work the way Newton believed it did when he wrote that "Absolute, true, and mathematical time, of itself, and from its own nature flows equably without regard to anything external..." We are used to thinking of time as absolute and universal, so it is disturbing to find that it can flow at a different rate for observers in different frames of reference. Nevertheless, the effects that Hafele and Keating observed were small. This makes sense: 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.



a / 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 rate aboard a GPS satellite than it does on the bike, and the GPS software has to take this into account.



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



c / Newton's laws do not distinguish past from future. The football could travel in either direction while obeying Newton's laws.

17.1.2 Causality

It's also reassuring that the effects on time were small compared to the three-day lengths of the plane trips. There was therefore no opportunity for paradoxical scenarios such as one in which the east-going experimenter arrived back in Washington before he left and then convinced himself not to take the trip. A theory that maintains this kind of orderly relationship between cause and effect is said to satisfy causality.

Causality is like a water-hungry front-yard lawn in Los Angeles: we know we want it, but it's not easy to explain why. Even in plain old Newtonian physics, there is no clear distinction between past and future. In figure c, number 18 throws the football to number 25, and the ball obeys Newton's laws of motion. If we took a video of the pass and played it backward, we would see the ball flying from 25 to 18, and Newton's laws would still be satisfied. Nevertheless, we have a strong psychological impression that there is a forward arrow of time. I can remember what the stock market did last year, but I can't remember what it will do next year. Joan of Arc's military victories against England caused the English to burn her at the stake; it's hard to accept that Newton's laws provide an equally good description of a process in which her execution in 1431 caused her to win a battle in 1429. There is no consensus at this point among physicists on the origin and significance of time's arrow, and for our present purposes we don't need to solve this mystery. Instead, we merely note the empirical fact that, regardless of what causality really means and where it really comes from, its behavior is consistent. Specifically, experiments show that if an observer in a certain frame of reference observes that event A causes event B, then observers in other frames agree that A causes B, not the other way around. This is merely a generalization about a large body of experimental results, not a logically necessary assumption. If Keating had gone around the world and arrived back in Washington before he left, it would have disproved this statement about causality.

17.1.3 Time distortion arising from motion and gravity

Hafele and Keating were testing specific quantitative predictions of relativity, and they verified them to within their experiment's error bars. Let's work backward instead, and inspect the empirical results for clues as to how time works.

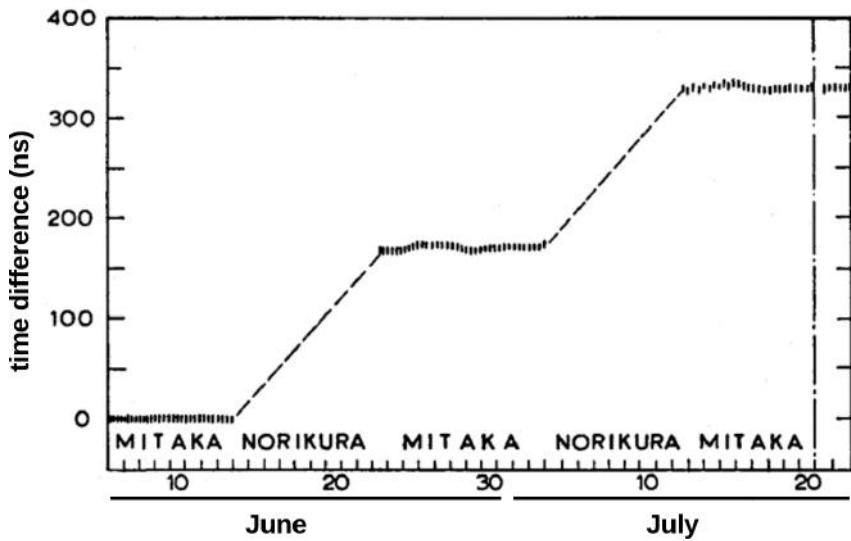
The two traveling clocks experienced effects in opposite directions, and this suggests that the rate at which time flows depends on the motion of the observer. The east-going clock was moving in the same direction as the earth's rotation, so its velocity relative to the earth's center was greater than that of the clock that remained in Washington, while the west-going clock's velocity was correspondingly reduced. The fact that the east-going clock fell behind, and the west-going one got ahead, shows that the effect of motion is to

make time go more slowly. This effect of motion on time was predicted by Einstein in his original 1905 paper on relativity, written when he was 26.

If this had been the only effect in the Hafele-Keating experiment, then we would have expected to see effects on the two flying clocks that were equal in size. Making up some simple numbers to keep the arithmetic transparent, suppose that the earth rotates from west to east at 1000 km/hr, and that the planes fly at 300 km/hr. Then the speed of the clock on the ground is 1000 km/hr, the speed of the clock on the east-going plane is 1300 km/hr, and that of the west-going clock 700 km/hr. Since the speeds of 700, 1000, and 1300 km/hr have equal spacing on either side of 1000, we would expect the discrepancies of the moving clocks relative to the one in the lab to be equal in size but opposite in sign.

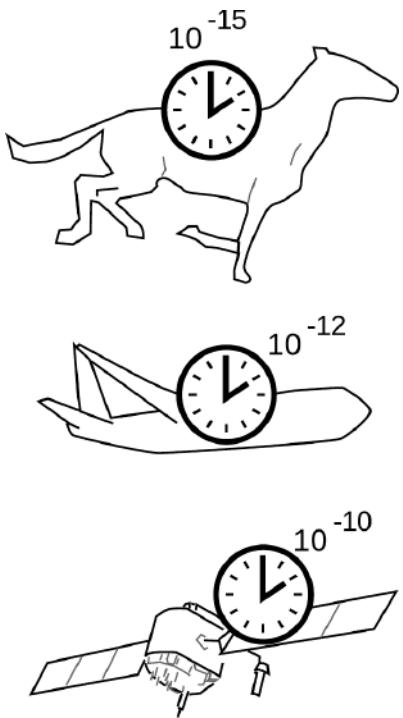


d / All three clocks are moving to the east. Even though the west-going plane is moving to the west relative to the air, the air is moving to the east due to the earth's rotation.



e / A graph showing the time difference between two atomic clocks. One clock was kept at Mitaka Observatory, at 58 m above sea level. The other was moved back and forth to a second observatory, Norikura Corona Station, at the peak of the Norikura volcano, 2876 m above sea level. The plateaus on the graph are data from the periods when the clocks were compared side by side at Mitaka. The difference between one plateau and the next shows a gravitational effect on the rate of flow of time, accumulated during the period when the mobile clock was at the top of Norikura.

In fact, the two effects are unequal in size: -59 ns and 273 ns. This implies that there is a second effect involved, simply due to the planes' being up in the air. This was verified more directly in a 1978 experiment by Iijima and Fujiwara, figure e, in which identical atomic clocks were kept at rest at the top and bottom of a



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

mountain near Tokyo. This experiment, unlike the Hafele-Keating one, isolates one effect on time, the gravitational one: time's rate of flow increases with height in a gravitational field. Einstein didn't figure out how to incorporate gravity into relativity until 1915, after much frustration and many false starts. The simpler version of the theory without gravity is known as special relativity, the full version as general relativity. We'll restrict ourselves to special relativity, and that means that what we want to focus on right now is the distortion of time due to motion, not gravity.

We can now see in more detail how to apply the correspondence principle. The behavior of the three clocks in the Hafele-Keating experiment shows that the amount of time distortion increases as the speed of the clock's motion increases. 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.

By the way, don't get an inflated idea of the importance of the Hafele-Keating experiment. Special relativity had already been confirmed by a vast and varied body of experiments decades before 1971. The only reason I'm giving such a prominent role to this experiment, which was actually more important as a test of general relativity, is that it is conceptually very direct.

17.2 Distortion of space and time

17.2.1 The Lorentz transformation

Relativity says that when two observers are in different frames of reference, each observer considers the other one's perception of time to be distorted. We'll also see that something similar happens to their observations of distances, so both space and time are distorted.

What exactly is this distortion? How do we even conceptualize it?

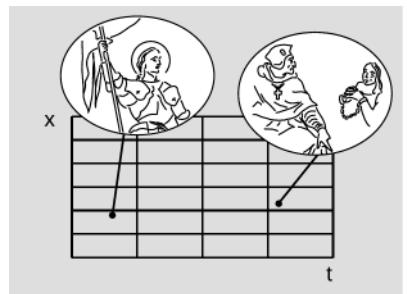
The idea isn't really as radical as it might seem at first. We can visualize the structure of space and time using a graph with position and time on its axes. These graphs are familiar by now, but we're going to look at them in a slightly different way. Before, we used them to describe the motion of objects. The grid underlying the graph was merely the stage on which the actors played their parts. Now the background comes to the foreground: it's time and space themselves that we're studying. We don't necessarily need to have a line or a curve drawn on top of the grid to represent a particular object. We may, for example, just want to talk about events, depicted as points on the graph as in figure g. A distortion of the Cartesian grid underlying the graph can arise for perfectly ordinary reasons that Isaac Newton would have readily accepted. For example, we can simply change the units used to measure time and position, as in figure h.

We're going to have quite a few examples of this type, so I'll adopt the convention shown in figure i for depicting them. Figure i summarizes the relationship between figures g and h in a more compact form. The gray rectangle represents the original coordinate grid of figure g, while the grid of black lines represents the new version from figure h. Omitting the grid from the gray rectangle makes the diagram easier to decode visually.

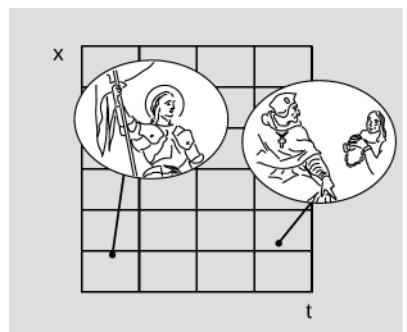
Our goal of unraveling the mysteries of special relativity amounts to nothing more than finding out how to draw a diagram like i in the case where the two different sets of coordinates represent measurements of time and space made by two different observers, each in motion relative to the other. Galileo and Newton thought they knew the answer to this question, but their answer turned out to be only approximately right. To avoid repeating the same mistakes, we need to clearly spell out what we think are the basic properties of time and space that will be a reliable foundation for our reasoning. I want to emphasize that there is no purely logical way of deciding on this list of properties. The ones I'll list are simply a summary of the patterns observed in the results from a large body of experiments. Furthermore, some of them are only approximate. For example, property 1 below is only a good approximation when the gravitational field is weak, so it is a property that applies to special relativity, not to general relativity.

Experiments show that:

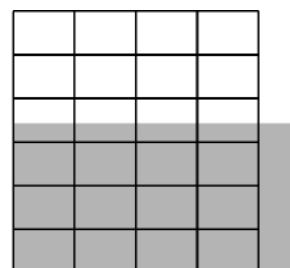
1. No point in time or space has properties that make it different from any other point.
2. Likewise, all directions in space have the same properties.
3. Motion is relative, i.e., all inertial frames of reference are



g / Two events are given as points on a graph of position versus time. Joan of Arc helps to restore Charles VII to the throne. At a later time and a different position, Joan of Arc is sentenced to death.



h / A change of units distorts an x - t graph. This graph depicts exactly the same events as figure g. The only change is that the x and t coordinates are measured using different units, so the grid is compressed in t and expanded in x .



i / A convention we'll use to represent a distortion of time and space.

equally valid.

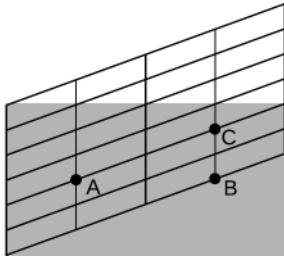
4. Causality holds, in the sense described on page 378.
5. Time depends on the state of motion of the observer.

Most of these are not very subversive. Properties 1 and 2 date back to the time when Galileo and Newton started applying the same universal laws of motion to the solar system and to the earth; this contradicted Aristotle, who believed that, for example, a rock would naturally want to move in a certain special direction (down) in order to reach a certain special location (the earth's surface). Property 3 is the reason that Einstein called his theory "relativity," but Galileo and Newton believed exactly the same thing to be true, as dramatized by Galileo's run-in with the Church over the question of whether the earth could really be in motion around the sun. Property 4 would probably surprise most people only because it asserts in such a weak and specialized way something that they feel deeply must be true. The only really strange item on the list is 5, but the Hafele-Keating experiment forces it upon us.

If it were not for property 5, we could imagine that figure j would give the correct transformation between frames of reference in motion relative to one another. Let's say that observer 1, whose grid coincides with the gray rectangle, is a hitch-hiker standing by the side of a road. Event A is a raindrop hitting his head, and event B is another raindrop hitting his head. He says that A and B occur at the same location in space. Observer 2 is a motorist who drives by without stopping; to him, the passenger compartment of his car is at rest, while the asphalt slides by underneath. He says that A and B occur at different points in space, because during the time between the first raindrop and the second, the hitch-hiker has moved backward. On the other hand, observer 2 says that events A and C occur in the same place, while the hitch-hiker disagrees. The slope of the grid-lines is simply the velocity of the relative motion of each observer relative to the other.

Figure j has familiar, comforting, and eminently sensible behavior, but it also happens to be wrong, because it violates property 5. The distortion of the coordinate grid has only moved the vertical lines up and down, so both observers agree that events like B and C are simultaneous. If this was really the way things worked, then all observers could synchronize all their clocks with one another for once and for all, and the clocks would never get out of sync. This contradicts the results of the Hafele-Keating experiment, in which all three clocks were initially synchronized in Washington, but later went out of sync because of their different states of motion.

It might seem as though we still had a huge amount of wiggle room available for the correct form of the distortion. It turns out, however, that properties 1-5 are sufficient to prove that there is only



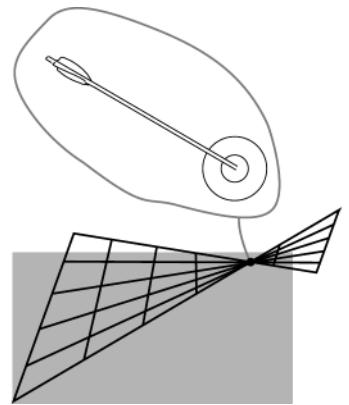
j / A Galilean version of the relationship between two frames of reference. As in all such graphs in this chapter, the original coordinates, represented by the gray rectangle, have a time axis that goes to the right, and a position axis that goes straight up.

one answer, which is the one found by Einstein in 1905. To see why this is, let's work by a process of elimination.

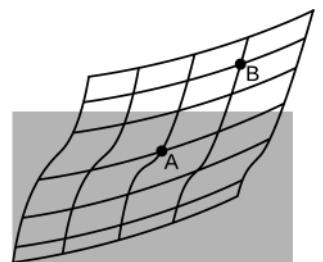
Figure k shows a transformation that might seem at first glance to be as good a candidate as any other, but it violates property 3, that motion is relative, for the following reason. In observer 2's frame of reference, some of the grid lines cross one another. This means that observers 1 and 2 disagree on whether or not certain events are the same. For instance, suppose that event A marks the arrival of an arrow at the bull's-eye of a target, and event B is the location and time when the bull's-eye is punctured. Events A and B occur at the same location and at the same time. If one observer says that A and B coincide, but another says that they don't, we have a direct contradiction. Since the two frames of reference in figure k give contradictory results, one of them is right and one is wrong. This violates property 3, because all inertial frames of reference are supposed to be equally valid. To avoid problems like this, we clearly need to make sure that none of the grid lines ever cross one another.

The next type of transformation we want to kill off is shown in figure l, in which the grid lines curve, but never cross one another. The trouble with this one is that it violates property 1, the uniformity of time and space. The transformation is unusually "twisty" at A, whereas at B it's much more smooth. This can't be correct, because the transformation is only supposed to depend on the relative state of motion of the two frames of reference, and that given information doesn't single out a special role for any particular point in spacetime. If, for example, we had one frame of reference *rotating* relative to the other, then there would be something special about the axis of rotation. But we're only talking about *inertial* frames of reference here, as specified in property 3, so we can't have rotation; each frame of reference has to be moving in a straight line at constant speed. For frames related in this way, there is nothing that could single out an event like A for special treatment compared to B, so transformation l violates property 1.

The examples in figures k and l show that the transformation we're looking for must be linear, meaning that it must transform lines into lines, and furthermore that it has to take parallel lines to parallel lines. Einstein wrote in his 1905 paper that "...on account of the property of homogeneity [property 1] which we ascribe to time and space, the [transformation] must be linear."¹ Applying this to our diagrams, the original gray rectangle, which is a special type of parallelogram containing right angles, must be transformed into another parallelogram. There are three types of transformations, figure m, that have this property. Case I is the Galilean transformation of figure j on page 382, which we've already ruled out.

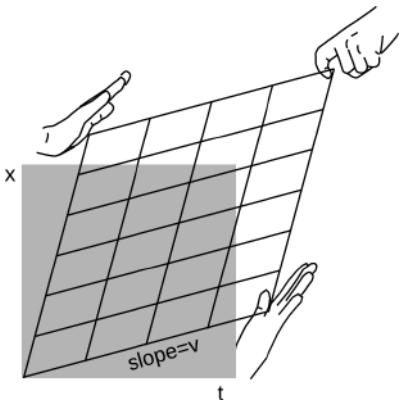
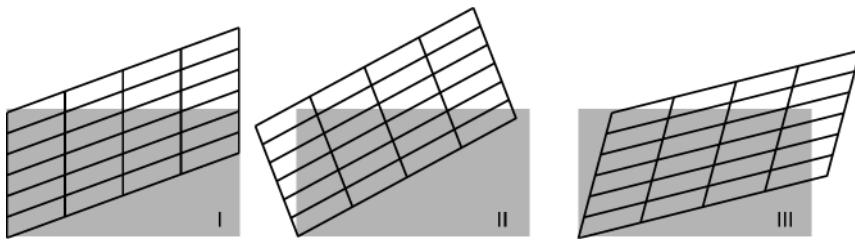


k / A transformation that leads to disagreements about whether two events occur at the same time and place. This is not just a matter of opinion. Either the arrow hit the bull's-eye or it didn't.

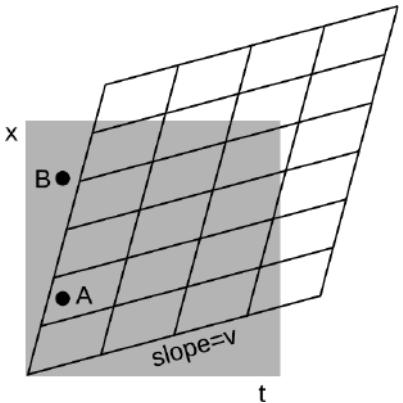


l / A nonlinear transformation.

¹A. Einstein, "On the Electrodynamics of Moving Bodies," *Annalen der Physik* 17 (1905), p. 891, tr. Saha and Bose.



n / In the units that are most convenient for relativity, the transformation has symmetry about a 45-degree diagonal line.



o / Interpretation of the Lorentz transformation. The slope indicated in the figure gives the relative velocity of the two frames of reference. Events A and B that were simultaneous in frame 1 are not simultaneous in frame 2, where event A occurs to the right of the $t = 0$ line represented by the left edge of the grid, but event B occurs to its left.

m / Three types of transformations that preserve parallelism. Their distinguishing feature is what they do to simultaneity, as shown by what happens to the left edge of the original rectangle. In I, the left edge remains vertical, so simultaneous events remain simultaneous. In II, the left edge turns counterclockwise. In III, it turns clockwise.

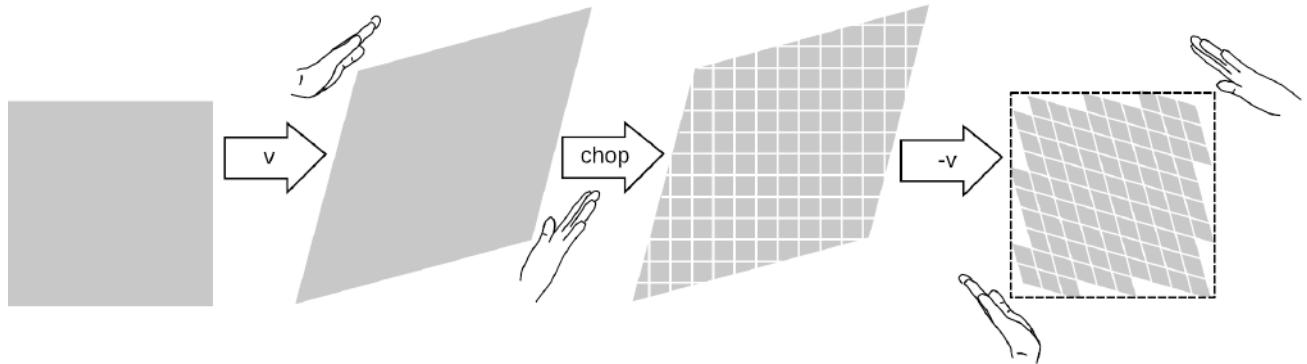
Case II can also be discarded. Here every point on the grid rotates counterclockwise. What physical parameter would determine the amount of rotation? The only thing that could be relevant would be v , the relative velocity of the motion of the two frames of reference with respect to one another. But if the angle of rotation was proportional to v , then for large enough velocities the grid would have left and right reversed, and this would violate property 4, causality: one observer would say that event A caused a later event B, but another observer would say that B came first and caused A.

The only remaining possibility is case III, which I've redrawn in figure n with a couple of changes. This is the one that Einstein predicted in 1905. The transformation is known as the Lorentz transformation, after Hendrik Lorentz (1853-1928), who partially anticipated Einstein's work, without arriving at the correct interpretation. The distortion is a kind of smooshing and stretching, as suggested by the hands. Also, we've already seen in figures g-i on page 381 that we're free to stretch or compress everything as much as we like in the horizontal and vertical directions, because this simply corresponds to choosing different units of measurement for time and distance. In figure n I've chosen units that give the whole drawing a convenient symmetry about a 45-degree diagonal line. Ordinarily it wouldn't make sense to talk about a 45-degree angle on a graph whose axes had different units. But in relativity, the symmetric appearance of the transformation tells us that space and time ought to be treated on the same footing, and measured in the same units.

As in our discussion of the Galilean transformation, slopes are interpreted as velocities, and the slope of the near-horizontal lines in figure o is interpreted as the relative velocity of the two observers. The difference between the Galilean version and the relativistic one is that now there is smooshing happening from the other side as well. Lines that were vertical in the original grid, representing si-

multaneous events, now slant over to the right. This tells us that, as required by property 5, different observers do not agree on whether events that occur in different places are simultaneous. The Hafele-Keating experiment tells us that this non-simultaneity effect is fairly small, even when the velocity is as big as that of a passenger jet, and this is what we would have anticipated by the correspondence principle. The way that this is expressed in the graph is that if we pick the time unit to be the second, then the distance unit turns out to be hundreds of thousands of miles. In these units, the velocity of a passenger jet is an extremely small number, so the slope v in figure o is extremely small, and the amount of distortion is tiny — it would be much too small to see on this scale.

The only thing left to determine about the Lorentz transformation is the size of the transformed parallelogram relative to the size of the original one. Although the drawing of the hands in figure n may suggest that the grid deforms like a framework made of rigid coat-hanger wire, that is not the case. If you look carefully at the figure, you'll see that the edges of the smooshed parallelogram are actually a little longer than the edges of the original rectangle. In fact what stays the same is not lengths but *areas*, as proved in the caption to figure p.



p / Proof that Lorentz transformations don't change area: We first subject a square to a transformation with velocity v , and this increases its area by a factor $R(v)$, which we want to prove equals 1. We chop the resulting parallelogram up into little squares and finally apply a $-v$ transformation; this changes each little square's area by a factor $R(-v)$, so the whole figure's area is also scaled by $R(-v)$. The final result is to restore the square to its original shape and area, so $R(v)R(-v) = 1$. But $R(v) = R(-v)$ by property 2 of spacetime on page 381, which states that all directions in space have the same properties, so $R(v) = 1$.

17.2.2 The γ factor

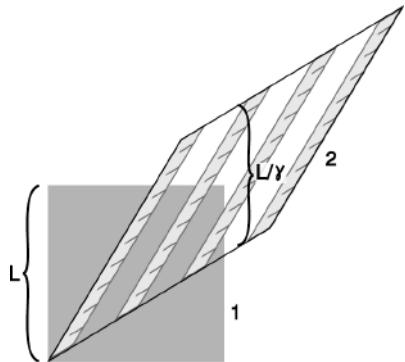
With a little algebra and geometry (homework problem 3, page 421), one can use the equal-area property to show that the factor γ (Greek letter gamma) defined in figure q is given by the equation

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

If you've had good training in physics, the first thing you probably think when you look at this equation is that it must be nonsense, because its units don't make sense. How can we take something with units of velocity squared, and subtract it from a unitless 1? But remember that this is expressed in our special relativistic units, in which the same units are used for distance and time. We refer to these as *natural* units. In this system, velocities are always unitless. This sort of thing happens frequently in physics. For instance, before James Joule discovered conservation of energy, nobody knew that heat and mechanical energy were different forms of the same thing, so instead of measuring them both in units of joules as we would do now, they measured heat in one unit (such as calories) and mechanical energy in another (such as foot-pounds). In ordinary metric units, we just need an extra conversion factor c , and the equation becomes

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

q / The γ factor.

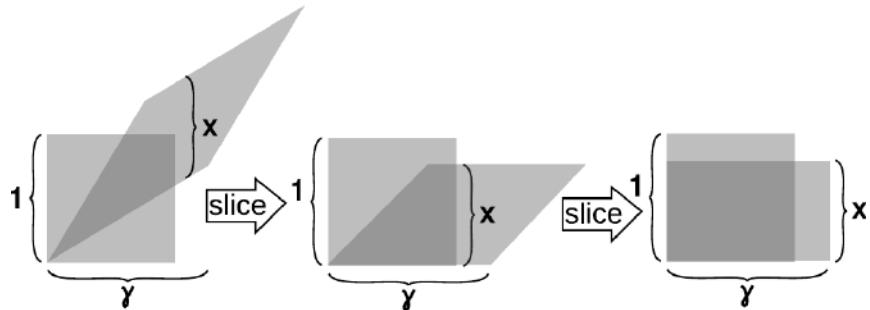


r / The ruler is moving in frame 1, represented by a square, but at rest in frame 2, shown as a parallelogram. Each picture of the ruler is a snapshot taken at a certain moment as judged according to frame 2's notion of simultaneity. An observer in frame 1 judges the ruler's length instead according to frame 1's definition of simultaneity, i.e., using points that are lined up vertically on the graph. The ruler appears shorter in the frame in which it is moving. As proved in figure s, the length contracts from L to L/γ .

Here's why we care about γ . Figure q defines it as the ratio of two times: the time between two events as expressed in one coordinate system, and the time between the same two events as measured in the other one. The interpretation is:

Time dilation

A clock runs fastest in the frame of reference of an observer who is at rest relative to the clock. An observer in motion relative to the clock at speed v perceives the clock as running more slowly by a factor of γ .



s / This figure proves, as claimed in figure r, that the length contraction is $x = 1/\gamma$. First we slice the parallelogram vertically like a salami and slide the slices down, making the top and bottom edges horizontal. Then we do the same in the horizontal direction, forming a rectangle with sides γ and x . Since both the Lorentz transformation and the slicing processes leave areas unchanged, the area γx of the rectangle must equal the area of the original square, which is 1.

As proved in figures r and s, lengths are also distorted:

Length contraction

A meter-stick appears longest to an observer who is at rest relative to it. An observer moving relative to the meter-stick at v observes the stick to be shortened by a factor of γ .

self-check A

What is γ when $v = 0$? What does this mean? ▷ Answer, p. 434

Figure t shows the behavior of γ as a function of v .

Changing an equation from natural units to SI example 1

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. For example, suppose we want to solve for v in terms of γ . In natural units, we have $\gamma = 1/\sqrt{1 - v^2}$, so $\gamma^{-2} = 1 - v^2$, and $v = \sqrt{1 - \gamma^{-2}}$.

This form of the result might be fine for many purposes, but if we wanted to find a value of v in SI units, we would need to reinsert factors of c in the final result. There is no need to do this throughout the whole derivation. By looking at the final result, we see that there is only one possible way to do this so that the results make sense in SI, which is to write $v = c\sqrt{1 - \gamma^{-2}}$.

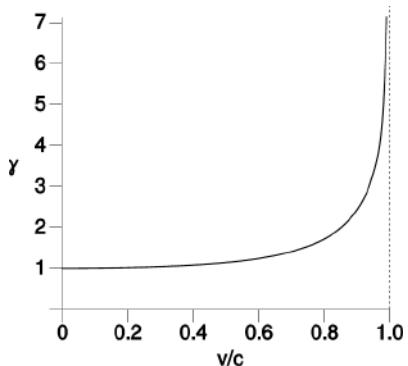
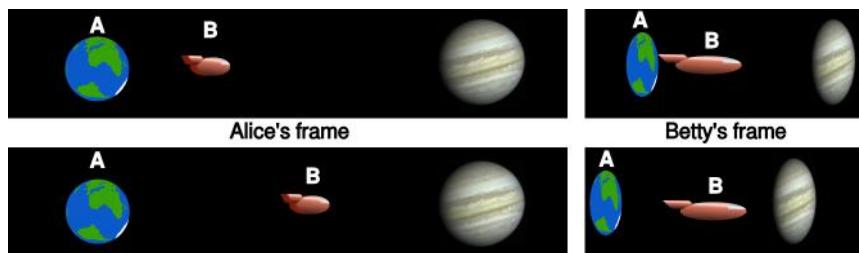
Motion of a ray of light example 2

▷ 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: $x = -ct$.

An interstellar road trip example 3

Alice stays on earth while her twin Betty heads off in a spaceship for Tau Ceti, a nearby star. Tau Ceti is 12 light-years away, so even though Betty travels at 87% of the speed of light, it will take her a long time to get there: 14 years, according to Alice.



t / A graph of γ as a function of v .

Betty experiences time dilation. At this speed, her γ is 2.0, so that

u / Example 3.

the voyage will only seem to her to last 7 years. But there is perfect symmetry between Alice's and Betty's frames of reference, so Betty agrees with Alice on their relative speed; Betty sees herself as being at rest, while the sun and Tau Ceti both move backward at 87% of the speed of light. How, then, can she observe Tau Ceti to get to her in only 7 years, when it should take 14 years to travel 12 light-years at this speed?

We need to take into account length contraction. Betty sees the distance between the sun and Tau Ceti to be shrunk by a factor of 2. The same thing occurs for Alice, who observes Betty and her spaceship to be foreshortened.

The correspondence principle

example 4

The correspondence principle requires that γ be close to 1 for the velocities much less than c encountered in everyday life. In natural units, $\gamma = (1 - v^2)^{-1/2}$. For small values of ϵ , the approximation $(1 + \epsilon)^p \approx 1 + p\epsilon$ holds (see p. 436). Applying this approximation, we find $\gamma \approx 1 + v^2/2$.

As expected, this gives approximately 1 when v is small compared to 1 (i.e., compared to c , which equals 1 in natural units).

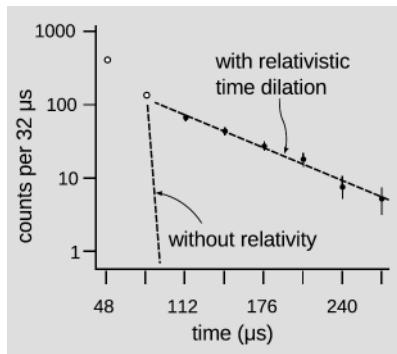
In problem 16 on p. 423 we rewrite this in SI units.

Figure t on p. 387 shows that the approximation is *not* valid for large values of v/c . In fact, γ blows up to infinity as v gets closer and closer to c .

Large time dilation

example 5

The time dilation effect in the Hafele-Keating experiment was very small. If we want to see a large time dilation effect, we can't do it with something the size of the atomic clocks they used; the kinetic energy would be greater than the total megatonnage of all the world's nuclear arsenals. We can, however, accelerate subatomic particles to speeds at which γ is large. For experimental particle physicists, relativity is something you do all day before heading home and stopping off at the store for milk. An early, low-precision experiment of this kind was performed by Rossi and Hall in 1941, using naturally occurring cosmic rays. Figure w shows a 1974 experiment² of a similar type which verified the time dilation predicted by relativity to a precision of about one part per thousand.



v / Muons accelerated to nearly c undergo radioactive decay much more slowly than they would according to an observer at rest with respect to the muons. The first two data-points (unfilled circles) were subject to large systematic errors.

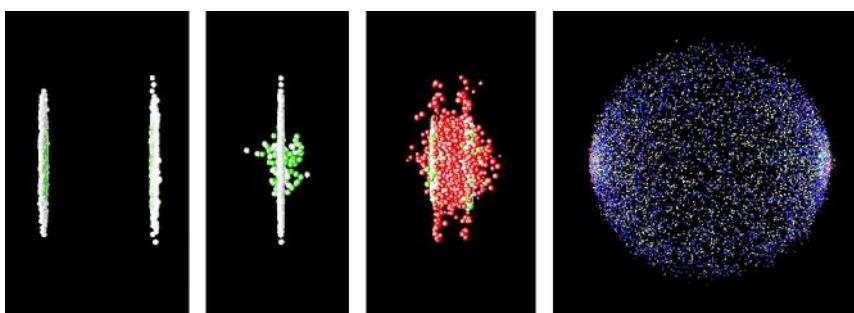
Particles called muons (named after the Greek letter μ , "myoo") were produced by an accelerator at CERN, near Geneva. A muon is essentially a heavier version of the electron. Muons undergo radioactive decay, lasting an average of only 2.197 μ s before they

²Bailey et al., Nucl. Phys. B150(1979) 1



w / Apparatus used for the test of relativistic time dilation described in example 5. The prominent black and white blocks are large magnets surrounding a circular pipe with a vacuum inside. (c) 1974 by CERN.

evaporate into an electron and two neutrinos. The 1974 experiment was actually built in order to measure the magnetic properties of muons, but it produced a high-precision test of time dilation as a byproduct. Because muons have the same electric charge as electrons, they can be trapped using magnetic fields. Muons were injected into the ring shown in figure w, circling around it until they underwent radioactive decay. At the speed at which these muons were traveling, they had $\gamma = 29.33$, so on the average they lasted 29.33 times longer than the normal lifetime. In other words, they were like tiny alarm clocks that self-destructed at a randomly selected time. Figure v shows the number of radioactive decays counted, as a function of the time elapsed after a given stream of muons was injected into the storage ring. The two dashed lines show the rates of decay predicted with and without relativity. The relativistic line is the one that agrees with experiment.

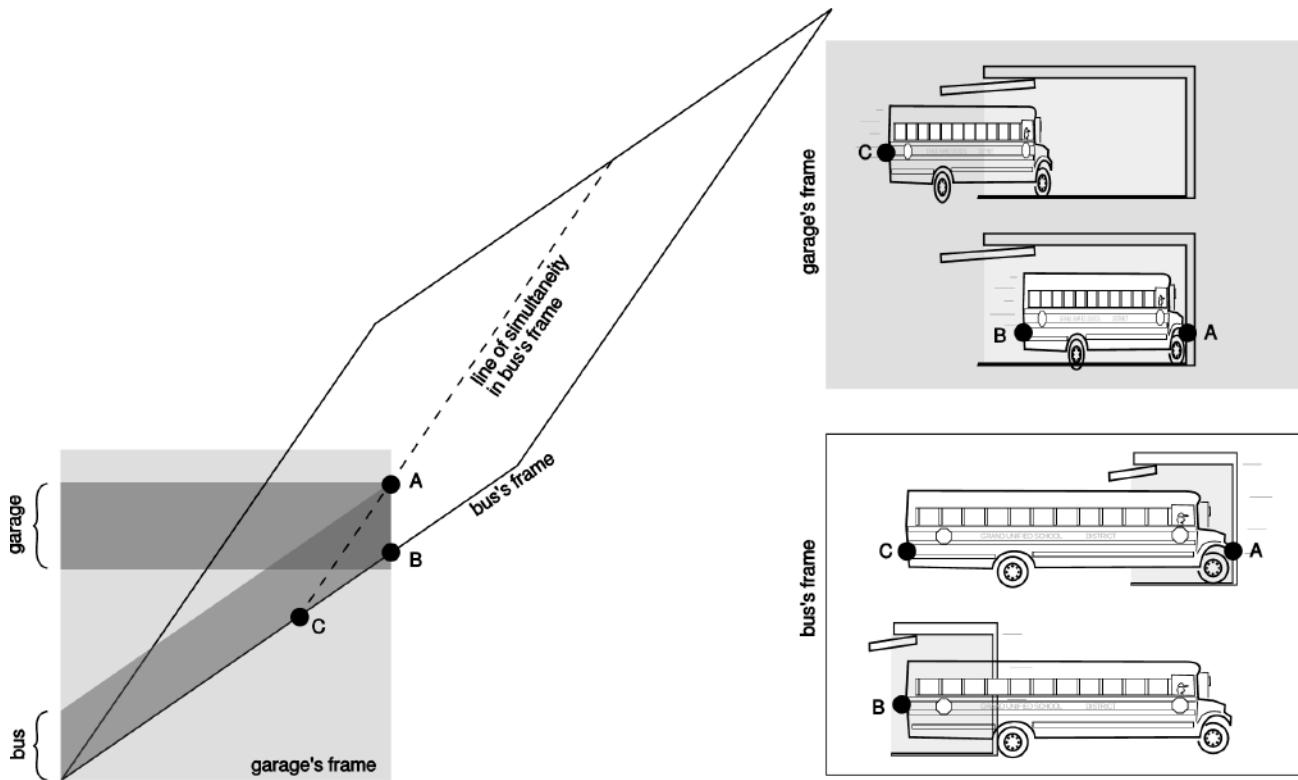


x / Colliding nuclei show relativistic length contraction.

An example of length contraction

example 6

Figure x 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.



y / Example 7: In the garage's frame of reference, the bus is moving, and can fit in the garage due to its length contraction. In the bus's frame of reference, the garage is moving, and can't hold the bus due to *its* length contraction.

The garage paradox

example 7

One of the most famous of all the so-called relativity paradoxes has to do with our incorrect feeling that simultaneity is well defined. The idea is that one could take a schoolbus and drive it at relativistic speeds into a garage of ordinary size, in which it normally would not fit. Because of the length contraction, the bus would supposedly fit in the garage. The driver, however, will perceive the *garage* as being contracted and thus even less able to contain the bus.

The paradox is resolved when we recognize that the concept of

fitting the bus in the garage “all at once” contains a hidden assumption, the assumption that it makes sense to ask whether the front and back of the bus can *simultaneously* be in the garage. Observers in different frames of reference moving at high relative speeds do not necessarily agree on whether things happen simultaneously. As shown in figure y, the person in the garage’s frame can shut the door at an instant B he perceives to be simultaneous with the front bumper’s arrival A at the back wall of the garage, but the driver would not agree about the simultaneity of these two events, and would perceive the door as having shut long after she plowed through the back wall.

17.2.3 The universal speed c

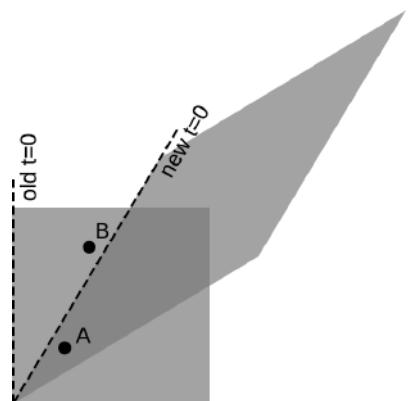
Let’s think a little more about the role of the 45-degree diagonal in the Lorentz transformation. Slopes on these graphs are interpreted as velocities. This line has a slope of 1 in relativistic units, but that slope corresponds to c in ordinary metric units. We already know that the relativistic distance unit must be extremely large compared to the relativistic time unit, so c must be extremely large. Now note what happens when we perform a Lorentz transformation: this particular line gets stretched, but the new version of the line lies right on top of the old one, and its slope stays the same. In other words, if one observer says that something has a velocity equal to c , every other observer will agree on that velocity as well. (The same thing happens with $-c$.)

Velocities don’t simply add and subtract.

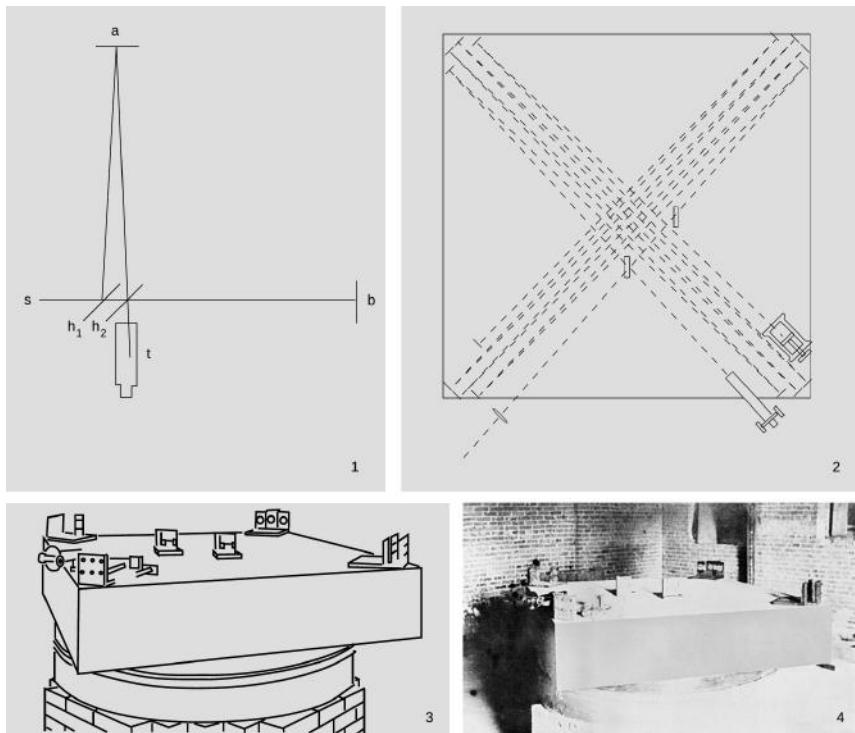
This is counterintuitive, since we expect velocities to add and subtract in relative motion. If a dog is running away from me at 5 m/s relative to the sidewalk, and I run after it at 3 m/s, the dog’s velocity in my frame of reference is 2 m/s. According to everything we have learned about motion, the dog must have different speeds in the two frames: 5 m/s in the sidewalk’s frame and 2 m/s in mine. But velocities are measured by dividing a distance by a time, and both distance and time are distorted by relativistic effects, so we actually shouldn’t expect the ordinary arithmetic addition of velocities to hold in relativity; it’s an approximation that’s valid at velocities that are small compared to c .

A universal speed limit

For example, suppose Janet takes a trip in a spaceship, and accelerates until she is moving at $0.6c$ relative to the earth. She then launches a space probe in the forward direction at a speed relative to her ship of $0.6c$. We might think that the probe was then moving at a velocity of $1.2c$, but in fact the answer is still less than c (problem 1, page 420). This is an example of a more general fact about relativity, which is that c represents a universal speed limit. This is required by causality, as shown in figure z.



z / A proof that causality imposes a universal speed limit. In the original frame of reference, represented by the square, event A happens a little before event B. In the new frame, shown by the parallelogram, A happens after $t = 0$, but B happens before $t = 0$; that is, B happens before A. The time ordering of the two events has been reversed. This can only happen because events A and B are very close together in time and fairly far apart in space. The line segment connecting A and B has a slope greater than 1, meaning that if we wanted to be present at both events, we would have to travel at a speed greater than c (which equals 1 in the units used on this graph). You will find that if you pick any two points for which the slope of the line segment connecting them is less than 1, you can never get them to straddle the new $t = 0$ line in this funny, time-reversed way. Since different observers disagree on the time order of events like A and B, causality requires that information never travel from A to B or from B to A; if it did, then we would have time-travel paradoxes. The conclusion is that c is the maximum speed of cause and effect in relativity.



aa / The Michelson-Morley experiment, shown in photographs, and drawings from the original 1887 paper. 1. A simplified drawing of the apparatus. A beam of light from the source, s , is partially reflected and partially transmitted by the half-silvered mirror h_1 . The two half-intensity parts of the beam are reflected by the mirrors at a and b , reunited, and observed in the telescope, t . If the earth's surface was supposed to be moving through the ether, then the times taken by the two light waves to pass through the moving ether would be unequal, and the resulting time lag would be detectable by observing the interference between the waves when they were reunited. 2. In the real apparatus, the light beams were reflected multiple times. The effective length of each arm was increased to 11 meters, which greatly improved its sensitivity to the small expected difference in the speed of light. 3. In an earlier version of the experiment, they had run into problems with its "extreme sensitiveness to vibration," which was "so great that it was impossible to see the interference fringes except at brief intervals ... even at two o'clock in the morning." They therefore mounted the whole thing on a massive stone floating in a pool of mercury, which also made it possible to rotate it easily. 4. A photo of the apparatus.

Light travels at c .

Now consider a beam of light. We're used to talking casually about the "speed of light," but what does that really mean? Motion is relative, so normally if we want to talk about a velocity, we have to specify what it's measured relative to. A sound wave has a certain speed relative to the air, and a water wave has its own speed relative

to the water. If we want to measure the speed of an ocean wave, for example, we should make sure to measure it in a frame of reference at rest relative to the water. But light isn't a vibration of a physical medium; it can propagate through the near-perfect vacuum of outer space, as when rays of sunlight travel to earth. This seems like a paradox: light is supposed to have a specific speed, but there is no way to decide what frame of reference to measure it in. The way out of the paradox is that light must travel at a velocity equal to c . Since all observers agree on a velocity of c , regardless of their frame of reference, everything is consistent.

The Michelson-Morley experiment

The constancy of the speed of light had in fact already been observed when Einstein was an 8-year-old boy, but because nobody could figure out how to interpret it, the result was largely ignored. In 1887 Michelson and Morley set up a clever apparatus to measure any difference in the speed of light beams traveling east-west and north-south. The motion of the earth around the sun at 110,000 km/hour (about 0.01% of the speed of light) is to our west during the day. Michelson and Morley believed that light was a vibration of a mysterious medium called the ether, so they expected that the speed of light would be a fixed value relative to the ether. As the earth moved through the ether, they thought they would observe an effect on the velocity of light along an east-west line. For instance, if they released a beam of light in a westward direction during the day, they expected that it would move away from them at less than the normal speed because the earth was chasing it through the ether. They were surprised when they found that the expected 0.01% change in the speed of light did not occur.

The ring laser gyroscope

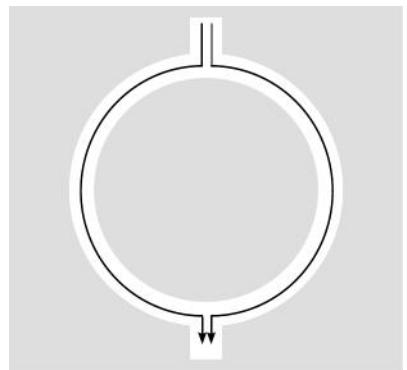
example 8

If you've flown in a jet plane, you can thank relativity for helping you to avoid crashing into a mountain or an ocean. Figure ab shows a standard piece of navigational equipment called a ring laser gyroscope. A beam of light is split into two parts, sent around the perimeter of the device, and reunited. Since the speed of light is constant, we expect the two parts to come back together at the same time. If they don't, it's evidence that the device has been rotating. The plane's computer senses this and notes how much rotation has accumulated.

No frequency-dependence

example 9

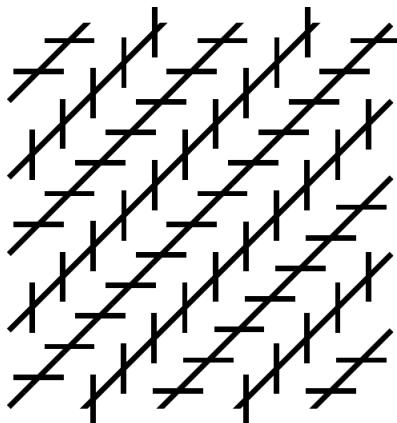
Relativity has only one universal speed, so it requires that all light waves travel at the same speed, regardless of their frequency and wavelength. Presently the best experimental tests of the invariance of the speed of light with respect to wavelength come from astronomical observations of gamma-ray bursts, which are sudden outpourings of high-frequency light, believed to originate from a supernova explosion in another galaxy. One such obser-



ab / A ring laser gyroscope.

vation, in 2009,³ found that the times of arrival of all the different frequencies in the burst differed by no more than 2 seconds out of a total time in flight on the order of ten billion years!

Discussion questions



Discussion question B

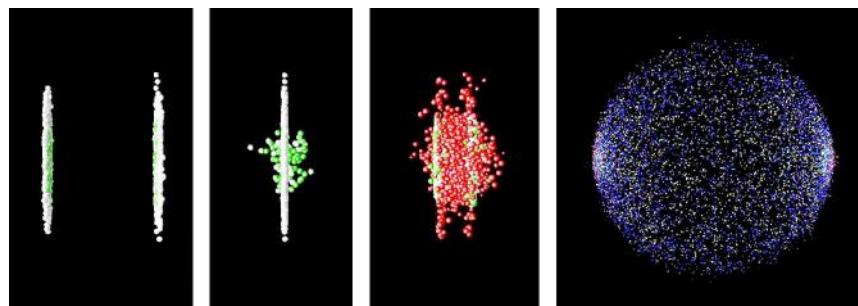
A A person in a spaceship moving at 99.99999999% of the speed of light relative to Earth shines a flashlight forward through dusty air, so the beam is visible. What does she see? What would it look like to an observer on Earth?

B A question that students often struggle with is whether time and space can really be distorted, or whether it just seems that way. Compare with optical illusions or magic tricks. How could you verify, for instance, that the lines in the figure are actually parallel? Are relativistic effects the same, or not?

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

D 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?

E Figure x from page 389, depicting the collision of two nuclei at the RHIC accelerator, is reproduced below. 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?



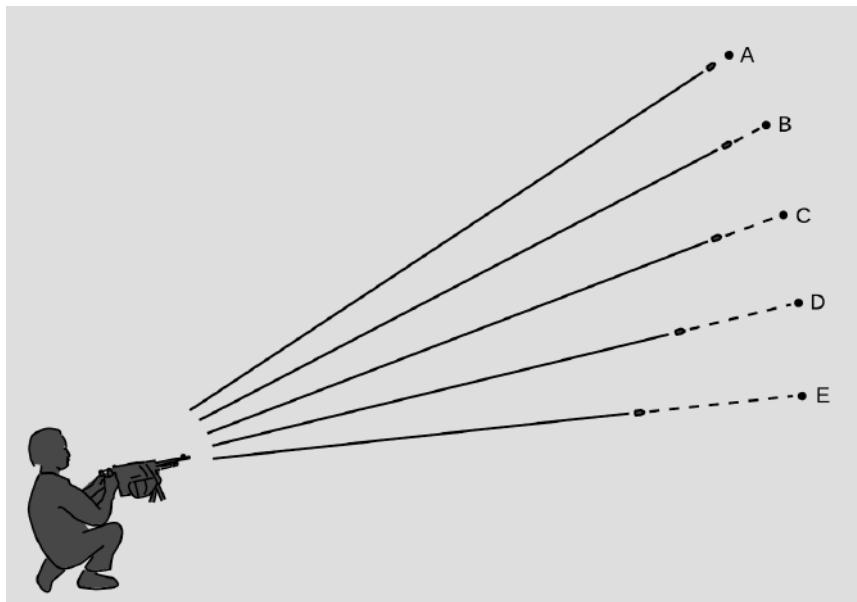
ac / Discussion question E: colliding nuclei show relativistic length contraction.

F If you stick a piece of foam rubber out the window of your car while driving down the freeway, the wind may compress it a little. Does it make sense to interpret the relativistic length contraction as a type of strain that pushes an object's atoms together like this? How does this relate to discussion question E?

G The machine-gunner in the figure sends out a spray of bullets.

³<http://arxiv.org/abs/0908.1832>

Suppose that the bullets are being shot into outer space, and that the distances traveled are trillions of miles (so that the human figure in the diagram is not to scale). After a long time, the bullets reach the points shown with dots which are all equally far from the gun. Their arrivals at those points are events A through E, which happen at different times. Sketch these events on a position-time graph. The chain of impacts extends across space at a speed greater than c . Does this violate special relativity?



Discussion question G.

17.3 No action at a distance

17.3.1 The Newtonian picture

The Newtonian picture of the universe has particles interacting with each other by exerting forces from a distance, and these forces are imagined to occur without any time delay. For example, suppose that super-powerful aliens, angered when they hear disco music in our AM radio transmissions, come to our solar system on a mission to cleanse the universe of our aesthetic contamination. They apply a force to our sun, causing it to go flying out of the solar system at a gazillion miles an hour. According to Newton's laws, the gravitational force of the sun on the earth will *immediately* start dropping off. This will be detectable on earth, and since sunlight takes eight minutes to get from the sun to the earth, the change in gravitational force will, according to Newton, be the first way in which earthlings learn the bad news — the sun will not visibly start receding until a little later. Although this scenario is fanciful, it shows a real feature of Newton's laws: that information can be transmitted from one place in the universe to another with zero time delay, so that transmission and reception occur at exactly the same instant. Newton was sharp enough to realize that this required a nontrivial assumption, which was that there was some completely objective and well-defined way of saying whether two things *happened* at exactly the same instant. He stated this assumption explicitly: “Absolute, true, and mathematical time, of itself, and from its own nature flows at a constant rate without regard to anything external...”

17.3.2 Time delays in forces exerted at a distance

Relativity forbids Newton's instantaneous action at a distance. For suppose that instantaneous action at a distance existed. It would then be possible to send signals from one place in the universe to another without any time lag. This would allow perfect synchronization of all clocks. But the Hafele-Keating experiment demonstrates that clocks A and B that have been initially synchronized will drift out of sync if one is in motion relative to the other. With instantaneous transmission of signals, we could determine, without having to wait for A and B to be reunited, which was ahead and which was behind. Since they don't need to be reunited, neither one needs to undergo any acceleration; each clock can fix an inertial frame of reference, with a velocity vector that changes neither its direction nor its magnitude. But this violates the principle that constant-velocity motion is relative, because each clock can be considered to be at rest, in its own frame of reference. Since no experiment has ever detected any violation of the relativity of motion, we conclude that instantaneous action at a distance is impossible.

Since forces can't be transmitted instantaneously, it becomes natural to imagine force-effects spreading outward from their source

like ripples on a pond, and we then have no choice but to impute some physical reality to these ripples. We call them fields, and they have their own independent existence. Gravity is transmitted through a field called the gravitational field. Besides gravity, there are other fundamental fields of force such as electricity and magnetism (). Ripples of the electric and magnetic fields turn out to be light waves. This tells us that the speed at which electric and magnetic field ripples spread must be c , and by an argument similar to the one in subsection 17.2.3 the same must hold for any other fundamental field, including the gravitational field.

Fields don't have to wiggle; they can hold still as well. The earth's magnetic field, for example, is nearly constant, which is why we can use it for direction-finding.

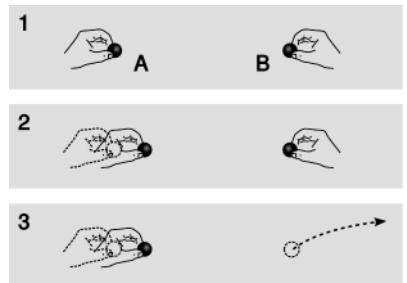
Even empty space, then, is not perfectly featureless. It has measurable properties. For example, we can drop a rock in order to measure the direction of the gravitational field, or use a magnetic compass to find the direction of the magnetic field. This concept made a deep impression on Einstein as a child. He recalled that when he was five years old, the gift of a magnetic compass convinced him that there was "something behind things, something deeply hidden."

17.3.3 More evidence that fields of force are real: they carry energy.

The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy. In figure ae/1, Alice and Betty hold balls A and B at some distance from one another. These balls make a force on each other; it doesn't really matter for the sake of our argument whether this force is gravitational, electrical, or magnetic. Let's say it's electrical, i.e., that the balls have the kind of electrical *charge* that sometimes causes your socks to cling together when they come out of the clothes dryer. We'll say the force is repulsive, although again it doesn't really matter.

If Alice chooses to move her ball closer to Betty's, ae/2, Alice will have to do some mechanical work against the electrical repulsion, burning off some of the calories from that chocolate cheesecake she had at lunch. This reduction in her body's chemical energy is offset by a corresponding increase in the electrical interaction energy. Not only that, but Alice feels the resistance stiffen as the balls get closer together and the repulsion strengthens. She has to do a little extra work, but this is all properly accounted for in the interaction energy.

But now suppose, ae/3, that Betty decides to play a trick on Alice by tossing B far away just as Alice is getting ready to move A. We have already established that Alice can't feel B's motion instantaneously, so the electric forces must actually be propagated by an



ae / Fields carry energy.

electric field. Of course this experiment is utterly impractical, but suppose for the sake of argument that the time it takes the change in the electric field to propagate across the diagram is long enough so that Alice can complete her motion before she feels the effect of B's disappearance. She is still getting stale information about B's position. As she moves A to the right, she feels a repulsion, because the field in her region of space is still the field caused by B in its *old* position. She 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.

If we hope to preserve the law of conservation of energy, then the only possible conclusion is that the electric field itself carries away the cheesecake energy. In fact, this example represents an impractical method of transmitting radio waves. Alice 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 Alice would still have had to do work in order to create them.

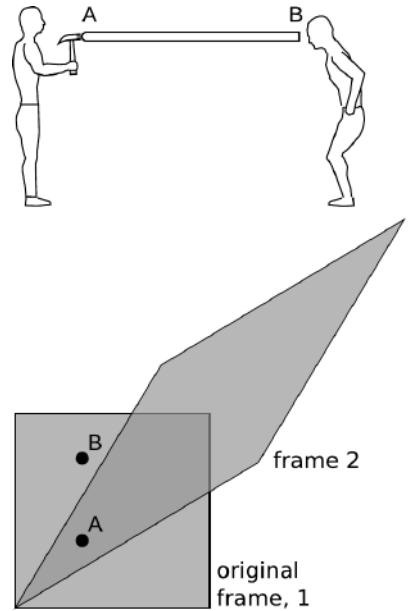
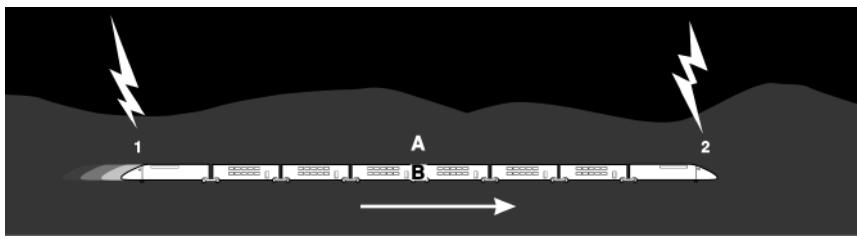
Discussion questions

A Amy and Bill are flying on spaceships in opposite directions at such high velocities that the relativistic effect on time's rate of flow is easily noticeable. Motion is relative, so Amy considers herself to be at rest and Bill to be in motion. She says that time is flowing normally for her, but Bill is slow. But Bill can say exactly the same thing. How can they *both* think the other is slow? Can they settle the disagreement by getting on the radio and seeing whose voice is normal and whose sounds slowed down and Darth-Vadery?



B 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 graph.

C Resolve the following paradox by drawing a spacetime diagram (i.e., a graph of x versus t). Andy and Beth are in motion relative to one another at a significant fraction of c . As they pass by each other, they exchange greetings, and Beth tells Andy that she is going to blow up a stick of dynamite one hour later. One hour later by Andy's clock, she



still hasn't exploded the dynamite, and he says to himself, "She hasn't exploded it because of time dilation. It's only been 40 minutes for her." He now accelerates suddenly so that he's moving at the same velocity as Beth. The time dilation no longer exists. If he looks again, does he suddenly see the flash from the explosion? How can this be? Would he see her go through 20 minutes of her life in fast-motion?

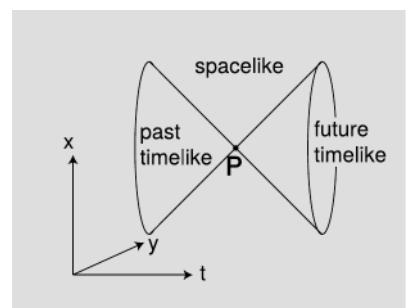
D Use a graph to resolve the following relativity paradox. Relativity says that in one frame of reference, event A could happen before event B, but in someone else's frame B would come before A. How can this be? Obviously the two people could meet up at A and talk as they cruised past each other. Wouldn't they have to agree on whether B had already happened?

E The rod in the figure is perfectly rigid. At event A, the hammer strikes one end of the rod. At event B, the other end moves. Since the rod is perfectly rigid, it can't compress, so A and B are simultaneous. In frame 2, B happens before A. Did the motion at the right end *cause* the person on the left to decide to pick up the hammer and use it?

17.4 The light cone

Given an event P, we can now classify all the causal relationships in which P can participate. In Newtonian physics, these relationships fell into two classes: P could potentially cause any event that lay in its future, and could have been caused by any event in its past. In relativity, we have a three-way distinction rather than a two-way one. There is a third class of events that are too far away from P in space, and too close in time, to allow any cause and effect relationship, since causality's maximum velocity is c . Since we're working in units in which $c = 1$, the boundary of this set is formed by the lines with slope ± 1 on a (t, x) plot. This is referred to as the light cone, for reasons that become more visually obvious when we consider more than one spatial dimension, figure ah.

Events lying inside one another's light cones are said to have a timelike relationship. Events outside each other's light cones are spacelike in relation to one another, and in the case where they lie on the surfaces of each other's light cones the term is lightlike.



ah / The light cone.

17.5 ★ The spacetime interval

The light cone is an object of central importance in both special and general relativity. It relates the *geometry* of spacetime to possible *cause-and-effect* relationships between events. This is fundamentally how relativity works: it's a geometrical theory of causality.

These ideas naturally lead us to ask what fruitful analogies we can form between the bizarre geometry of spacetime and the more familiar geometry of the Euclidean plane. The light cone cuts spacetime into different regions according to certain measurements of relationships between points (events). Similarly, a circle in Euclidean geometry cuts the plane into two parts, an interior and an exterior, according to the measurement of the distance from the circle's center. A circle stays the same when we rotate the plane. A light cone stays the same when we change frames of reference. Let's build up the analogy more explicitly.

Measurement in Euclidean geometry

We say that two line segments are congruent, $AB \cong CD$, if the distance between points A and B is the same as the distance between C and D, as measured by a rigid ruler.

Measurement in spacetime

We define $AB \cong CD$ if:

1. AB and CD are both spacelike, and the two distances are equal as measured by a rigid ruler, in a frame where the two events touch the ruler simultaneously.
2. AB and CD are both timelike, and the two time intervals are equal as measured by clocks moving inertially.
3. AB and CD are both lightlike.

The three parts of the relativistic version each require some justification.

Case 1 has to be the way it is because space is part of spacetime. In special relativity, this space is Euclidean, so the definition of congruence has to agree with the Euclidean definition, in the case where it is possible to apply the Euclidean definition. The spacelike relation between the points is both necessary and sufficient to make this possible. If points A and B are spacelike in relation to one another, then a frame of reference exists in which they are simultaneous, so we can use a ruler that is at rest in that frame to measure their distance. If they are lightlike or timelike, then no such frame of reference exists. For example, there is no frame of reference in which Charles VII's restoration to the throne is simultaneous with Joan of Arc's execution, so we can't arrange for both of these events to touch the same ruler at the same time.

The definition in case 2 is the only sensible way to proceed if we are to respect the symmetric treatment of time and space in relativity. The timelike relation between the events is necessary and sufficient to make it possible for a clock to move from one to the other. It makes a difference that the clocks move inertially, because the twins in example 3 on p. 387 disagree on the clock time between the traveling twin's departure and return.

Case 3 may seem strange, since it says that *any* two lightlike intervals are congruent. But this is the only possible definition, because this case can be obtained as a limit of the timelike one. Suppose that AB is a timelike interval, but in the planet earth's frame of reference it would be necessary to travel at almost the speed of light in order to reach B from A. The required speed is less than c (i.e., less than 1) by some tiny amount ϵ . In the earth's frame, the clock referred to in the definition suffers extreme time dilation. The time elapsed on the clock is very small. As ϵ approaches zero, and the relationship between A and B approaches a lightlike one, this clock time approaches zero. In this sense, the relativistic notion of "distance" is very different from the Euclidean one. In Euclidean geometry, the distance between two points can only be zero if they are the same point.

The case splitting involved in the relativistic definition is a little ugly. Having worked out the physical interpretation, we can now consolidate the definition in a nicer way by appealing to Cartesian coordinates.

Cartesian definition of distance in Euclidean geometry

Given a vector $(\Delta x, \Delta y)$ from point A to point B, the square of the distance between them is defined as $\overline{AB}^2 = \Delta x^2 + \Delta y^2$.

Definition of the interval in relativity

Given points separated by coordinate differences $\Delta x, \Delta y, \Delta z$, and Δt , the spacetime interval \mathcal{I} (cursive letter "I") between them is defined as $\mathcal{I} = \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$.

This is stated in natural units, so all four terms on the right-hand side have the same units; in metric units with $c \neq 1$, appropriate factors of c should be inserted in order to make the units of the terms agree. The interval \mathcal{I} is positive if AB is timelike (regardless of which event comes first), zero if lightlike, and negative if spacelike. Since \mathcal{I} can be negative, we can't in general take its square root and define a real number \overline{AB} as in the Euclidean case. When the interval is timelike, we can interpret $\sqrt{\mathcal{I}}$ as a time, and when it's spacelike we can take $\sqrt{-\mathcal{I}}$ to be a distance.

The Euclidean definition of distance (i.e., the Pythagorean theorem) is useful because it gives the same answer regardless of how we rotate the plane. Although it is stated in terms of a certain coordinate system, its result is unambiguously defined because it is

the same regardless of what coordinate system we arbitrarily pick. Similarly, \mathcal{I} is useful because, as proved in example 11 below, it is the same regardless of our frame of reference, i.e., regardless of our choice of coordinates.

Pioneer 10

example 10

▷ The Pioneer 10 space probe was launched in 1972, and in 1973 was the first craft to fly by the planet Jupiter. It crossed the orbit of the planet Neptune in 1983, after which telemetry data were received until 2002. The following table gives the spacecraft's position relative to the sun at exactly midnight on January 1, 1983 and January 1, 1995. The 1983 date is taken to be $t = 0$.

t (s)	x	y	z
0	1.784×10^{12} m	3.951×10^{12} m	0.237×10^{12} m
3.7869120000×10^8 s	2.420×10^{12} m	8.827×10^{12} m	0.488×10^{12} m

Compare the time elapsed on the spacecraft to the time in a frame of reference tied to the sun.

▷ We can convert these data into natural units, with the distance unit being the second (i.e., a light-second, the distance light travels in one second) and the time unit being seconds. Converting and carrying out this subtraction, we have:

Δt (s)	Δx	Δy	Δz
3.7869120000×10^8 s	0.2121×10^4 s	1.626×10^4 s	0.084×10^4 s

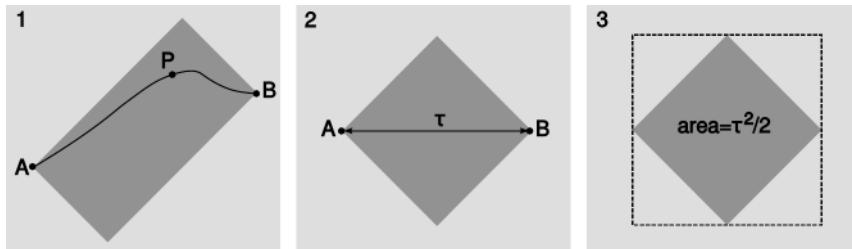
Comparing the exponents of the temporal and spatial numbers, we can see that the spacecraft was moving at a velocity on the order of 10^{-4} of the speed of light, so relativistic effects should be small but not completely negligible.

Since the interval is timelike, we can take its square root and interpret it as the time elapsed on the spacecraft. The result is $\sqrt{\mathcal{I}} = 3.786911996 \times 10^8$ s. This is 0.4 s less than the time elapsed in the sun's frame of reference.

ai / Light-rectangles, example 11.
1. The gray light-rectangle represents the set of all events such as P that could be visited after A and before B.

2. The rectangle becomes a square in the frame in which A and B occur at the same location in space.

3. The area of the dashed square is τ^2 , so the area of the gray square is $\tau^2/2$.



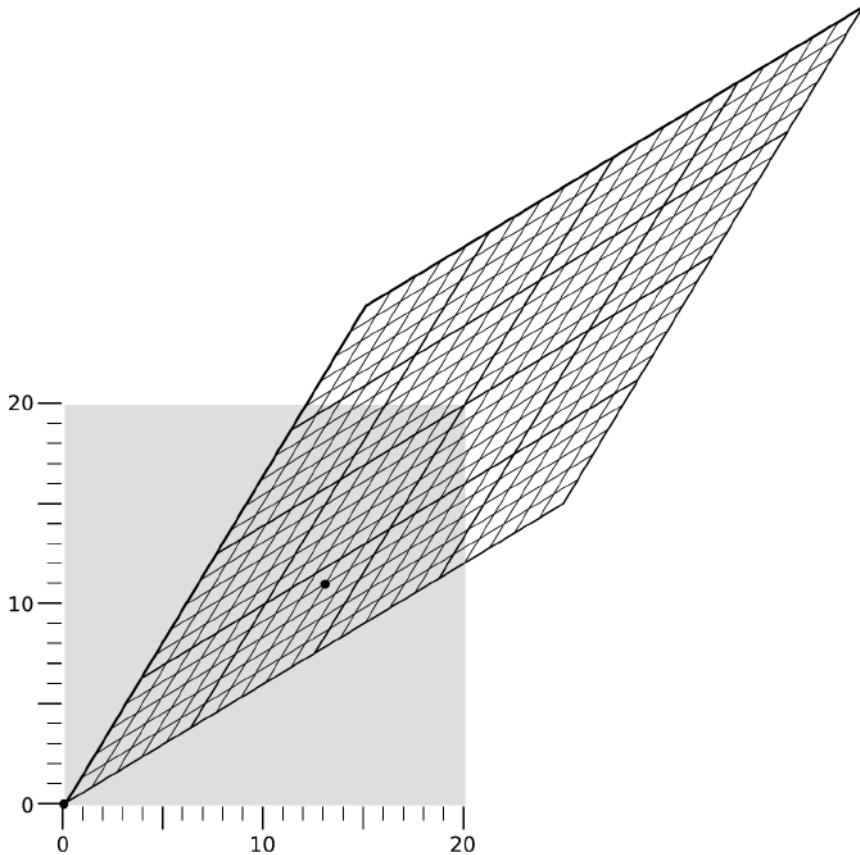
Invariance of the interval

example 11

In this example we prove that the interval is the same regardless of what frame of reference we compute it in. This is called "Lorentz invariance." The proof is limited to the timelike case.

Given events A and B, construct the light-rectangle as defined in figure ai/1. On p. 385 we proved that the Lorentz transformation doesn't change the area of a shape in the x - t plane. Therefore the area of this rectangle is unchanged if we switch to the frame of reference ai/2, in which A and B occurred at the same location and were separated by a time interval τ . This area equals half the interval \mathcal{I} between A and B. But a straightforward calculation shows that the rectangle in ai/1 also has an area equal to half the interval calculated in *that* frame. Since the area in any frame equals half the interval, and the area is the same in all frames, the interval is equal in all frames as well.

aj / Example 12.



A numerical example of invariance

example 12

Figure aj shows two frames of reference in motion relative to one another at $v = 3/5$. (For this velocity, the stretching and squishing of the main diagonals are both by a factor of 2.) 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 interval between these events is $13^2 - 11^2 = 48$. In the

frame represented by the parallelogram, the same two events lie at coordinates

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

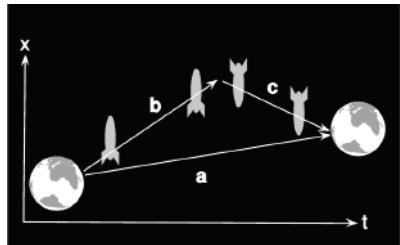
Calculating the interval using these values, the result is $8^2 - 4^2 = 48$, which comes out the same as in the other frame.

17.6 Four-vectors and the inner product

Example 10 makes it natural that we define a type of vector with four components, the first one relating to time and the others being spatial. These are known as four-vectors. It's clear how we should define the equivalent of a dot product in relativity:

$$\mathbf{A} \cdot \mathbf{B} = A_t B_t - A_x B_x - A_y B_y - A_z B_z$$

The term “dot product” has connotations of referring only to three-vectors, so the operation of taking the scalar product of two four-vectors is usually referred to instead as the “inner product.” The spacetime interval can then be thought of as the inner product of a four-vector with itself. We care about the relativistic inner product for exactly the same reason we care about its Euclidean version; both are scalars, so they have a fixed value regardless of what coordinate system we choose.



ak / Example 13.

The twin paradox

example 13

Alice and Betty are identical twins. Betty goes on a space voyage at relativistic speeds, traveling away from the earth and then turning around and coming back. Meanwhile, Alice stays on earth. When Betty returns, she is younger than Alice because of relativistic time dilation (example 3, p. 387).

But isn't it valid to say that Betty's spaceship is standing still and the earth moving? In that description, wouldn't Alice end up younger and Betty older? This is referred to as the “twin paradox.” It can't really be a paradox, since it's exactly what was observed in the Hafele-Keating experiment (p. 377).

Betty's track in the x - t plane (her “world-line” in relativistic jargon) consists of vectors **b** and **c** strung end-to-end (figure ak). We could adopt a frame of reference in which Betty was at rest during **b** (i.e., $b_x = 0$), but there is no frame in which **b** and **c** are parallel, so there is no frame in which Betty was at rest during *both* **b** and **c**. This resolves the paradox.

We have already established by other methods that Betty ages less than Alice, but let's see how this plays out in a simple numerical example. Omitting units and making up simple numbers, let's

say that the vectors in figure ak are

$$\begin{aligned}\mathbf{a} &= (6, 1) \\ \mathbf{b} &= (3, 2) \\ \mathbf{c} &= (3, -1),\end{aligned}$$

where the components are given in the order (t, x) . The time experienced by Alice is then

$$|\mathbf{a}| = \sqrt{6^2 - 1^2} = 5.9,$$

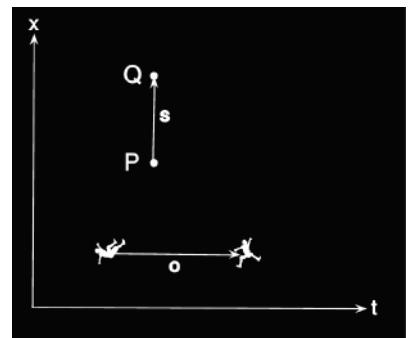
which is greater than the Betty's elapsed time

$$|\mathbf{b}| + |\mathbf{c}| = \sqrt{3^2 - 2^2} + \sqrt{3^2 - (-1)^2} = 5.1.$$

Simultaneity using inner products

example 14

Suppose that an observer O moves inertially along a vector \mathbf{o} , and let the vector separating two events P and Q be \mathbf{s} . O judges these events to be simultaneous if $\mathbf{o} \cdot \mathbf{s} = 0$. To see why this is true, suppose we pick a coordinate system as defined by O. In this coordinate system, O considers herself to be at rest, so she says her vector has only a time component, $\mathbf{o} = (\Delta t, 0, 0, 0)$. If she considers P and Q to be simultaneous, then the vector from P to Q is of the form $(0, \Delta x, \Delta y, \Delta z)$. The inner product is then zero, since each of the four terms vanishes. Since the inner product is independent of the choice of coordinate system, it doesn't matter that we chose one tied to O herself. Any other observer O' can look at O's motion, note that $\mathbf{o} \cdot \mathbf{s} = 0$, and infer that O must consider P and Q to be simultaneous, even if O' says they weren't.



al / Example 14.

17.7 Dynamics

So far we have said nothing about how to predict motion in relativity. Do Newton's laws still work? Do conservation laws still apply? The answer is yes, but many of the definitions need to be modified, and certain entirely new phenomena occur, such as the equivalence of energy and mass, as described by the famous equation $E = mc^2$.

17.7.1 Momentum

Consider the following scheme for traveling faster than the speed of light. The basic idea can be demonstrated by dropping a ping-pong ball and a baseball stacked on top of each other like a snowman. They separate slightly in mid-air, and the baseball therefore has time to hit the floor and rebound before it collides with the ping-pong ball, which is still on the way down. The result is a surprise if you haven't seen it before: the ping-pong ball flies off at high speed and hits the ceiling! A similar fact is known to people who investigate the scenes of accidents involving pedestrians. If a car moving at 90 kilometers per hour hits a pedestrian, the pedestrian flies off at nearly double that speed, 180 kilometers per hour. Now suppose the car was moving at 90 percent of the speed of light. Would the pedestrian fly off at 180% of c ?

To see why not, we have to back up a little and think about where this speed-doubling result comes from. For any collision, there is a special frame of reference, the center-of-mass frame, in which the two colliding objects approach each other, collide, and rebound with their velocities reversed. In the center-of-mass frame, the total momentum of the objects is zero both before and after the collision.

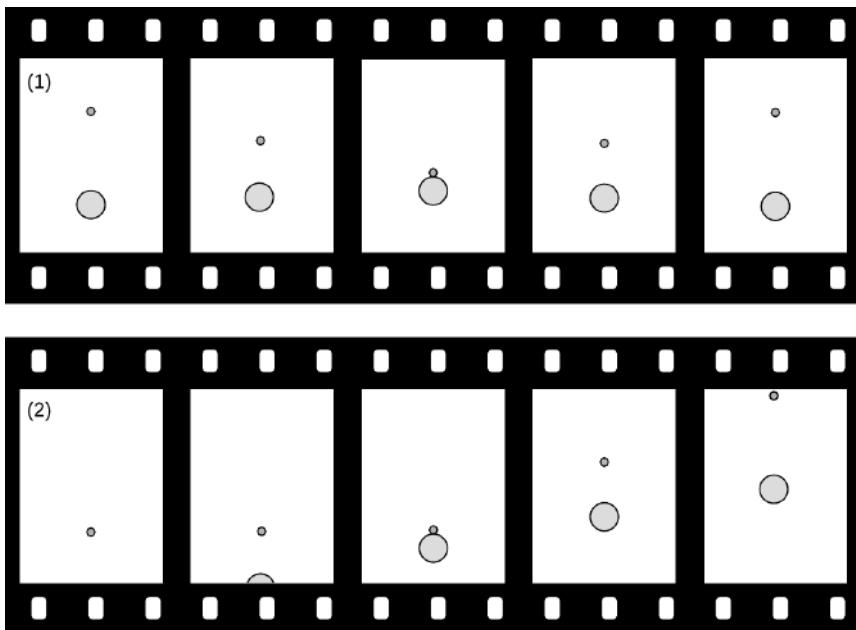


Figure am/1 shows such a frame of reference for objects of very unequal mass. Before the collision, the large ball is moving relatively slowly toward the top of the page, but because of its greater mass, its momentum cancels the momentum of the smaller ball, which is moving rapidly in the opposite direction. The total momentum is zero. After the collision, the two balls just reverse their directions of motion. We know that this is the right result for the outcome of the collision because it conserves both momentum and kinetic energy, and everything not forbidden is compulsory, i.e., in any experiment, there is only one possible outcome, which is the one that obeys all the conservation laws.

self-check B

How do we know that momentum and kinetic energy are conserved in figure am/1?

▷ Answer, p. 434

Let's make up some numbers as an example. Say the small ball has a mass of 1 kg, the big one 8 kg. In frame 1, let's make the velocities as follows:

	before the collision	after the collision
•	-0.8	0.8
○	0.1	-0.1

Figure am/2 shows the same collision in a frame of reference where the small ball was initially at rest. To find all the velocities in this frame, we just add 0.8 to all the ones in the previous table.

	before the collision	after the collision
•	0	1.6
○	0.9	0.7

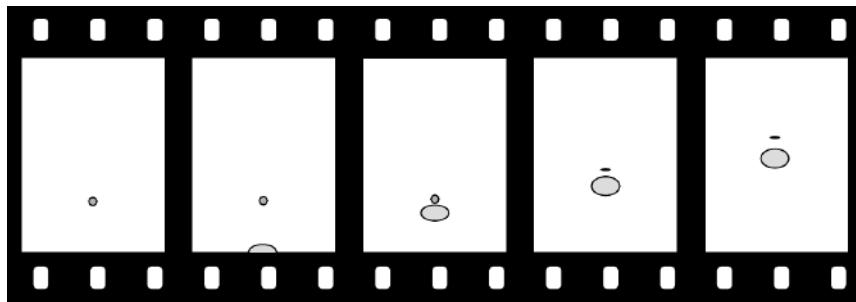
am / An unequal collision, viewed in the center-of-mass frame, 1, and in the frame where the small ball is initially at rest, 2. The motion is shown as it would appear on the film of an old-fashioned movie camera, with an equal amount of time separating each frame from the next. Film 1 was made by a camera that tracked the center of mass, film 2 by one that was initially tracking the small ball, and kept on moving at the same speed after the collision.

In this frame, as expected, the small ball flies off with a velocity, 1.6, that is almost twice the initial velocity of the big ball, 0.9.

If all those velocities were in meters per second, then that's exactly what happened. But what if all these velocities were in units of the speed of light? Now it's no longer a good approximation just to add velocities. We need to combine them according to the relativistic rules. For instance, the technique used in problem 1 on p. 420 can be used to show that combining a velocity of 0.8 times the speed of light with another velocity of 0.8 results in 0.98, not 1.6. The results are very different:

	before the collision	after the collision
•	0	0.98
○	0.83	0.76

an / An 8-kg ball moving at 83% of the speed of light hits a 1-kg ball. The balls appear foreshortened due to the relativistic distortion of space.



We can interpret this as follows. Figure am/1 is one in which the big ball is moving fairly slowly. This is very nearly the way the scene would be seen by an ant standing on the big ball. According to an observer in frame an, however, both balls are moving at nearly the speed of light after the collision. Because of this, the balls appear foreshortened, but the distance between the two balls is also shortened. To this observer, it seems that the small ball isn't pulling away from the big ball very fast.

Now here's what's interesting about all this. The outcome shown in figure am/2 was supposed to be the only one possible, the only one that satisfied both conservation of energy and conservation of momentum. So how can the *different* result shown in figure an be possible? The answer is that relativistically, momentum must not equal mv . The old, familiar definition is only an approximation that's valid at low speeds. If we observe the behavior of the small ball in figure an, it looks as though it somehow had some extra inertia. It's as though a football player tried to knock another player down without realizing that the other guy had a three-hundred-pound bag full of lead shot hidden under his uniform — he just doesn't seem to react to the collision as much as he should. As proved in section 17.7.4, this extra inertia is described by redefining momentum as

$$p = m\gamma v.$$

At very low velocities, γ is close to 1, and the result is very nearly mv , as demanded by the correspondence principle. But at very high velocities, γ gets very big — the small ball in figure an has a γ of 5.0, and therefore has five times more inertia than we would expect nonrelativistically.

This also explains the answer to another paradox often posed by beginners at relativity. Suppose you keep on applying a steady force to an object that's already moving at $0.9999c$. Why doesn't it just keep on speeding up past c ? The answer is that force is the rate of change of momentum. At $0.9999c$, an object already has a γ of 71, and therefore has already sucked up 71 times the momentum you'd expect at that speed. As its velocity gets closer and closer to c , its γ approaches infinity. To move at c , it would need an infinite momentum, which could only be caused by an infinite force.

Push as hard as you like ...

example 15

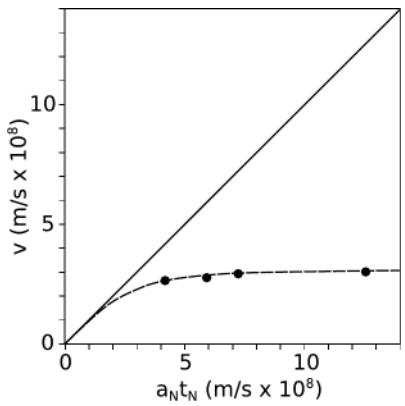
We don't have to depend on our imaginations to see what would happen if we kept on applying a force to an object indefinitely and tried to accelerate it past c . A nice experiment of this type was done by Bertozzi in 1964. In this experiment, electrons were accelerated by an electric field E through a distance ℓ_1 . Applying Newton's laws gives Newtonian predictions a_N for the acceleration and t_N for the time required.⁴

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.

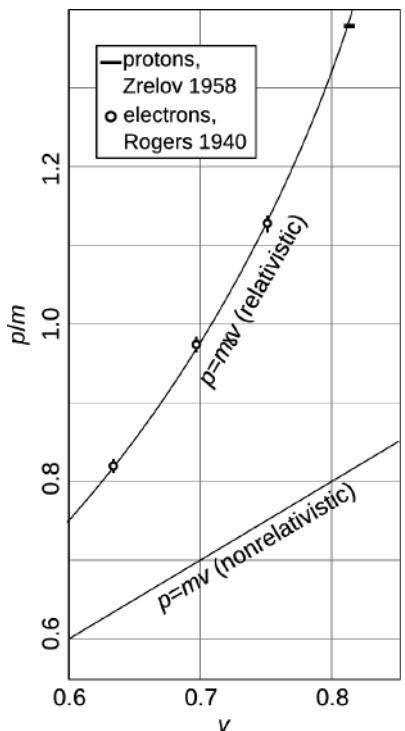
Figure ao shows the results.⁵ According to Newton, an acceleration a_N acting for a time t_N should produce a final velocity $a_N t_N$. The solid line in the graph shows the prediction of Newton's laws, which is that a constant force exerted steadily over time will produce a velocity that rises linearly 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, which are in good agreement with the experimental results.

Figure ap shows experimental data confirming the relativistic equation for momentum.



ao / Example 15.



ap / Two early high-precision tests of the relativistic equation $p = mv$ 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.

⁴Newton's second law gives $a_N = F/m = eE/m$. The constant-acceleration equation $\Delta x = (1/2)at^2$ then gives $t_N = \sqrt{2m\ell_1/eE}$.

⁵To make the low-energy portion of the graph legible, Bertozzi's highest-energy data point is omitted.

17.7.2 Equivalence of mass and energy

Now we're ready to see why mass and energy must be equivalent as claimed in the famous $E = mc^2$. So far we've only considered collisions in which none of the kinetic energy is converted into any other form of energy, such as heat or sound. Let's consider what happens if a blob of putty moving at velocity v hits another blob that is initially at rest, sticking to it. The nonrelativistic result is that to obey conservation of momentum the two blobs must fly off together at $v/2$. Half of the initial kinetic energy has been converted to heat.⁶

Relativistically, however, an interesting thing happens. A hot object has more momentum than a cold object! This is because the relativistically correct expression for momentum is mv/γ , and the more rapidly moving atoms in the hot object have higher values of γ . In our collision, the final combined blob must therefore be moving a little more slowly than the expected $v/2$, since otherwise the final momentum would have been a little greater than the initial momentum. To an observer who believes in conservation of momentum and knows only about the overall motion of the objects and not about their heat content, the low velocity after the collision would seem to be the result of a magical change in the mass, as if the mass of two combined, hot blobs of putty was more than the sum of their individual masses.

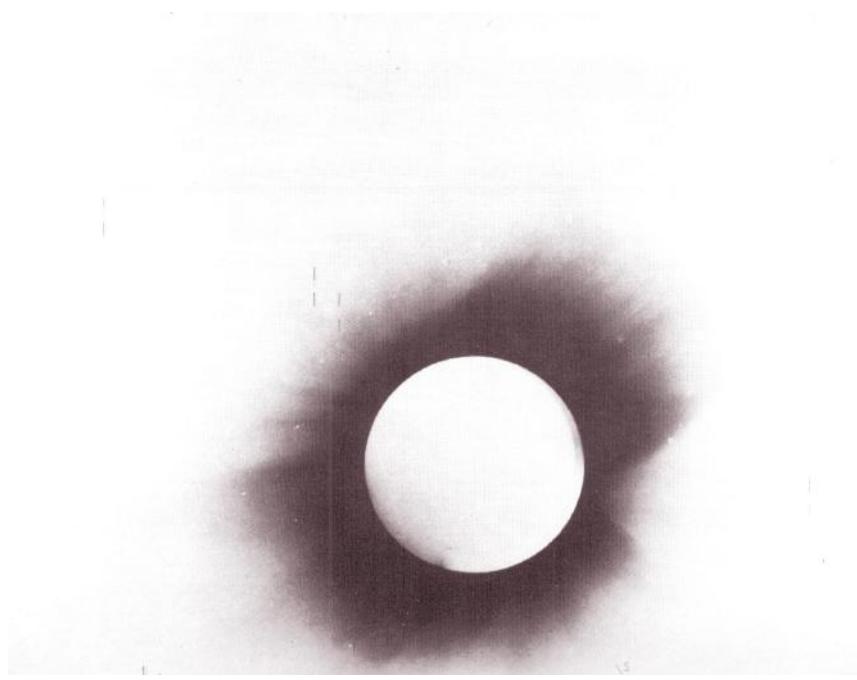
Now we know that the masses of all the atoms in the blobs must be the same as they always were. The change is due to the change in γ with heating, not to a change in mass. The heat energy, however, seems to be acting as if it was equivalent to some extra mass.

But this whole argument was based on the fact that heat is a form of kinetic energy at the atomic level. Would $E = mc^2$ apply to other forms of energy as well? Suppose a rocket ship contains some electrical energy stored in a battery. If we believed that $E = mc^2$ applied to forms of kinetic energy but not to electrical energy, then we would have to believe that the pilot of the rocket could slow the ship down by using the battery to run a heater! This would not only be strange, but it would violate the principle of relativity, 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).

The equation $E = mc^2$ tells us how much energy is equivalent

⁶A double-mass object moving at half the speed does not have the same kinetic energy. Kinetic energy depends on the square of the velocity, so cutting the velocity in half reduces the energy by a factor of $1/4$, which, multiplied by the doubled mass, makes $1/2$ the original energy.

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.



aq / Example 16, page 411.

Gravity bending light

Gravity is a universal attraction between things that have mass, and since the energy in a beam of light is equivalent to some very small amount of mass, we expect that light will be affected by gravity, although the effect should be very small. The first important experimental confirmation of relativity came in 1919 when stars next to the sun during a solar eclipse were observed to have shifted a little from their ordinary position. (If there was no eclipse, the glare of the sun would prevent the stars from being observed.) Starlight had been deflected by the sun's gravity. Figure aq is a photographic negative, so the circle that appears bright is actually the dark face of the moon, and the dark area is really the bright corona of the sun. The stars, marked by lines above and below them, appeared at positions slightly different than their normal ones.

example 16

Black holes

A star with sufficiently strong gravity can prevent light from leaving. Quite a few black holes have been detected via their gravitational forces on neighboring stars or clouds of gas and dust.

example 17

LIGHTS ALL ASKEW IN THE HEAVENS

**Men of Science More or Less
Agog Over Results of Eclipse
Observations.**

EINSTEIN THEORY TRIUMPHS

**Stars Not Where They Seemed
or Were Calculated to be,
but Nobody Need Worry.**

A BOOK FOR 12 WISE MEN

**No More in All the World Could
Comprehend It, Said Einstein When
His Daring Publishers Accepted It.**

ar / A New York Times headline from November 10, 1919, describing the observations discussed in example 16.

You've learned about conservation of mass and conservation of energy, but now we see that they're not even separate conservation laws. As a consequence of the theory of relativity, mass and energy are equivalent, and are not separately conserved — one can be converted into the other. Imagine that a magician waves his wand, and changes a bowl of dirt into a bowl of lettuce. You'd be impressed, because you were expecting that both dirt and lettuce would be conserved quantities. Neither one can be made to vanish, or to appear out of thin air. However, there are processes that can change one into the other. A farmer changes dirt into lettuce, and a compost heap changes lettuce into dirt. At the most fundamental level, lettuce and dirt aren't really different things at all; they're just collections of the same kinds of atoms — carbon, hydrogen, and so on. Because mass and energy are like two different sides of the same coin, we may speak of mass-energy, a single conserved quantity, found by adding up all the mass and energy, with the appropriate conversion factor: $E + mc^2$.

A rusting nail

example 18

- ▷ 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 19

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.

One commonly hears some misinterpretations of $E = mc^2$, one being that the equation tells us how much kinetic energy an object would have if it was moving at the speed of light. This wouldn't make much sense, both because the equation for kinetic energy has $1/2$ in it, $KE = (1/2)mv^2$, and because a material object can't be made to move at the speed of light. However, this naturally leads to the question of just how much mass-energy a moving object has. We know that when the object is at rest, it has no kinetic energy, so its mass-energy is simply equal to the energy-equivalent of its mass, mc^2 ,

$$\mathcal{E} = mc^2 \text{ when } v = 0,$$

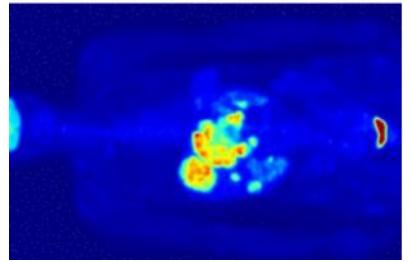
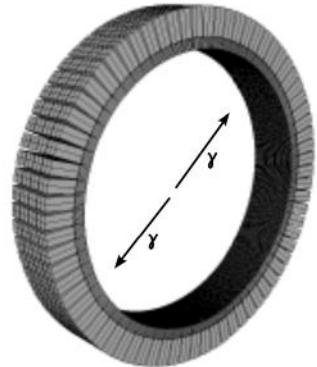
where the symbol \mathcal{E} (cursive “E”) stands for mass-energy. The point of using the new symbol is simply to remind ourselves that we’re talking about relativity, so an object at rest has $\mathcal{E} = mc^2$, not $E = 0$ as we’d assume in nonrelativistic physics.

Suppose we start accelerating the object with a constant force. A constant force means a constant rate of transfer of momentum, but $p = myv$ approaches infinity as v approaches c , so the object will only get closer and closer to the speed of light, but never reach it. Now what about the work being done by the force? The force keeps doing work and doing work, which means that we keep on using up energy. Mass-energy is conserved, so the energy being expended must equal the increase in the object’s mass-energy. We can continue this process for as long as we like, and the amount of mass-energy will increase without limit. We therefore conclude that an object’s mass-energy approaches infinity as its speed approaches the speed of light,

$$\mathcal{E} \rightarrow \infty \text{ when } v \rightarrow c.$$

Now that we have some idea what to expect, what is the actual equation for the mass-energy? As proved in section 17.7.4, it is

$$\mathcal{E} = myc^2.$$



as / 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.

self-check C

Verify that this equation has the two properties we wanted. ▷

Answer, p. 434

「 *KE compared to mc^2 at low speeds*

example 20

▷ An object is moving at ordinary nonrelativistic speeds. Compare its kinetic energy to the energy mc^2 it has purely because of its mass.

▷ The speed of light is a very big number, so mc^2 is a huge number of joules. The object has a gigantic amount of energy because of its mass, and only a relatively small amount of additional kinetic energy because of its motion.

Another way of seeing this is that at low speeds, γ is only a tiny bit greater than 1, so \mathcal{E} is only a tiny bit greater than mc^2 .

「 *The correspondence principle for mass-energy*

example 21

▷ Show that the equation $\mathcal{E} = myc^2$ obeys the correspondence principle.

▷ As we accelerate an object from rest, its mass-energy becomes greater than its resting value. Nonrelativistically, we interpret this excess mass-energy as the object's kinetic energy,

$$\begin{aligned} KE &= \mathcal{E}(v) - \mathcal{E}(v = 0) \\ &= myc^2 - mc^2 \\ &= m(\gamma - 1)c^2. \end{aligned}$$

Expressing γ as $(1 - v^2/c^2)^{-1/2}$ and making use of the approximation $(1 + \epsilon)^p \approx 1 + p\epsilon$ for small ϵ , we have $\gamma \approx 1 + v^2/2c^2$, so

$$\begin{aligned} KE &\approx m\left(1 + \frac{v^2}{2c^2} - 1\right)c^2 \\ &= \frac{1}{2}mv^2, \end{aligned}$$

which is the nonrelativistic expression. As demanded by the correspondence principle, relativity agrees with newtonian physics at speeds that are small compared to the speed of light.

17.7.3 ★ The energy-momentum four-vector

Starting from $\mathcal{E} = my$ and $p = myv$, a little algebra allows one to prove the identity

$$m^2 = \mathcal{E}^2 - p^2.$$

We can define an energy-momentum four-vector,

$$\mathbf{p} = (\mathcal{E}, p_x, p_y, p_z),$$

and the relation $m^2 = \mathcal{E}^2 - p^2$ then arises from the inner product $\mathbf{p} \cdot \mathbf{p}$. Since \mathcal{E} and p are separately conserved, the energy-momentum four-vector is also conserved.

v	Y
0.9870	1.0002(5)
0.9881	1.0012(5)
0.9900	0.9998(5)

A high-precision test of this fundamental relativistic relationship was carried out by Meyer *et al.* in 1963 by studying the motion of electrons in static electric and magnetic fields. They define the quantity

$$Y^2 = \frac{\mathcal{E}^2}{m^2 + p^2},$$

which according to special relativity should equal 1. Their results, tabulated in the sidebar, show excellent agreement with theory.

Energy and momentum of light

example 22

Light has $m = 0$ and $\gamma = \infty$, so if we try to apply $\mathcal{E} = my$ and $p = myv$ to light, or to any massless particle, we get the indeterminate form $0 \cdot \infty$, which can't be evaluated without a delicate and laborious evaluation of limits as in problem 5 on p. 422.

Applying $m^2 = \mathcal{E}^2 - p^2$ yields the same result, $\mathcal{E} = |p|$, much more easily. This example demonstrates that although we encountered the relations $\mathcal{E} = my$ and $p = myv$ first, the identity $m^2 = \mathcal{E}^2 - p^2$ is actually more fundamental.

Figure i on p. 125 shows an experiment that verified $\mathcal{E} = |p|$ empirically.

For the reasons given in example 22, we take $m^2 = \mathcal{E}^2 - p^2$ to be the *definition* of mass in relativity. One thing to be careful about is that this definition is not additive. Suppose that we lump two systems together and call them one big system, adding their mass-energies and momenta. When we do this, the mass of the combination is not the same as the sum of the masses. For example, suppose we have two rays of light moving in opposite directions, with energy-momentum vectors $(\mathcal{E}, \mathcal{E}, 0, 0)$ and $(\mathcal{E}, -\mathcal{E}, 0, 0)$. Adding these gives $(2\mathcal{E}, 0, 0)$, which implies a mass equal to $2\mathcal{E}$. In fact, in the early universe, where the density of light was high, the universe's ambient gravitational fields were mainly those caused by the light it contained.

Mass-energy, not energy, goes in the energy-momentum four-vector

example 23

When we say that something is a four-vector, we mean that it behaves properly under a Lorentz transformation: we can draw such a four-vector on graph paper, and then when we change frames of reference, we should be able to measure the vector in the new frame of reference by using the new version of the graph-paper grid derived from the old one by a Lorentz transformation.

If we had used the energy E rather than the mass-energy \mathcal{E} to construct the energy-momentum four-vector, we wouldn't have gotten a valid four-vector. An easy way to see this is to consider the case where a noninteracting object is at rest in some frame of reference. Its momentum and kinetic energy are both zero. If we'd defined $\mathbf{p} = (E, p_x, p_y, p_z)$ rather than $\mathbf{p} = (\mathcal{E}, p_x, p_y, p_z)$, we

would have had $\mathbf{p} = 0$ in this frame. But when we draw a zero vector, we get a point, and a point remains a point regardless of how we distort the graph paper we use to measure it. That wouldn't have made sense, because in other frames of reference, we have $E \neq 0$.

Metric units

example 24

The relation $m^2 = \mathcal{E}^2 - p^2$ is only valid in relativistic units. If we tried to apply it without modification to numbers expressed in metric units, we would have

$$\text{kg}^2 = \text{kg}^2 \cdot \frac{\text{m}^4}{\text{s}^4} - \text{kg}^2 \cdot \frac{\text{m}^2}{\text{s}^2},$$

which would be nonsense because the three terms all have different units. As usual, we need to insert factors of c to make a metric version, and these factors of c are determined by the need to fix the broken units:

$$m^2 c^4 = \mathcal{E}^2 - p^2 c^2$$

Pair production requires matter

example 25

Example 19 on p. 412 discussed the annihilation of an electron and a positron into two gamma rays, which is an example of turning matter into pure energy. An opposite example is pair production, a process in which a gamma ray disappears, and its energy goes into creating an electron and a positron.

Pair production cannot happen in a vacuum. For example, gamma rays from distant black holes can travel through empty space for thousands of years before being detected on earth, and they don't turn into electron-positron pairs before they can get here. Pair production can only happen in the presence of matter. When lead is used as shielding against gamma rays, one of the ways the gamma rays can be stopped in the lead is by undergoing pair production.

To see why pair production is forbidden in a vacuum, consider the process in the frame of reference in which the electron-positron pair has zero total momentum. In this frame, the gamma ray would have to have had zero momentum, but a gamma ray with zero momentum must have zero energy as well (example 22). This means that conservation of *four*-momentum has been violated: the timelike component of the four-momentum is the mass-energy, and it has increased from 0 in the initial state to at least $2mc^2$ in the final state.

17.7.4 * Proofs

This optional section proves some results claimed earlier.

Ultrarelativistic motion

We start by considering the case of a particle, described as “ultrarelativistic,” that travels at very close to the speed of light. A good way of thinking about such a particle is that it’s one with a very small mass. For example, the subatomic particle called the neutrino has a very small mass, thousands of times smaller than that of the electron. Neutrinos are emitted in radioactive decay, and because the neutrino’s mass is so small, the amount of energy available in these decays is always enough to accelerate it to very close to the speed of light. Nobody has ever succeeded in observing a neutrino that was *not* ultrarelativistic. When a particle’s mass is very small, the mass becomes difficult to measure. For almost 70 years after the neutrino was discovered, its mass was thought to be zero. Similarly, we currently believe that a ray of light has no mass, but it is always possible that its mass will be found to be nonzero at some point in the future. A ray of light can be modeled as an ultrarelativistic particle.

Let’s compare ultrarelativistic particles with train cars. A single car with kinetic energy E has different properties than a train of two cars each with kinetic energy $E/2$. The single car has half the mass and a speed that is greater by a factor of $\sqrt{2}$. But the same is not true for ultrarelativistic particles. Since an idealized ultrarelativistic particle has a mass too small to be detectable in any experiment, we can’t detect the difference between m and $2m$. Furthermore, ultrarelativistic particles move at close to c , so there is no observable difference in speed. Thus we expect that a single ultrarelativistic particle with energy E compared with two such particles, each with energy $E/2$, should have all the same properties as measured by a mechanical detector.

An idealized zero-mass particle also has no frame in which it can be at rest. It always travels at c , and no matter how fast we chase after it, we can never catch up. We can, however, observe it in different frames of reference, and we will find that its energy is different. For example, distant galaxies are receding from us at substantial fractions of c , and when we observe them through a telescope, they appear very dim not just because they are very far away but also because their light has less energy in our frame than in a frame at rest relative to the source. This effect must be such that changing frames of reference according to a specific Lorentz transformation always changes the energy of the particle by a fixed factor, regardless of the particle’s original energy; for if not, then the effect of a Lorentz transformation on a single particle of energy E would be different from its effect on two particles of energy $E/2$.

How does this energy-shift factor depend on the velocity v of the Lorentz transformation? Rather than v , it becomes more convenient to express things in terms of the Doppler shift factor D , which multiplies when we change frames of reference. Let's write $f(D)$ for the energy-shift factor that results from a given Lorentz transformation. Since a Lorentz transformation D_1 followed by a second transformation D_2 is equivalent to a single transformation by D_1D_2 , we must have $f(D_1D_2) = f(D_1)f(D_2)$. This tightly constrains the form of the function f ; it must be something like $f(D) = s^n$, where n is a constant. We postpone until p. 419 the proof that $n = 1$, which is also in agreement with experiments with rays of light.

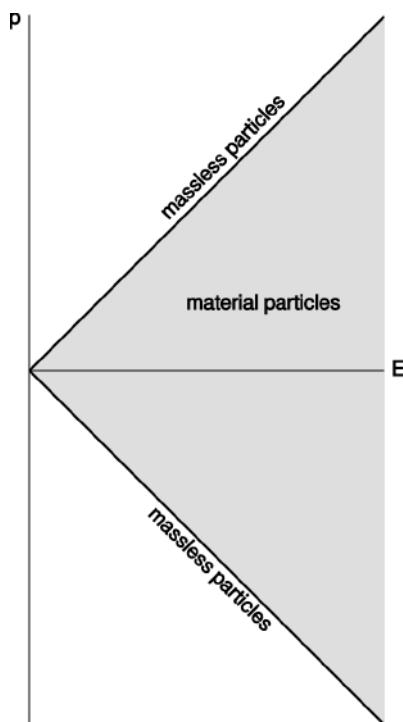
Our final result is that the energy of an ultrarelativistic particle is simply proportional to its Doppler shift factor D . Even in the case where the particle is truly massless, so that D doesn't have any finite value, we can still find how the energy differs according to different observers by finding the D of the Lorentz transformation between the two observers' frames of reference.

Energy

The following argument is due to Einstein. Suppose that a material object O of mass m , initially at rest in a certain frame A, emits two rays of light, each with energy $E/2$. By conservation of energy, the object must have lost an amount of energy equal to E . By symmetry, O remains at rest.

We now switch to a new frame of reference moving at a certain velocity v in the z direction relative to the original frame. We assume that O 's energy is different in this frame, but that the change in its energy amounts to multiplication by some unitless factor x , which depends only on v , since there is nothing else it could depend on that could allow us to form a unitless quantity. In this frame the light rays have energies $ED(v)$ and $ED(-v)$. If conservation of energy is to hold in the new frame as it did in the old, we must have $2xE = ED(v) + ED(-v)$. After some algebra, we find $x = 1/\sqrt{1 - v^2}$, which we recognize as γ . This proves that $E = m\gamma$ for a material object.

Momentum



at / In the p - E plane, massless particles lie on the two diagonals, while particles with mass lie to the right.

We've seen that ultrarelativistic particles are “generic,” in the sense that they have no individual mechanical properties other than an energy and a direction of motion. Therefore the relationship between energy and momentum must be *linear* for ultrarelativistic particles. Indeed, experiments verify that light has momentum, and doubling the energy of a ray of light doubles its momentum rather than quadrupling it. On a graph of p versus E , massless particles, which have $E \propto |p|$, lie on two diagonal lines that connect at the origin. If we like, we can pick units such that the slopes of these lines are plus and minus one. Material particles lie to the right of

these lines. For example, a car sitting in a parking lot has $p = 0$ and $E = mc^2$.

Now what happens to such a graph when we change to a different frame or reference that is in motion relative to the original frame? A massless particle still has to act like a massless particle, so the diagonals are simply stretched or contracted along their own lengths. In fact the transformation must be linear (p. 383), because conservation of energy and momentum involve addition, and we need these laws to be valid in all frames of reference. By the same reasoning as in figure p on p. 385, the transformation must be area-preserving. We then have the same three cases to consider as in figure m on p. 384. Case I is ruled out because it would imply that particles keep the same energy when we change frames. (This is what would happen if c were infinite, so that the mass-equivalent E/c^2 of a given energy was zero, and therefore E would be interpreted purely as the mass.) Case II can't be right because it doesn't preserve the $E = |p|$ diagonals. We are left with case III, which establishes the fact that the p - E plane transforms according to exactly the same kind of Lorentz transformation as the x - t plane. That is, (E, p_x, p_y, p_z) is a four-vector.

The only remaining issue to settle is whether the choice of units that gives invariant 45-degree diagonals in the x - t plane is the same as the choice of units that gives such diagonals in the p - E plane. That is, we need to establish that the c that applies to x and t is equal to the c' needed for p and E , i.e., that the velocity scales of the two graphs are matched up. This is true because in the Newtonian limit, the total mass-energy E is essentially just the particle's mass, and then $p/E \approx p/m \approx v$. This establishes that the velocity scales are matched at small velocities, which implies that they coincide for all velocities, since a large velocity, even one approaching c , can be built up from many small increments. (This also establishes that the exponent n defined on p. 418 equals 1 as claimed.)

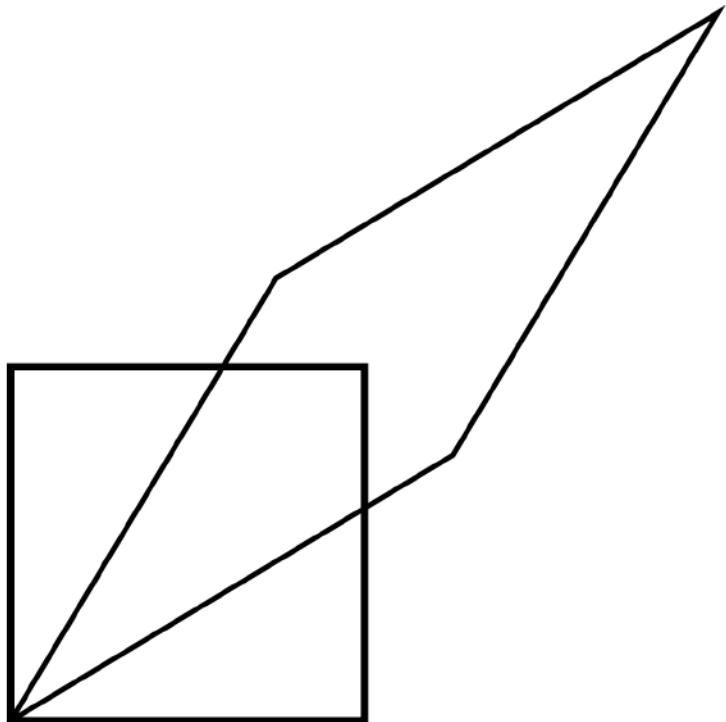
Since $m^2 = E^2 - p^2$, it follows that for a material particle, $p = m\gamma v$.

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 17.2. 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

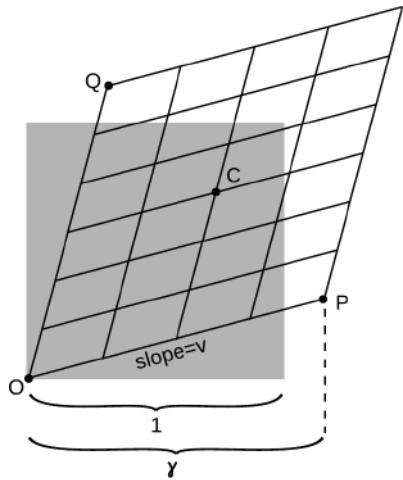
In example 5 on page 388, I remarked that accelerating a macroscopic (i.e., not microscopic) object to close to the speed of light would require an unreasonable amount of energy. Suppose that the starship Enterprise from Star Trek has a mass of 8.0×10^7 kg, about the same as the Queen Elizabeth 2. Compute the kinetic energy it would have to have if it was moving at half the speed of light. Compare with the total energy content of the world's nuclear arsenals, which is about 10^{21} J.

✓

- 3** In this homework problem, you'll fill in the steps of the algebra required in order to find the equation for γ on page 385. To keep the algebra simple, let the time t in figure q equal 1, as suggested in the figure accompanying this homework problem. The original square then has an area of 1, and the transformed parallelogram must also have an area of 1. (a) Prove that point P is at $x = v\gamma$, so that its (t, x) coordinates are $(1, v\gamma)$. (b) Find the (t, x) coordinates of point Q. (c) Find the length of the short diagonal connecting P and Q. (d) Average the coordinates of P and Q to find the coordinates of the midpoint C of the parallelogram, and then find distance OC. (e) Find the area of the parallelogram by computing twice the area of triangle PQO. [Hint: You can take PQ to be the base of the triangle.] (f) Set this area equal to 1 and solve for γ to prove $\gamma = 1/\sqrt{1 - v^2}$.

✓

- 4** (a) Find a relativistic equation for the velocity of an object in terms of its mass and momentum (eliminating γ). 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 3.

5 An object moving at a speed very close to the speed of light is referred to as ultrarelativistic. Ordinarily (luckily) the only ultrarelativistic objects in our universe are subatomic particles, such as cosmic rays or particles that have been accelerated in a particle accelerator.

- (a) What kind of number is γ for an ultrarelativistic particle?
- (b) Repeat example 20 on page 414, but instead of very low, non-relativistic speeds, consider ultrarelativistic speeds.
- (c) Find an equation for the ratio \mathcal{E}/p . The speed may be relativistic, but don't assume that it's ultrarelativistic. ✓
- (d) Simplify your answer to part c for the case where the speed is ultrarelativistic. ✓
- (e) We can think of a beam of light as an ultrarelativistic object — it certainly moves at a speed that's sufficiently close to the speed of light! Suppose you turn on a one-watt flashlight, leave it on for one second, and then turn it off. Compute the momentum of the recoiling flashlight, in units of $\text{kg}\cdot\text{m/s}$. ✓
- (f) Discuss how part e relates to the correspondence principle.

6 (a) A charged particle is surrounded by a uniform electric field. Starting from rest, it is accelerated by the field to speed v after traveling a distance d . Now it is allowed to continue for a further distance $3d$, for a total displacement from the start of $4d$. What speed will it reach, assuming newtonian physics?

- (b) Find the relativistic result for the case of $v = c/2$.

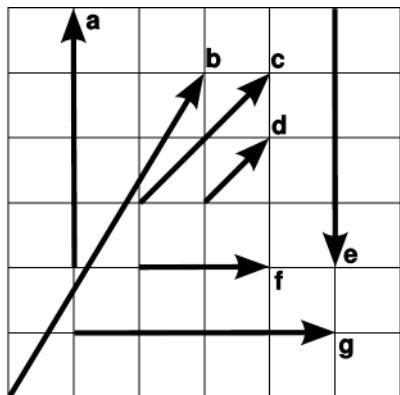
7 Problem 7 has been deleted.

8 Expand the equation $K = m(\gamma - 1)$ 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 nonrelativistic expression for kinetic energy.

9 Consider the relativistic relation for momentum as a function of velocity (for a particle with nonzero mass). 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.

10 The figure shows seven four-vectors, represented in a two-dimensional plot of x versus t . All the vectors have y and z components that are zero. Which of these vectors are congruent to others, i.e., which represent spacetime intervals that are equal to one another? If you reason based on Euclidean geometry, you will get the wrong answers. ▷ Solution, p. 431

11 Four-vectors can be timelike, lightlike, or spacelike. What can you say about the inherent properties of particles whose momentum four-vectors fall in these various categories?



Problem 10.

12 The following are the three most common ways in which gamma rays interact with matter:

Photoelectric effect: The gamma ray hits an electron, is annihilated, and gives all of its energy to the electron.

Compton scattering: The gamma ray bounces off of an electron, exiting in some direction with some amount of energy.

Pair production: The gamma ray is annihilated, creating an electron and a positron.

Example 25 on p. 416 shows that pair production can't occur in a vacuum due to conservation of the energy-momentum four-vector. What about the other two processes? Can the photoelectric effect occur without the presence of some third particle such as an atomic nucleus? Can Compton scattering happen without a third particle?

13 Expand the relativistic equation for the longitudinal Doppler shift of light $D(v)$ in a Taylor series, and find the first two nonvanishing terms. Show that these two terms agree with the nonrelativistic expression, so that any relativistic effect is of higher order in v .

14 (a) In this chapter we've represented Lorentz transformations as distortions of a square into various parallelograms, with the degree of distortion depending on the velocity of one frame of reference relative to another. Suppose that one frame of reference was moving at c relative to another. Discuss what would happen in terms of distortion of a square, and show that this is impossible by using an argument similar to the one used to rule out transformations like the one in figure k on page 383.

(b) Resolve the following paradox. Two pen-pointer lasers are placed side by side and aimed in parallel directions. Their beams both travel at c relative to the hardware, but each beam has a velocity of zero relative to the neighboring beam. But the speed of light can't be zero; it's supposed to be the same in all frames of reference.

15 The products of a certain radioactive decay are a massive particle and a gamma ray, which is massless. See example 22 on p. 415 for a discussion of the energy and momentum of a gamma ray. (a) Show that, in the center of mass frame, the energy of the gamma is less than the mass-energy of the massive particle.

(b) Show that the opposite inequality holds if we compare the *kinetic* energy of the massive particle to the energy of the gamma. [Problem by B. Shotwell.]

16 Natural relativistic units were introduced on p. 386, and examples 1 and 2 on pp. 387 and 387 gave examples of how to convert an equation from natural units to SI units. In example 4 on p. 388, we derived the approximation

$$\gamma \approx 1 + \frac{v^2}{2}$$

for values of v that are small compared to 1 (i.e., small compared to the speed of light in natural units). As in the other examples, convert this equation to SI units. ▷ Solution, p. 431

17 We want to throw a ball of diameter b through a hole of diameter h in a thin wall. Clearly this is possible if $b < h$, but consider the case where $b > h$. If the motion is relativistic, then is it unambiguous whether the ball fits through the hole, or is this frame-dependent, as in example 7 on p. 390? If the former, then is there some velocity v that is required, expressible in terms of b and h ?

18 (a) Let L be the diameter of our galaxy. Suppose that a person in a spaceship of mass m wants to travel across the galaxy at constant speed, taking proper time τ . Find the kinetic energy of the spaceship. (b) Your friend is impatient, and wants to make the voyage in an hour. For $L = 10^5$ light years, estimate the energy in units of megatons of TNT (1 megaton = 4×10^9 J).

Hints

Hints for chapter 1

Page 33, problem 7:

You can use the geometric interpretation of the dot product.

Hints for chapter 2

Page 69, problem 9:

There are many possible ways of approaching the approximation in part b, either with or without calculus. Most scientists and engineers would probably proceed by constructing a small, unitless parameter out of ℓ and r , then expanding the expression as a Taylor series in that parameter.

Hints for chapter 3

Page 83, problem 7:

The force on the lithium ion is the vector sum of all the forces of all the quadrillions of sodium and chlorine atoms, which would obviously be too laborious to calculate. Nearly all of these forces, however, are canceled by a force from an ion on the opposite side of the lithium.

Hints for chapter 4

Page 107, problem 13:

Since we have $t \ll r$, the volume of the membrane is essentially the same as if it was unrolled and flattened out, and the field's magnitude is nearly constant.

Page 110, problem 22:

The approach is similar to the one used for the other problem, but you want to work with potential and electrical energy rather than force.

Hints for chapter 10

Page 252, problem 9:

First find the energy stored in a spherical shell extending from r to $r + dr$, then integrate to find the total energy.

Page 252, problem 12:

The math is messy if you put the origin of your polar coordinates at the center of the disk. It comes out much simpler if you put the origin at the edge, right on top of the point at which we're trying to compute the voltage.

Hints for chapter 11

Page 280, problem 4:

The analysis gets simpler if you imagine the gears with their axes coinciding rather than parallel. This is impossible mechanically, but easier electrically.

Hints for chapter 12

Page 306, problem 8:

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 15

Page 358, problem 4:

Use Faraday's law, and choose an Ampèrian surface that is a disk of radius R sandwiched

between the plates.

Solutions to selected problems

Solutions for chapter 1

Page 33, problem 6:

$\mathbf{E} \times \mathbf{E}$ wouldn't make sense because it's a vector, whereas energy is a scalar. It also wouldn't make sense because the vector cross product of any two collinear vectors is zero.

Page 36, problem 17:

Statement 1 is false because if one or both of the vectors is zero, then their dot product is zero, but the angle between them is undefined. Statement 2 is false for the same reason, and also because the two vectors could be in opposite directions.

Page 36, problem 18:

The method using the cross product is by far the easier way. The area is simply $(1/2)|\mathbf{u} \times \mathbf{v}|$. This just requires taking one vector cross product and then computing its magnitude.

To do this using only vector dot products, you would first have to find the angle between the two vectors from $\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos \theta$. Then you would find the area as $(1/2)|\mathbf{u}||\mathbf{v}| \sin \theta$. This requires far more computation.

Solutions for chapter 2

Page 67, problem 1:

(a) There are positive charges on the surfaces where the field lines begin: on the top of the box and the bottom of the sphere. There are negative charges where the field lines terminate: on the bottom of the box and the top of the sphere. (b) The field lines are more dense at the star, so the electric field is stronger there. There is no charge at these points, because no field lines begin or end there. (c) The total charge appears to be exactly zero, because of the symmetry of the charge distribution with respect to reflection across the horizontal mid-plane.

Page 69, problem 10:

(a) Let the Gaussian pillbox of example 9, p. 59, have area A on each flat face. Gauss's law is

$$4\pi kq_{in} = \sum \mathbf{E}_j \cdot \mathbf{A}_j$$
$$4\pi k\sigma A = \sum |\mathbf{E}_j||\mathbf{A}_j|.$$

By symmetry, we would expect the field to be perpendicular to the plane and the same on both sides of the plane. This is a rather subtle point, which we'll come back to at the end. Assuming this to be true, there is no flux through the sides of the pillbox, only through the flat faces, and the magnitude of the field is the same everywhere on the faces, so we can take it outside the sum.

$$4\pi k\sigma A = |\mathbf{E}| \cdot 2A$$
$$E = 2\pi k\sigma.$$

Returning to the symmetry issue, the tricky point is that we could add on to our solution any other solution that was valid in a vacuum. For instance, we could add a uniform electric field that would cancel out the sheet's field on one side, while doubling it on the other.

(b) The most efficient way to do this is to make use of a similar equation whose units we already know to be correct, such as the equation $E = kq/r^2$ for a point charge. Since q/r^2 has the same units as σ , the units do check out.

Solutions for chapter 4

Page 103, problem 1:

The work done against gravity is equal to the change in gravitational potential energy, which is the same in the two cases if the person's body weight is the same. Stating the number in units of J/kg rather than J would be nicer, because then the number would be a property of the paths themselves, regardless of who was climbing along them. In the electrical analogy, charge plays the role of mass, so the units would be J/C.

Page 103, problem 3:

By symmetry, the field is always directly toward or away from the center. We can therefore calculate it along the x axis, where $r = x$, and the result will be valid for any location at that distance from the center. The electric field is minus the derivative of the potential,

$$\begin{aligned} E &= -\frac{d\phi}{dx} \\ &= -\frac{d}{dx}(x^{-1}e^{-x}) \\ &= x^{-2}e^{-x} + x^{-1}e^{-x} \end{aligned}$$

At small x , near the proton, the first term dominates, and the exponential is essentially 1, so we have $E \propto x^{-2}$, as we expect from the Coulomb force law. At large x , the second term dominates, and the field approaches zero faster than an exponential.

Page 104, problem 5:

In figure 1, the voltmeter is reading the difference in electrical potential between the battery's two terminals: the potential of the positive nipple minus the potential of the negative flat end. The meter is set up to read in millivolts, and the reading is equivalent to about 1.3 V, which is reasonable for a nominally 1.5 V battery. In 2, the wires have been reversed, so the sign of the potential difference is reversed. In 3, only one probe is connected. There is no way to define an absolute voltage, only a voltage difference between two points in space, so the meter can't possibly give a meaningful result here. The way the voltage measurement is implemented in this particular meter causes it to read zero in this situation, but the result is not really zero, it's an error.

Page 106, problem 9:

The field may not be exactly constant, but roughly speaking, we have $\Delta\phi = E\Delta x$, where Δx is the gap and $\Delta\phi$ is the voltage difference that has to be applied. Because E is big (a typical number was given in example 4 on p. 52), the voltage difference would tend to be big. Making the gap smaller reduces the voltage difference that is required.

Page 109, problem 19:

(a) Conservation of energy gives

$$\begin{aligned}U_A &= U_B + K_B \\K_B &= U_A - U_B \\\frac{1}{2}mv^2 &= e\Delta V \\v &= \sqrt{\frac{2e\Delta V}{m}}\end{aligned}$$

(b) Plugging in numbers, we get 5.9×10^7 m/s. This is about 20% of the speed of light.

Solutions for chapter 5

Page 143, problem 2:

(a) The region between the spheres contains an intense magnetic field (field lines close together), which is because both spheres' contribution to the field are downward, causing them to reinforce and become stronger. This strong field is oriented in such a way that there is a horizontal pressure. There is pressure at the outside edges of the spheres as well, but the field is weaker there (because of partial cancellation of the fields), so the pressure there is not as strong. The result is that the balls make forces on each other that have repulsive horizontal components. The vertical forces cancel by symmetry, because they can't depend on whether we flip the arrowheads or not.

(b) This is similar to the first example, but now the orientation is such that there is tension between the balls. The force is attractive.

Page 146, problem 15:

We have $D = q\ell$ and $F_x = q\ell b = Db$. Since b is the same thing, in this example, as $\partial E_x / \partial x$, our equation for F is consistent with the result involving a derivative on p. 136 and also, as claimed, depends on q and ℓ only via D .

Page 149, problem 22:

To apply the right-hand rule to this case, flip your hand so that you're looking at your palm, with your thumb pointing to the right. Curling your fingers traces a circule oriented in the direction such that the field at P comes out of the page.

Solutions for chapter 6

Page 169, problem 3:

(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 169, problem 5:

Let the wave be propagating in the $+x$ direction, and let the fields be parallel to the y axis. Then sticking a paddlewheel into the wave to measure $\text{curl } \mathbf{E}$ will make the wheel spin if its axle is along the z axis, but the paddlewheel won't spin at all if its axle is along y . That is, $\text{curl } \mathbf{E}$ has no y component. Maxwell's equations then require that $\partial \mathbf{B} / \partial t$ have no y component. But this is false, because \mathbf{B} is in the y direction and varying as a function of time.

Page 170, problem 6:

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 8

Page 203, problem 1:

$$\Delta t = \Delta q/I = e/I = 0.160 \text{ } \mu\text{s}$$

Page 203, problem 6:

It's much more practical to measure voltage differences. To measure a current, you have to break the circuit somewhere and insert the meter there, but it's not possible to disconnect the circuits sealed inside the board.

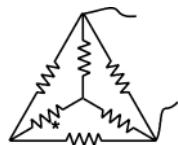
Solutions for chapter 9

Page 224, problem 10:

In series, they give $11 \text{ k}\Omega$. In parallel, they give $(1/1 \text{ k}\Omega + 1/10 \text{ k}\Omega)^{-1} = 0.9 \text{ k}\Omega$.

Page 225, problem 15:

The actual shape is irrelevant; all we care about is what's connected to what. Therefore, we can draw the circuit flattened into a plane. Every vertex of the tetrahedron is adjacent to every other vertex, so any two vertices to which we connect will give the same resistance. Picking two arbitrarily, we have this:



This is unfortunately a circuit that cannot be converted into parallel and series parts, and that's what makes this a hard problem! However, we can recognize that by symmetry, there is zero current in the resistor marked with an asterisk. Eliminating this one, we recognize the whole arrangement as a triple parallel circuit consisting of resistances R , $2R$, and $2R$. The resulting resistance is $R/2$.

Solutions for chapter 10

Page 251, problem 2:

The moment of inertia is $I = \int r^2 dm$. Let the ring have total mass M and radius b . The proportionality

$$\frac{M}{2\pi} = \frac{dm}{d\theta}$$

gives a change of variable that results in

$$I = \frac{M}{2\pi} \int_0^{2\pi} r^2 d\theta.$$

If we measure θ from the axis of rotation, then $r = b \sin \theta$, so this becomes

$$I = \frac{Mb^2}{2\pi} \int_0^{2\pi} \sin^2 \theta \, d\theta.$$

The integrand averages to $1/2$ over the 2π range of integration, so the integral equals π . We therefore have $I = \frac{1}{2}Mb^2$. This is, as claimed, half the value for rotation about the symmetry axis.

Solutions for chapter 11

Page 283, problem 12:

Note that in the Biot-Savart law, the variable \mathbf{r} is defined as a vector that points from the current to the point at which the field is being calculated, whereas in the polar coordinates used to express the equation of the spiral, the vector more naturally points the opposite way. This requires some fiddling with signs, which I'll suppress, and simply identify $d\ell$ with $d\mathbf{r}$.

$$\mathbf{B} = \frac{kI}{c^2} \int \frac{d\ell \times \mathbf{r}}{r^3}$$

The vector $d\mathbf{r}$ has components $dx = w(\cos \theta - \theta \sin \theta)$ and $dy = w(\sin \theta + \theta \cos \theta)$. Evaluating the vector cross product, and substituting θ/w for r , we find

$$\begin{aligned} \mathbf{B} &= \frac{kI}{c^2 w} \int \frac{\theta(\cos \theta \sin \theta - \theta \sin^2 \theta - \cos \theta \sin \theta - \theta \cos^2 \theta) \, d\theta}{\theta^3} \\ &= \frac{kI}{c^2 w} \int \frac{d\theta}{\theta} \\ &= \frac{kI}{c^2 w} \ln \frac{\theta_2}{\theta_1} \\ &= \frac{kI}{c^2 w} \ln \frac{b}{a} \end{aligned}$$

Solutions for chapter 12

Page 305, problem 2:

$$\begin{aligned} \sin(a+b) &= \left(e^{i(a+b)} - e^{-i(a+b)} \right) / 2i \\ &= \left(e^{ia} e^{ib} - e^{-ia} e^{-ib} \right) / 2i \\ &= [(\cos a + i \sin a)(\cos b + i \sin b) - (\cos a - i \sin a)(\cos b - i \sin b)] / 2i \\ &= \cos a \sin b + \sin a \cos b \end{aligned}$$

By a similar computation, we find $\cos(a+b) = \cos a \cos b - \sin a \sin b$.

Page 305, problem 3:

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 306, problem 5:

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 $1(2i)^4 = 16$. Satisfying the differential equation then amounts to having $16 - 16i = 0$, which is true.

Solutions for chapter 13

Page 325, problem 2:

By definition, the henry is 1 J/A^2 . The units of L/R are then

$$\frac{\text{J}/\text{A}^2}{\text{V}/\text{A}} = \frac{\text{J}}{\text{V}\cdot\text{A}} = \frac{\text{J}}{\text{J}/\text{s}} = \text{s}$$

Page 327, problem 9:

The resonant frequency is proportional to $L^{-1/2}$, so increasing L by a factor of 100 will change the resonant frequency by a factor of $1/10$.

Solutions for chapter 14

Page 343, problem 9:

The two examples are both in series. Impedances in series add. The order in which we add complex numbers doesn't affect the result. (Addition of complex numbers is "commutative," just as it is for the real numbers.) Therefore the two impedances are the same.

Solutions for chapter 17

Page 422, problem 10:

Among the spacelike vectors, **a** and **e** are clearly congruent, because they're the same except for a rotation in space; this is the same as the definition of congruence in ordinary Euclidean geometry, where rotation doesn't matter. Vector **b** is also congruent to these, since it represents an interval $3^2 - 5^2 = -4^2$, just like the other two.

The lightlike vectors **c** and **d** both represent intervals of zero, so they're congruent, even though **c** is a double-scale version of **d**.

The timelike vectors **f** and **g** are not congruent to each other or to any of the others; **f** represents an interval of 2^2 , while **g**'s interval is 4^2 .

Page 423, problem 16:

To make the units make sense, we need to make sure that both sides of the \approx sign have the same units, and also that both terms on the right-hand side have the same units. Everything is unitless except for the second term on the right, so we add a factor of c^{-2} to fix it:

$$\gamma \approx 1 + \frac{v^2}{2c^2}.$$

Answers to self-checks

Answers to self-checks for chapter 5

Page 122, self-check A:

We would expect the field to be stronger closer to the wire, and in fact example 2 shows that

$B \propto 1/r$. The spacing of the magnetic field lines reflects this: more closely spaced field lines show a stronger field.

Page 136, self-check B:

If the dipole had been oriented horizontally, then the two attractive forces and the two repulsive forces from the fixed charges would have all ended up having zero total vertical component. (This is easiest to see based on symmetry.) The actual orientation of that dipole was such that its negative charge was closer to the fixed positive charge on top, and its positive charge was closer to the fixed positive below. This caused the net force on it from the upper fixed charge to be upward, and also caused the net force from the lower fixed charge to be upward.

Answers to self-checks for chapter 8

Page 198, self-check A:

The large amount of power means a high rate of conversion of the battery's chemical energy into heat. The battery will quickly use up all its energy, i.e., "burn out."

Answers to self-checks for chapter 10

Page 235, self-check A:

The two-dimensional idea is actually unimportant, so the sum can be simplified to

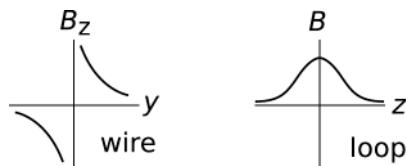
$$\sum_{m=0}^{63} 2^m.$$

In binary, this is 1111111...1111111, a string of 64 ones, which is the same as $2^{64} - 1$, or approximately $(2^{10})^6 \cdot 2^4 \approx (1000)^6 \cdot 16 = 1.6 \times 10^{19}$, which is certainly more grains of wheat than have ever existed in the world.

Answers to self-checks for chapter 11

Page 273, self-check A:

1. The field pattern consists of circular loops in planes perpendicular to the wire. This means that on the y axis, it's the z component that's nonzero. Using the right-hand rule, and assuming a right-handed coordinate system, we find that the field has a positive z component for positive y .
2. The expression looks like $(\dots + z^2)^{-3/2}$. Because the exponent is negative, the function is maximized when the quantity inside the parentheses is minimized, at $z = 0$. We get a bell-shaped curve.



Answers to self-checks for chapter 12

Page 293, self-check A:

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 297, self-check B:

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 300, self-check C:

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 13

Page 314, self-check A:

Yes. The mass has the same kinetic energy regardless of which direction it's moving. Friction converts mechanical energy into heat at the same rate whether the mass is sliding to the right or to the left. The spring has an equilibrium length, and energy can be stored in it either by compressing it ($x < 0$) or stretching it ($x > 0$).

Answers to self-checks for chapter 14

Page 332, self-check A:

The impedance depends on the frequency at which the capacitor is being driven. It isn't just a single value for a particular capacitor.

Answers to self-checks for chapter 15

Page 350, self-check A:

The circulation around the Ampèrian surface we used was counterclockwise, since the field on the bottom was to the right. Applying the right-hand rule, the current I_{through} must have been out of the page at the top of the solenoid, and into the page at the bottom.

Answers to self-checks for chapter 16

Page 364, self-check A:

An (idealized) battery is a circuit element that always maintains the same voltage difference across itself, so by the loop rule, the voltage difference across the capacitor must remain unchanged, even while the dielectric is being withdrawn. The bound charges on the surfaces of the dielectric have been attracting the free charges in the plates, causing them to charge up more than they ordinarily would have. As the dielectric is withdrawn, the capacitor will be partially discharged, and we will observe a current in the ammeter. Since the dielectric is attracted to the plates, positive work is done in extracting it, indicating that there must be an increase in the electrical energy stored in the capacitor. This may seem paradoxical, since the energy stored in a capacitor is $(1/2)CV^2$, and we are decreasing the capacitance. However, the energy $(1/2)CV^2$ is calculated in terms of the work required to deposit the free charge on the plates. In addition to this energy, there is also energy stored in the dielectric itself. By moving its bound charges farther away from the free charges in the plates, to which they are attracted, we have increased their electrical energy. This energy of the bound charges is inaccessible to the electric circuit.

Answers to self-checks for chapter 17

Page 387, self-check A:

At $v = 0$, we get $\gamma = 1$, so $t = T$. There is no time distortion unless the two frames of reference are in relative motion.

Page 407, self-check B:

The total momentum is zero before the collision. After the collision, the two momenta have reversed their directions, but they still cancel. Neither object has changed its kinetic energy, so the total energy before and after the collision is also the same.

Page 414, self-check C:

At $v = 0$, we have $\gamma = 1$, so the mass-energy is mc^2 as claimed. As v approaches c , γ approaches infinity, so the mass energy becomes infinite as well.

Answers

Mathematical Review

Algebra

Quadratic equation:

The solutions of $ax^2 + bx + c = 0$
are $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

Logarithms and exponentials:

$$\ln(ab) = \ln a + \ln b$$

$$e^{a+b} = e^a e^b$$

$$\ln e^x = e^{\ln x} = x$$

$$\ln(a^b) = b \ln a$$

$$\frac{d}{dx}(cf) = c \frac{df}{dx}$$

$$\frac{d}{dx}(f + g) = \frac{df}{dx} + \frac{dg}{dx}$$

The chain rule:

$$\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$$

Geometry, area, and volume

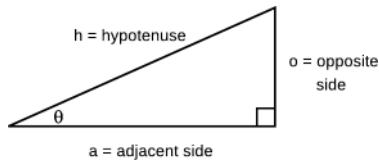
area of a triangle of base b and height h
circumference of a circle of radius r
area of a circle of radius r
surface area of a sphere of radius r
volume of a sphere of radius r

$$\begin{aligned} &= \frac{1}{2}bh \\ &= 2\pi r \\ &= \pi r^2 \\ &= 4\pi r^2 \\ &= \frac{4}{3}\pi r^3 \end{aligned}$$

$$\frac{d}{dx}(fg) = \frac{df}{dx}g + \frac{dg}{dx}f$$

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{f'}{g} - \frac{fg'}{g^2}$$

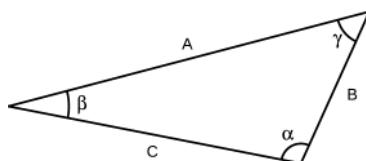
Trigonometry with a right triangle



$$\sin \theta = o/h \quad \cos \theta = a/h \quad \tan \theta = o/a$$

Pythagorean theorem: $h^2 = a^2 + o^2$

Trigonometry with any triangle



Law of Sines:

$$\frac{\sin \alpha}{A} = \frac{\sin \beta}{B} = \frac{\sin \gamma}{C}$$

Law of Cosines:

$$C^2 = A^2 + B^2 - 2AB \cos \gamma$$

Properties of the derivative and integral

Let f and g be functions of x , and let c be a constant.

Linearity of the derivative:

$$\frac{d}{dx}(cf) = c \frac{df}{dx}$$

$$\frac{d}{dx}(f + g) = \frac{df}{dx} + \frac{dg}{dx}$$

The chain rule:

Derivatives of products and quotients:

Some derivatives:

$$\begin{aligned} \frac{d}{dx}x^m &= mx^{m-1}, \text{ except for } m = 0 \\ \frac{d}{dx}\sin x &= \cos x & \frac{d}{dx}\cos x &= -\sin x \\ \frac{d}{dx}e^x &= e^x & \frac{d}{dx}\ln x &= \frac{1}{x} \end{aligned}$$

Linearity of the integral:

$$\int cf(x) dx = c \int f(x) dx$$

$$\int [f(x) + g(x)] dx = \int f(x) dx + \int g(x) dx$$

The fundamental theorem of calculus:

The derivative and the integral undo each other, in the following sense:

$$\int_a^b f'(x) dx = f(b) - f(a)$$

Approximations to Exponents and Logarithms

It is often useful to have certain approximations involving exponents and logarithms. As a simple numerical example, suppose that your bank balance grows by 1% for two years in a row. Then the result of compound interest is growth by a factor of $1.01^2 = 1.0201$, but the compounding effect is quite small, and the result is essentially 2% growth. That is, $1.01^2 \approx 1.02$. This is a special case of the more general approximation

$$(1 + \epsilon)^p \approx 1 + p\epsilon,$$

which holds for small values of ϵ and is used in example 4 on p. 388 relating to relativity. Proof: Any real exponent p can be approximated to the desired precision as $p = a/b$, where a and b are integers. Let $(1+\epsilon)^p = 1+x$. Then $(1+\epsilon)^a = (1+x)^b$. Multiplying out both sides gives $1 + a\epsilon + \dots = 1 + bx + \dots$, where \dots indicates higher powers. Neglecting these higher powers gives $x \approx (a/b)\epsilon \approx p\epsilon$.

We have considered an approximation that can be found by restricting the *base* of an exponential to be close to 1. It is often of interest as well to consider the case where the *exponent* is restricted to be small. Consider the base- e case. One way of defining e is that when we use it as a base, the rate of growth of the function e^x , for small x , equals 1. That is,

$$e^x \approx 1 + x$$

for small x . This can easily be generalized to other bases, since $a^x = e^{\ln(a^x)} = e^{x \ln a}$, giving

$$a^x \approx 1 + x \ln a.$$

Finally, since $e^x \approx 1 + x$, we also have

$$\ln(1 + x) \approx x.$$

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.

Cover Crystal: Hans-Joachim Engelhardt, CC-BY-SA licensed.

15 *Coronal mass ejection*: NASA, CC-BY licensed. **16** *Ripples*: Scott Robinson, CC-BY. **18** *Man on phone*: Ian Camp, CC-BY-SA. **18** *Compass*: Wikimedia Commons user Arpingstone, public domain. **18** *Photovoltaics*: Klaus Holl, CC-BY-SA. **18** *Multimeter*: Wikimedia Commons user Binarysequence, CC-BY-SA. **18** *Geiger counter*: Wikipedia user Changlc, public domain. **18** *X-ray*: Roentgen, ca. 1895. **19** *Girl*: Georges Hébert, L'éducation Physique féminine, 1921, public domain. **19** *M100*: European Southern Observatory, CC-BY-SA. **19** *Saturn*: NASA, public domain. **19** *Converging rays*: Wikipedia user Fir0002, CC-BY-SA. **26** *Iron filings around bar magnet*: Windell Oskay, CC-BY. **26** *Faraday*: Painting by Thomas Phillips, 1842. **27** *Bamboo*: Max Pixel, maxpixel.net, CC0 license. **28** *Coulomb*: Louis Hierle, 1894. **44** *New York City*: Petr Kratochvil, publicdomainpictures.net, CC0 license. **46** *Duel*: Ilya Repin, 1899. **48** *New York City*: Petr Kratochvil, publicdomainpictures.net, CC0 license. **49** *Snowplow*: Antti Leppanen, CC-BY-SA. **52** *Spark plug*: Norris Wong, CC-BY. **58** *Coaxial cable*: Wikimedia Commons user FDominec, CC-BY-SA licensed. **60** *Variable capacitor*: Wikimedia Commons user Solaris2006, CC-BY-SA licensed. **61** *Roman brick arch*: Heinz-Josef Lucking, CC-BY-SA. **61** *Cell phone*: Rafael Fernandez, CC-BY-SA. **63** *Earth*: NASA, Apollo 17. Public domain. **67** *Weather map of Ireland*: Janjic et al., University College Dublin, CC-BY. **73** *Photo of salt crystal*: Hans-Joachim Engelhardt, CC-BY-SA. **73** *Diagram of salt crystal's atomic structure*: Wikimedia Commons user Benjah-bmm27, public domain. **74** *Molecule*: Wikimedia Commons user Benjah-bmm27, public domain. **87** *Electric quadrupole field*: Wikimedia Commons user Geek3, CC-BY-SA. **90** *Multimeter*: Wikimedia Commons user Binarysequence, CC-BY-SA. **90** *Using multimeter*: US Navy/Robert Winn, public domain. **91** *Europe*: Julio Reis, CC-BY-SA. **94** *Skateboarder on top of pipe*: Oula Lehtinen, Wikimedia Commons, CC-BY-SA. **94** *Skater in pool*: Courtesy of J.D. Rogge. **97** *Topographic map*: USGS, public domain. **106** *Wikimedia Commons user Peeperman*: CC-BY-SA. **108** *Based on work by Wikimedia Commons user Geek3*: CC-BY-SA. **119** *Golden gate bridge*: Wikimedia Commons user Calibas, CC-BY-SA. **121** *Salt crystal*: Wikimedia Commons user Benjah-bmm27, public domain. **122** *Hand*: Wikimedia Commons user Janolaf30, public domain. **125** *Halley's comet*: NASA/W. Liller, public domain. **130** *Jello*: Photo by Mark Fickett, CC-BY-SA. **143** *Wikimedia Commons user Geek3*: CC-BY-SA. **152** *Terrier*: Based on a painting by Frederick August Wenderoth, 1875. **153** *Eve and the serpent*: Lemercier and Co., after Walter Crane, 1899. **155** *PSSC Physics*: . **161** *Window*: Redrawn from a photo by Wikimedia Commons user Rudiger Muller, CC-BY-SA. **165** *Dipole radiation pattern*: Redrawn from an animation by Wikimedia Commons user Chetvorno, public domain. **165** *Antenna*: Wikimedia Commons user Schwarzbek Mess-Elektronik, CC-BY-SA. **169** *Map plane by OpenStreetMap contributors and wmfabs.org*: CC-BY-SA.

173 *Crossed polarizing films*: Niels Bosboom, CC-BY-SA. **389** *Muon storage ring at CERN*: (c) 1974 by CERN; used here under the U.S. fair use doctrine. **192** *Multimeter*: Wikimedia Commons user Wdwd, CC-BY-SA. **200** *Chain drive*: Based on art by Wikimedia Commons user Keithonearth, CC-BY-SA. **200** *Railroad switching yard*: Arne Hueckelheim, CC-BY-SA. **216** *Impossible staircase*: Philip Ronan, public domain. **234** *Knitting*: Wikimedia Commons user Ikiwaner, CC-BY-SA. **234** *Puzzle*: Sam Loyd, 1914, PD. **235** *Russian dolls*: Wikimedia Commons users Fanghong and Gnomz007, CC-BY-SA. **257** *Doorbell ringer*: Wikimedia Commons user MNH, CC-BY-SA. **261** *Freeway*: Denys Nevohai, goodfreephotos.com, public domain. **261** *Einzel lens*: Based on a figure by Wikimedia Commons user Schnieri, CC-BY-SA. **271** *Twin lead cable*: Wikimedia Commons user LuckyLouie, CC-BY-SA. **279** *Transmission line*: Dave Bryant, CC-BY-SA. **325** *Oscilloscope trace*: Wikimedia Commons user Xato, public domain.

Index

- Q* (quality factor), 302
LRC circuit, 324
mechanical oscillator, 300
g factor, 269
- AC, *see* alternating current
aether, 156
alternating current (AC)
defined, 123
Ampère's law, 350
ampere (unit), 122
Archimedean spiral, 283
atom
energy scale, 78
planetary model, 79
raisin cookie model, 78
atomic clock, 16
- ballast, 319
Biot-Savart law, 274
black hole, 411
boundary conditions, 95
- capacitor, 43, 309
capacitance, 309
how electrostatic fields are uniquely determined, 85
causality, 378
charge, 29
conservation, 57
continuity equation, 259
junction rule, 210
fundamental (*e*), 56
invariance, 56
quantization, 56
- circuit, 191
complete, 191
open, 191
short, 198
- classical electron radius, 140
comet, 125
complete circuit, 191
complex numbers, 291
component, 22
Compton scattering, 423
- conductivity, 362
conductor, 92
defined, 91
correspondence principle
defined, 377
for mass-energy, 414
for relativistic momentum, 409
for time dilation, 377
- Coulomb constant, 29
Coulomb's law, 50
cross product, 25
curl, 86
component form, 101
curl-meter, 86
curl-meter, *see* curl
current, 122
density, 257
current density
transformation of, 262
cyclotron, 145
cyclotron frequency, 145
- DC, *see* direct current
DC circuit
flow of energy, 199
diamagnetism, 368
differential mode, 366
dipole, 132
electric, 43
field in mid-plane, 51
field of, 270
field on axis, 69
potential of, 106
energy due to orientation, 132
far field
equations, 271
universality of, 134
force in nonuniform field, 136
- magnetic
field of, 271
further properties, 267
- direct current (DC), 123
dispersion, 158, 302
div-meter, *see* divergence

divergence
 component form, 65
 div-meter, 61
 one dimension, 61
 three dimensions, 61
 dot product, 25
 relativistic, 404
 Earnshaw's theorem, 62
 raisin cookie model evades, 78
 Einstein, Albert, 166, 175
 Einzel lens, 261
 electric dipole
 field of, 270
 electric field
 related to potential, 93
 static
 conservative, 85
 transformation between frames of reference, 176
 unit, 29
 electromagnetic wave
 energy, 156
 geometry, 156
 momentum, 161
 not a vibration of a medium, 155
 propagation at c , 158
 electron
 classical radius, 140
 discovery, 76
 size, 140
 energy
 equivalence to mass, 137, 410
 of fields, 27
 energy-momentum four vector, 414
 equipotential, 97
 equivalent resistance
 of resistors in parallel, 211
 ether, 393
 Euler's formula, 294, 295
 Euler, Leonhard, 294
 farad
 defined, 309
 Faraday's law, 351
 Faraday, Michael, 26, 351
 ferrite bead, 366
 ferromagnetism, 368
 field
 basic properties, 18
 energy content, 27
 field-line representation, 26
 gravitational
 not observable, 20
 inertia of, 136
 observability, 19
 scalar, 19
 sea of arrows representation, 26
 vector, 18
 fluorescent light, 319
 flux, 46
 force
 on a charge in a magnetic field, 131
 on a charge in an electric field, 52
 on a wire in a magnetic field, 131, 146
 four-vector, 404
 energy-momentum, 414
 Franklin, Benjamin, 41
 fundamental charge e , 56
 fundamental theorem of algebra, 294
 fundamental theorem of calculus
 for div, grad, and curl, 347
 FWHM (full width at half maximum), 302
 g factor, 269
 gamma ray
 pair production, 416
 garage paradox, 390
 gas discharge tube, 319
 Gauss's law
 field lines, in vacuum, 45
 global form, 46
 local form, 60
 generator, 164
 gradient, 97
 fundamental theorem of calculus, 347
 Hafele-Keating experiment, 16
 Halley's comet, 125
 Hertz, Heinrich, 159
 homogeneity of spacetime, 383
 hysteresis, 370
 images
 method of, 100
 impedance, 331
 of an inductor, 332
 impedance matching, 336, 367

- inductance
 - defined, 311
- induction
 - static electric and magnetic, 135
- inductor, 309
 - inductance, 309
- inner product, 404
- junction rule, 210
- Kirchoff's current law, *see* junction rule
- Kirchoff's voltage law, *see* loop rule
- Laplace's equation, 95
- length contraction, 180
- light
 - electromagnetic wave, 17
 - momentum of, 422
 - speed, 17
- light cone, 399
- lightlike, 399
- loop rule, 216
 - in AC circuits, 324
- Lorentz force law, 131
- Lorentz invariance, 402
- Lorentz transformation, 384
- Lorentz, Hendrik, 384
- LRC circuit, 338
- lumped-circuit approximation
 - for capacitors, 310
- magnetic dipole
 - field of, 271
- magnetic field
 - circular loop of current, 273
 - long, straight wire, 121, 181, 273
 - found using Ampère's law, 357
 - found using Biot-Savart law, 282
 - solenoid, 350
 - transformation between frames of reference, 176
 - unit, 28
- magnitude, 21
- mass
 - equivalence to energy, 410
- mass-energy
 - conservation of, 412
 - correspondence principle, 414
 - of a moving particle, 413
- Maxwell's equations
 - full version, 263
 - in a vacuum, 163
 - in equations, words, and pictures, 440
 - in matter, 368
 - restricted to electrostatics, 99
- Michelson-Morley experiment, 393
- Millikan, Robert, 56
- momentum
 - of electromagnetic fields, 123
 - of light, 422
 - relativistic, 406, 414
- multimeter, 89
- natural units, 386
- neutron, 57
- Nichols-Hull experiment on momentum of light, 125
- nucleus, 79
- Oersted, Hans Christian, 120
- ohm (unit), 196
- Ohm's law, 196
- Ohm, Georg Simon, 196
- ohmic
 - defined, 196
- open circuit, 191
- overdamping, 324
- pair production, 416, 423
- paramagnetism, 368
- permeability-sideways, 365
- permittivity, 363
- planetary model, 79
- Poisson's equation, 95
- positron, 138, 412
- potential, 89
 - related to field, 93
- power
 - DC circuit, 194
- Poynting vector, 161
- Poynting, John Henry, 161
- pressure
 - of fields, 126
- quality factor (Q)
 - LRC circuit, 324
 - mechanical oscillator, 300
- quark, 270

radiation, *see* electromagnetic wave
 raisin cookie model, 78
 RC circuit, 321
 RC time constant, 321
 resistance
 defined, 196
 in parallel, 210
 in series, 215
 resistivity
 defined, 217
 resistor, 198
 resistors
 in parallel, 211
 resonance, 300
 RHIC accelerator, 390
 RL circuit, 322
 RMS (root mean square), 335
 root mean square, 335
 rotational invariance, 24
 Rutherford
 discovery of nucleus, 79
 scalar, 21
 defined, 21
 schematic, 209
 schematics, 209
 shear, 130
 short circuit
 defined, 198
 simple harmonic motion, 296
 skin depth, 362
 solenoid, 311
 magnetic field, 273, 350
 spacelike, 399
 spark gap transmitter, 322
 spark plug, 322
 spin, 270
 spiral
 Archimedean, 283
 Stern-Gerlach experiment, 136, 371
 Stokes's theorem
 for zero curl, 86
 general form, 349
 superconductor, 197
 superposition, 19
 tension, *see* pressure
 tesla (unit), 28
 There is also the time derivative $\partial/\partial t.$, 99

Thomson, J.J., 77
 time
 not absolute, 16
 time constant
 RC, 321
 RL, 322
 time dilation, 182
 timelike, 399
 Tolman-Stewart experiment, 95
 transistor, 196
 triangle inequality, 404
 twin paradox, 404

units
 natural relativistic, 386

vector, 21
 addition, 22
 analytic, 23
 graphical, 22
 components, 22
 cross product, 25
 defined, 21
 dot product, 25
 four-vector, 404
 magnitude, 21
 multiplication by a scalar, 22
 unit, 23

voltmeter, 89
 Voyager space probe, 186

wave
 plane, 154
 spherical, 154

wave, electromagnetic, *see* electromagnetic wave
 world-line, 404

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	v
acceleration	m/s^2	a
force	$\text{N} = \text{kg} \cdot \text{m}/\text{s}^2$	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	$\text{kg} \cdot \text{m}/\text{s}$	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	$h = 6.63 \times 10^{-34} \text{ J} \cdot \text{s}$

Maxwell's equations: the complete laws of electricity and magnetism.

As equations:

$$\operatorname{div} \mathbf{E} = 4\pi k\rho$$

$$\operatorname{div} \mathbf{B} = 0$$

$$\operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\operatorname{curl} \mathbf{B} = \frac{1}{c^2} \frac{\partial \mathbf{E}}{\partial t} + \frac{4\pi k}{c^2} \mathbf{j}$$

\mathbf{E} = electric field

\mathbf{B} = magnetic field

$\underbrace{\quad}_{\text{fields of force}}$

ρ = density of electric charge

\mathbf{j} = density of electric current

$\underbrace{\quad}_{\text{properties of matter}}$

In words:

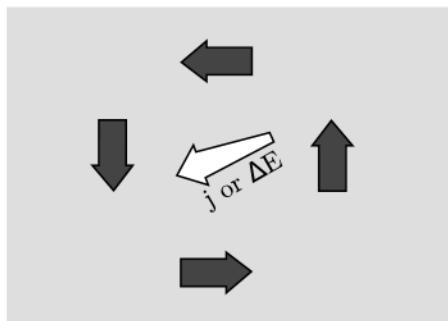
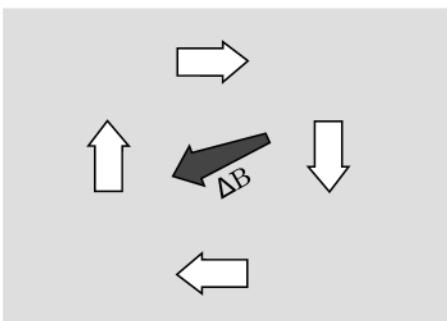
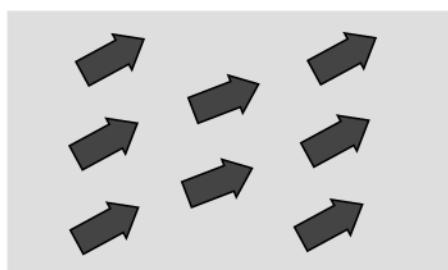
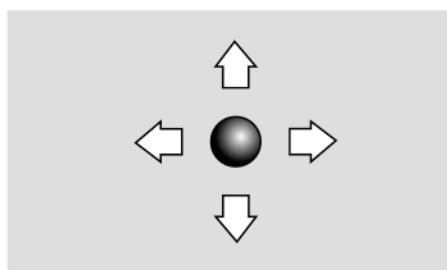
A charge creates an electric field that diverges from it.

Magnetic field patterns never diverge.

A changing magnetic field induces a curly electric field.

A changing electric field induces a curly magnetic field.
A current creates a curly magnetic field.

In pictures:



White arrows are electric fields. Dark arrows are magnetic fields.